DESTEK VEKTÖR MAKİNELERİ ÜZERİNE BİR ÇALIŞMA

A STUDY ON SUPPORT VECTOR MACHINES

MUSTAFA MURAT ARAT

PROF. DR. TURHAN MENTEŞ

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim Yönetmeliğinin

İstatistik Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2014

MUSTAFA MURAT ARAT' ın hazırladığı **"Destek Vektör Makineleri Üzerine Bir Çalışma"** adlı bu çalışma aşağıdaki jüri tarafından **İSTATİSTİK ANABİLİM DALI'** nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Ali YAZICI
Başkan ……

Prof. Dr. Turhan MENTEŞ
Danışman ……

Doç. Dr. Serpil AKTAŞ ALTUNAY
Üye ……..

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Fatma SEVİN DÜZ
Fen Bilimleri Enstitüsü Müdürü

Anneannem Düriye ÇİÇEK anısına…

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

/ /2014

MUSTAFA MURAT ARAT

ÖZET

DESTEK VEKTÖR MAKİNELERİ ÜZERİNE BİR ÇALIŞMA

Mustafa Murat ARAT

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Turhan MENTEŞ

Haziran 2014, 112 sayfa

Destek Vektör Makineleri, oluşturulan ve problemin çözümünde kullanılan etkili matematiksel optimizasyon yöntemleri sayesinde yüksek boyutlu ve küçük sayıda eğitim verisinden öğrenebilen yeni nesil bir öğrenme metodu olarak Vladamir Vapnik tarafından 1995 yılında önerilmiştir. Destek Vektör Makineleri'nin dayandığı teori olan "İstatistiksel Öğrenme Teorimi", Vladamir Vapnik ve Alexey Chervonenkis tarafından 1960'da önerilmiş ve 1970'li yıllarda geliştirilmiştir. Destek Vektör Makineleri, doğrusal olmayışı modellemek için çekirdek fonksiyonların kullanılması, genelleştirebilme yeteneğinin yüksek olması, teorik yapısının kuvvetli ve uygulamalarda hızlı performans göstermesi nedeniyle, son yıllarda, örüntü tanımlamada, bağlanım analizinde, yüz tanımlamada, resim ve metin sınıflandırmada, veri madenciliğinde, kalite kontrol yöntemlerinde, finans, ekonomi, genetik, biyoloji ve diğer biyoenformatik uygulamalarda sıklıkla kullanılmaya başlanmıştır. Bu nedenle bu tez çalışmasında Destek Vektör Makinelerinin teorik alt yapısı ve kompütasyonel yaklaşımlar detaylı bir şekilde ele alınıp incelenmiştir. Wisconsin Göğüs Kanseri veri seti üzerinde sınıflandırma analizi ve Hollanda'nın Brabant şehrine ait elektrik yük verisi ise bağlanım analizi için kullanılarak Destek Vektör Makineleri'nin performansı tanıtılmıştır.

Anahtar Kelimeler: Bağlanım analizi, çekirdek yöntemler, destek vektör makineleri, sınıflandırma

ABSTRACT

A STUDY ON SUPPORT VECTOR MACHINES

Mustafa Murat ARAT

Master of Science, Department of Statistics

Supervisor: Prof. Dr. Turhan MENTEŞ

June 2014, 112 pages

Support Vector Machines were introduced by Vladamir Vapnik in 1995 as a new-generation technique, learning from training set which is high-dimensional and whose sample size is much small, by creating and solving a quadratic programming problem using some novel mathematical optimization techniques. Statistical Learning Theory, which Support Vector Machines algorithm is based on, has been proposed in 1960s by Vladamir Vapnik and Alexey Chervonenkis and has been literally developed in 1970s. Due to the use of kernel functions to model non-linearity, high performance of generalization, powerful theoretical foundations and the ability to train relatively quickly, in the last decades, this method has been utilized frequently and mostly on pattern recognition, regression analysis, face recognition, image and text classification, data mining, quality control and applications of finance, economy, genetic, biology and bioinformatic. In this thesis, we discuss the theoretical basis and computational approaches to Support Vector Machines in details. Then, we discuss its performance on Wisconsin Breast Cancer dataset for classification and on an original electrical load data set of Brabant province of the Netherlands for regression.

Key Words: Kernel methods, support vector machines, regression, classification

TEŞEKKÜR

Öncelikli olarak, değerli tez danışmanım sayın Prof. Dr. Turhan MENTEŞ'e verdiği destek, harcadığı zaman ve bana olan güveni için teşekkürü bir borç bilirim.

Bu çalışmanın gerçekleşmesinde ve sonuca ulaştırılmasında büyük katkıları ve emeği olan, sonsuz sabırla beni her zaman çalışmaya teşvik eden ve tecrübelerinden yararlandığım sevgili hocalarım sayın Doç. Dr. Serpil AKTAŞ ALTUNAY'a, sayın Yrd. Doç. Dr. Semra TÜRKAN'a, sayın Doç. Dr. Gamze ÖZEL KADILAR'a, sayın Yrd. Doç. Dr. Nursel KOYUNCU'ya, sayın Arş. Gör. Dr. Nilgün ÖZGÜL'e ve sayın Arş. Gör. Dr. Esra POLAT'a teşekkürlerimi sunuyorum.

Tez sürecinde hoşgörüleri ve yardımları için çalışma arkadaşlarım sayın Arş. Gör. Emrah ALTUN'a, sayın Arş. Gör. Hatice YENİAY'a, sayın Arş. Gör. Nurbanu BURSA'ya ve ihtiyacım olan anlarda yanımda olan tüm değerli dostlarıma en kalbi şükranlarımı ve minnettarlığımı sunarım.

En önemlisi, hayatımın her anında sonsuz sevgilerini sunan, bugünlere gelmemde emek ve çaba harcayan, maddi ve manevi olarak desteklerini esirgemeyen ve her zaman yanımda olup beni yalnız bırakmayan biricik annem Fadime ARAT'a, babam Yalçın Kaya ARAT'a, ikizim Sinem ARAT'a, ablam İrem ARAT'a ve teyzelerim Muazzez ÇİÇEK, Mukadder ÇİÇEK ve Lale ÇİÇEK'e sonsuz teşekkürlerimi sunuyorum…

İÇİNDEKİLER DİZİNİ

Sayfa

ÖZET i
ABSTRACT ii
TEŞEKKÜR iii
İÇİNDEKİLER DİZİNİ iv
ŞEKİLLER DİZİNİ vi
ÇİZELGELER DİZİSİ vii
KISALTMALAR viii
1. GİRİŞ 1
2. DESTEK VEKTÖR MAKİNELERİ İÇİN ÇEKİRDEK YÖNTEMLER 5
2.1. Eğitimli Öğrenme 5
2.2. Özellik Uzayında Doğrudan Görüntüleme 6
2.3. Çekirdek Tabanlı Sonlu Özellik Uzayı 8
2.4. Hilbert Uzayları 10
2.5. Doğrusal Fonksiyoneller 13
2.6. Hilbert Uzayının Duali 16
2.7. Fonksiyonel Uzayın Arka Planı 17
2.8. Normlu Uzaylar 21
2.9. Kompakt ve Öz-eşlenik İşleçler – Spektral Teorem 22
2.10. İntegral İşleci 24
2.11. Çekirdeklerin Karakterize Edilmesi 27
2.12. Doğuran Çekirdekli Hilbert Uzayları 30
2.13. Uzayın ve Fonksiyonun Düzenlenmesi 35
2.13.1. Ivanov Düzenlemesi 38
2.13.2. Tikhonov Düzenlemesi 38
3. İSTATİSTİKSEL ÖĞRENME TEOREMİ 40
3.1. Deneysel Risk Minimizasyonu 41
3.2. Yapısal Risk Minimizasyonu ve Vapnik-Chervonenkis Teoremi 44
4. DESTEK VEKTÖR MAKİNELERİ 50
4.1. İkili Sınıflandırma İçin Destek Vektör Makineleri 50
4.1.1. Nokta Çarpımın Geometrisi 53
4.1.2. Hipotez Uzayını Düzenlemek ve Ayırıcı Düzlemler 55

4.1.3. Sert Sınır Sınıflandırıcıları 60
4.1.4. Yumuşak Sınır Sınıflandırıcıları 64
4.1.5. Doğrusal Olmayan Sınıflandırıcılar 69
4.2. Destek Vektör Regresyonu 72
5. UYGULAMA 79
5.1. Wisconsin Göğüs Kanseri Veri Seti 79
5.2. Kısa Dönemli Elektrik Yük Tahmini 82
6. SONUÇLAR 89
KAYNAKLAR 92
ÖZGEÇMİŞ 98

ŞEKİLLER DİZİNİ

Sayfa

Şekil 2.1. Özellik uzayına haritalama …….. 7
Şekil 3.1. İki boyutlu uzayda 3 ve 4 veri noktası için VC-boyutu ……. 45
Şekil 3.2. Büyüme fonksiyonu davranış grafiği …… 46
Şekil 3.3. Beklenen risk fonksiyonu yapısı …… 49
Şekil 4.1. Sınıflandırma problemi için kullanılan dört kayıp fonksiyonu …….. 52
Şekil 4.2. Dik izdüşüm olarak bir iç çarpım …….…...….... 54
Şekil 4.3. Sayısal izdüşüm …….……. 56
Şekil 4.4. Ayırıcı hiperdüzlem ve destek vektörleri ……... 59
Şekil 4.5. Yumuşak sınır sınıflandırıcısı …… 65
Şekil 4.6. ϕ dönüşüm operatörü ile haritalama ……..……... 70
Şekil 4.7. Bağlanım problemleri için kullanılan bazı kayıp fonksiyonları …….... 74
Şekil 4.8. Doğrusal DVR için yumuşak sınır kayıp düzeni …… 76
Şekil 5.1. C ve γ parametreleri için ağ-arama algoritması ……... 81
Şekil 5.2. 29 Eylül 2008 Perşembe ve 29 Ekim 2008 Çarşamba tarihleri arası
Brabant şehri saatlik elektrik yük talebi ……. 83
Şekil 5.3. 11.12.2008 günü içerisindeki saatlik yük değişimi ……... 85
Şekil 5.4. 03.11.2008-09.11.2008 haftasındaki ortalama yük tüketimi …….. 86
Şekil 5.5. DVR'nin performansı ……. 88

ÇİZELGELER DİZİSİ

Sayfa

Çizelge 5.1. Wisconsin göğüs kanseri veri seti …….. 79

Çizelge 5.2. Elektrik talebini etkileyen faktörler …….. 84

KISALTMALAR

Bu çalışmada kullanılmış bazı matematiksel simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Matematiksel Simgeler

$X \times Y$	Girdi-Çıktı (gözlem) uzayı
$S \in X \times Y$	Eğitim kümesi
n	Eğitim kümesinin büyüklüğü
$\mathcal{F}$	Özellik uzayı
$\phi : X \to \mathcal{F}$	Özellik uzayına doğrusal olmayan haritalama fonksiyonu
X	Tüm olası girdi vektörleri uzayı
$d = boyut(X)$	Girdi uzayının boyutu ($\vec{x}_i$ 'nin uzunluğu veya açıklayıcı değişkenlerin sayısı)
$\vec{x}_i \in X$	Bir girdi vektörü
$y_i \in \mathbb{R}$	Bağlanım analizi için çıktı değişkeni
$y_i \in \{+1, -1\}$	İkili sınıflandırma için çıktı değişkeni
y_t	x_t test örneği için çıktı değişkeni
Y	Açıklama (çıktı) uzayı
H	Hipotez (Hilbert) uzayı
$f \in H : X \to Y$	Bir hipotez (bağlanım, tahmin, karar) fonksiyonu
Y^X	X 'i, Y 'e haritalayan tüm olası fonksiyonlar uzayı
$K : X \times X \to \mathbb{R}$	Çekirdek fonksiyonu
K_S	K 'nın $S = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ eğitim kümesi ile sınırlandırılması
H_K	Doğuran çekirdekli Hilbert uzayı (DÇHU)
$\langle \cdot, \cdot \rangle_{H_K}$, $\lVert \cdot \rVert_{H_K}$	H_K DÇHU'da bir iç çarpım ve norm
$(\cdot \cdot)$	Öklid uzayında bir nokta çarpımı
$\forall g \in H, \mathcal{F}_g : H \to \mathbb{R}$	Doğrusal fonksiyonel
$\varepsilon_{\vec{x}} : H \to \mathbb{R}$	Değerlendirme fonksiyoneli
$P : H \to L$	H 'nin L alt uzayına izdüşüm işleci

$L^2(X)$	Karesi integrallenebilir fonksiyonlar uzayı
$T_K : L^2(X) \rightarrow L^2(X)$	İntegral İşleci
υ_i	T_K işlecinin özdeğeri
ς_i	T_K işlecinin υ_i özdeğerine ait özvektörü
$L(H, \mathrm{X})$	Kayıp sınıfı
$\ell(f, \{\vec{x}, y\})$	Kayıp fonksiyonu
R_X	Beklenen risk (hata)
R_{deney}	Deneysel risk (hata)
$V(\cdot)$	Vapnik-Chervonenkis (VC) boyutu
$\vec{w} \in \mathbb{R}^d$	Ağırlık vektörü
b	Eşik/eğilim değeri
$\mathfrak{H}$	Karar Hiperdüzlemi (Sınırı)
$\mathfrak{H}_+$ ve $\mathfrak{H}_-$	Sırasıyla pozitif ve negatif çıktı alan sınıflar tarafında bulunan destek hiperdüzlemleri
$\vec{x}_+^{sv}$ ve $\vec{x}_-^{sv}$	Sırasıyla pozitif ve negatif çıktı alan sınıfa ait destek vektörleri
$\vec{x}_0$	Karar hiperdüzlemi üzerindeki bir nokta
L_P	Primal Lagrange fonksiyonu
L_D	Dual Lagrange fonksiyonu
C	Ceza ve sınır arasında denge görevi gören düzenleme katsayısı
ξ	Gevşek değişken
ε	Regresyon tüpünün büyüklüğü
γ	Radyal tabanlı çekirdek fonksiyonunun parametresi

Kısaltmalar

ARMA	Otoregresif hareketli ortalamalar
ARIMA	Otoregresif bütünleşik hareketli ortalamalar
ARMAX	Dışsal açıklayıcı değişken içeren ARIMA
DÇHU	Doğuran Çekirdekli Hilbert Uzayı
DRM	Deneysel Risk Minimizasyonu
DVM	Destek Vektör Makineleri
DVR	Destek Vektör Regresyonu
GSYİH	Gayrisafi Yurtiçi Hâsıla
KKT	Karush-Kuhn-Tucker
SARMA	Mevsimsel ARMA
SARIMA	Mevsimsel ARIMA
VC	Vapnik-Chervonenkis
YRM	Yapısal Risk Minimizasyonu
YSA	Yapay Sinir Ağları

1. GİRİŞ

Genel anlamda veriden öğrenme, sonlu sayıda eğitim verisi kullanarak tahmin fonksiyonunun öğrenilmesini sağlayan öğrenme makinesi veya algoritmasının oluşturulması olarak tanımlanır. Öğrenme işleri genel olarak sınıflandırma, bağlanım (regresyon), kümeleme ve nitelik çıkarımı gibi işlemleri kapsar. Makine öğrenmesi yöntemleri istatistiksel açıdan dağılımdan bağımsız olarak çalışabilen yöntemlerdir. Bu teknikler de kendi içinde eğitimli ve eğitimsiz öğrenme yöntemleri olmak üzere sınıflandırılabilir. Klasik istatistiksel yöntemlere göre oluşturulan bağlanım ve sınıflandırma uygulamaları, olasılık dağılım fonksiyonlarına dayalı bir takım varsayımlara göre çalışır.

1930 ve 1960 yılları arasında geçen 30 yıl, parametrik çıkarsamalar için "altın çağ"dır. Bu süre boyunca istatistik çıkarsamalara tek bir yaklaşım hâkim olmuştur: olasılıksal bir dağılıma dayalı çıkarsama yaklaşımı. Ancak, 1960'larda, bilimsel ve uygulamalı problemlerin çözümünde bilgisayarların kullanımının yaygınlaşmasıyla, araştırmacılar, ilk defa, çok değişkenli özel modelleri analiz etmeye veya daha kesin yaklaşımlar elde etmeyi denemeye başlamışlardır. Yapılan tüm bu çalışmalar, olasılıksal dağılıma dayalı yöntemlerin bazı sebeplerden dolayı yetersiz olduğunu göstermiştir. Bu sebeplerden ilki yüksek boyutlu ve çok değişkenli problemlerin bilgisayarla analizinde ortaya çıkan ve R. Bellman tarafından isimlendirilen *boyut laneti (the curse of dimensionality)*'dir [1]. İkinci sebep ise, gerçek hayattan elde edilen verilerin analizinde, Tukey'in de belirttiği gibi, bu verilerin istatistiksel bileşenlerinin sadece klasik dağılım fonksiyonları ile tanımlanamayacağıdır. Uygulamada bu verilerin dağılımları farklılık gösterebilir ve analiz esnasında, etkili bir algoritma inşa etmek isteyenlerin bu farklılığı göz önüne alması gerekmektedir.

Dağılıma dayalı yöntemlerle yaşanan zorlukların keşfi istatistik biliminde bir dönüm noktası olmuştur. Bu keşif ile birlikte istatistik biliminde *veri analizi (data analysis)* denilen yeni bir yönelim doğmuştur. Veri analizinde amaç, araştırmacıların veriden tümevarımsal çıkarsamalar yapmasına yardımcı olmaktır. Bu nedenle, verinin görselleştirmesi, sınıflandırılması, tahmin edilmesi ve kümelendirilmesi gibi birçok amaç için çeşitli yeni teknikler geliştirilmiştir. Bu tekniklerin ilki 1958 yılında fizyolog F. Rosenblatt tarafından önerilen bir öğrenme makinesi (bilgisayar programı) olan *Algılayıcı (Perceptron)*'dır [2]. Algılayıcı algoritması, en basit öğrenme problemi olan doğrusal

sınıflandırma (örüntü tanımlama) problemini çözmek için kullanılmıştır. F. Rosenblatt en basit örneklerle bu algoritmanın genelleştirilebileceğini göstermiştir. Algılayıcı algoritmasından sonra birçok farklı öğrenme makinesi literatürde önerilmiştir. Hiç biri Algılayıcı algoritmasından daha kötü genelleştirilebilme potansiyeline sahip değildir ancak hiç birinin nörobiyolojik bir benzerliği bulunmamaktadır. Tüm bu makinelerin ortak özelliği ise Deneysel Risk Minimizasyonu (Empirical Risk Minimization - DRM)'na dayanmasıdır. DRM prensibi gerçek riskin yansız bir tahmin edicisi olan ve eğitim verisinden hesaplanan deneysel riskin minimize edilmesine dayanır. Ancak deneysel riski minimize eden birçok hatta sonsuz sayıda fonksiyonun bulunabilmesi ve bu tümevarımsal çıkarsama yönteminde aşırı uyumun sıklıkla karşılaşılması yeni bir risk minimizasyonunun oluşturulmasına neden olmuştur.

Doğası gereği, elde edilen veri sonlu ve kısıtlıdır. Genellikle bu verinin dağılımı *tek biçimli (uniform)* bir yapıya sahip değildir. Problemin çok boyutlu olması ise *girdi uzayında (input space)* verinin seyrek bir dağılım sergilemesine neden olmaktadır. Sonuç olarak, bu problemin *kötü konumlanmış (ill-posed)* bir problem olduğunu söylenebilir [3]. Ayrıca daha iyi bir eğitim için daha çok eğitim verisinin kullanılması gerekmektedir. Eğitim verisi yeterince büyük olduğu zaman (sonsuza yakın), eğitim hatası çok küçük (sıfıra yakın) olabilir. Fakat eğitim kümesi çok küçük ise, eğitim hatası oldukça büyük olacak ve öğrenme makinesinin verdiği sonuç güvenilir olmayacaktır. Eğer eğitim veri kümesi gürültü içeriyorsa, eğitim hatasını azaltmak için daha çok eğitim verisine ihtiyaç duyulur. Kecman (2001) [4], örneklem boyutunun belirlemek için eğitim veri sayısının Vapnik-Chervonenkis (VC) boyutuna oranını kullanmıştır. Ancak, uraya kadar dikkat edilen durumlar sadece eğitim hatası düşünülerek planlanmıştır. Genelleme hatası olarak bilinen, test verisi üzerindeki hata veya öğrenme yönteminin genelleme kapasitesi hiç hesaba katılmamıştır. Eğer giriş veri kümesi çok boyutlu ve veriye ait karakteristik fonksiyon çok karmaşık ise, öğrenme işleminde daha çok veriye ihtiyaç duyulur. Ayrıca boyut laneti, giriş uzayının boyutunu da arttıracaktır. Yapay Sinir Ağları (YSA) gibi geleneksel öğrenme makineleri yöntemlerinin yeterince eğitim verisine sahip olma ihtiyacı, düşük yakınsama oranı, yerel minimuma takılma problemi ve *aşırı uyum/eksik uyum (overfitting/underfitting)* gibi birçok eksiklikleri de vardır [5].

Bu gibi problemlerin ve eksikliklerin var olması yeni bir öğrenme makinesinin geliştirilmesi ihtiyacını ortaya çıkarmıştır. Tüm bu sebepler nedeniyle Destek Vektör Makineleri (DVM) yüksek boyutlu ve küçük sayıda eğitim verisinden öğrenebilen yeni nesil bir öğrenme metodu olarak Vladamir Vapnik tarafından 1995 yılında tanıtılmıştır. DVM'nin dayandığı İstatistiksel Öğrenme Teorisi, Vladamir Vapnik ve Alexey Chervonenkis tarafından 1960'lı yıllarda başlayıp 1970'li yıllarda gelişen başarılı bir çalışmanın ürünüdür. DVM, genelleştirebilme yeteneğinin yüksek olması, teorik yapısının kuvvetli ve uygulamalarda yüksek performans göstermesi bakımından son zamanlarda örüntü tanımlamada, bağlanım analizinde, yüz tanımlamada, resim ve metin sınıflandırmada, veri madenciliğinde, kalite kontrol yöntemlerinde, finans, ekonomi, genetik, biyoloji ve diğer biyoenformatik uygulamalarda sıklıkla kullanılmaya başlanmıştır.

Bu tez çalışmasının amacı DVM'nin teorik alt yapısının detaylı bir şekilde incelenmesidir. Tezin organizasyonu şu şekildedir. İkinci bölümde çekirdek yöntemler ve çekirdek yöntemlerinin tanımlandığı uzayların özellikleri detaylı bir şekilde anlatılmıştır.

Üçüncü bölümde, öğrenme algoritmaların teorik alt yapısını teşkil eden istatistiksel öğrenme teorisinden ve bazı öğrenme algoritmalarının dayandığı DRM'den bahsedilip, bu tümevarımsal prensibin avantajları ve dezavantajları tartışılmıştır. Daha sonra, neden DVM için yeni bir yaklaşıma ihtiyaç duyulduğu anlatılmış ve DVM'lerin dayandığı Yapısal Risk Minimizasyonu (YRM) prensibinden bahsedilip oluşacak karesel problemlerinin çözümü için dördüncü bölüme geçilmiştir.

Dördüncü bölümde DVM'nin temelleri hakkında ayrıntılı bilgi verilerek, sınıflandırma ve bağlanım problemlerinin oluşturulması gereken optimizasyon problemleri tanıtılıp bu problemlerin çözümleri gösterilerek global çözümün nasıl garanti edildiği ayrıntılı olarak açıklanmıştır. İlk olarak iki sınıf için doğrusal sınıflandırma problemi ele alınmış ve doğrusal sınıflandırma formülasyonları inşa edilmiştir. Doğrusal sınıflandırmanın mümkün olmadığı durumlarda kullanılan, DVM'nin temelini oluşturan çekirdek kavramından bahsedilmiştir. Çekirdek fonksiyonları ile, girdi uzayında doğrusal olarak sınıflandırılamayan verilerin, *özellik uzayına (feature space)* haritalama yapılarak doğrusal olarak sınıflandırılabilmeleri sağlanmıştır. Aynı işlemler bağlanım çözümlesi için gerçekleştirilmiştir ve bölüm sonlandırılmıştır.

Beşinci bölümde ise, iki sınıflı bir göğüs kanseri veri seti üzerinde DVM ile sınıflandırma ve Hollanda'nın Brabant şehrine ait elektrik yük verisi üzerinde Destek Vektör Regresyonu (DVR) kullanılarak kısa dönemli tahminlerin elde edilmesine ilişkin uygulamalar gösterilmiş olup, sonuçlar yorumlanıp tartışılmıştır.

2. DESTEK VEKTÖR MAKİNELERİ İÇİN ÇEKİRDEK YÖNTEMLER

Çekirdek yöntemlerinin tümü, bir çekirdek fonksiyonu yardımıyla, $\mathcal{X}$ *girdi uzayında* bulunan eğitim veri kümesini ayırıcı sınıflandırma ve bağlanım ile $\mathcal{F}$ *özellik uzayına* aktarır. Bu aktarma dolaylı haritalama veya izdüşüm ile yapılır. Haritalama ve izdüşümden kasıt, girdi uzayında doğrusal olarak ayrılamayan verinin, özellik uzayı olarak tanımlanan yüksek boyutlu bir uzayda görüntülenmesidir. Dolaylıdan kasıt ise orijinal girdi uzayında çalışırken, özellik uzayında gerekli hesaplamaların yapılmasıdır. Bu yüksek boyutlu özellik uzayının her bir koordinatı, araştırmacının elinde bulunan veriyi tanımlayan bir özellikten oluşmaktadır. Çekirdek fonksiyonu, bu koordinatları hesaplamak yerine veri noktalarından oluşan ikililer arasındaki bir iç çarpımı hesaplar. Yapılan bu işlem, koordinatların bulunmasından daha ucuz ve daha kolaydır. Bu yaklaşım *çekirdek hilesi (kernel trick)* olarak da bilinir. DVM, Gauss-tipi Süreçler, Fisher'in Doğrusal Diskriminant Analizi, Temel Bileşenler Analizi, Kanonik Korelasyon Analizi ve Ridge Regresyon çekirdek fonksiyonu ile çalışabilen bazı algoritmalardır.

2.1. Eğitimli Öğrenme

Eğitimli öğrenmenin ilk adımı,

$$S = \{\vec{x}_i, y_i\}_{i=1}^{n}, \qquad \vec{x}_i \in X \ , \ y_i \in Y$$

biçiminde gösterilen bir eğitim veri kümesi oluşturulabilecek bir olayın veya rastgele sürecin gözlemlenmesidir. Burada y_i, bağımlı değişken veya cevap değişkeni olarak adlandırılır. Yapılacak sınıflandırma veya bağlanım analizine göre Y, çıktı veya tanım uzayı, kesikli veya sürekli değer alır ve bu uzayın büyüklüğü $\mathbb{R}^k$ 'dır. Ancak, öğrenme algoritmasının birçoğunda $k = 1$ olarak kabul edilir [6]. $\vec{x}_i$ ise bağımsız değişken veya açıklayıcı değişken olarak isimlendirilir. X girdi uzayı, sonlu boyutlu bir reel uzaydır ve $\mathbb{R}^d$ ile gösterilir. Burada *d*, açıklayıcı değişkenlerin sayısıdır.

İkinci adım ise, X girdi uzayında tanımlı $\{\vec{x}_i\}_{i=1}^{n}$ gözlenen girdiler ile bu girdilere karşılık gelen ve Y açıklama uzayında tanımlı $\{y_i\}_{i=1}^{n}$ gözlenen çıktılar arasında $f : X \rightarrow Y$ olacak biçimde bir bağlantı bularak, bu olayı veya rastgele süreci modellemektir. f *hipotez/tahmin fonksiyonu (hypothesis/prediction function)* yapılacak olan analiz türüne

göre (sınıflandırma veya bağlanım) karar fonksiyonu veya bağlanım fonksiyonu olarak adlandırılır. Diğer bir ifadeyle, $f(\vec{x}_i) = \max_{y \in Y} P(y \mid \vec{x}_i)$ biçimindeki bilinmeyen koşullu olasılık yoğunluk fonksiyonu tahmin edilmek istenilmektedir. Bu hipotezin gözlenen eğitim kümesi üzerinde bir tür hata ölçüsünü minimize etmesi ve aynı zamanda basit bir fonksiyonel forma sahip olması istenir. Bunun için ilk koşul, $f : X \to Y$ bağlantısının gözlenen bir veriden çekilerek oluşturulmasının sağlanmasıdır. İkinci koşul ise, genelleştirilemeyen veya test örneklerinin çıktılarını doğru bir şekilde tahmin edemeyen karmaşık bir fonksiyonun oluşup, eğitim kümesinde aşırı uyuma neden olmasının engellenmesidir.

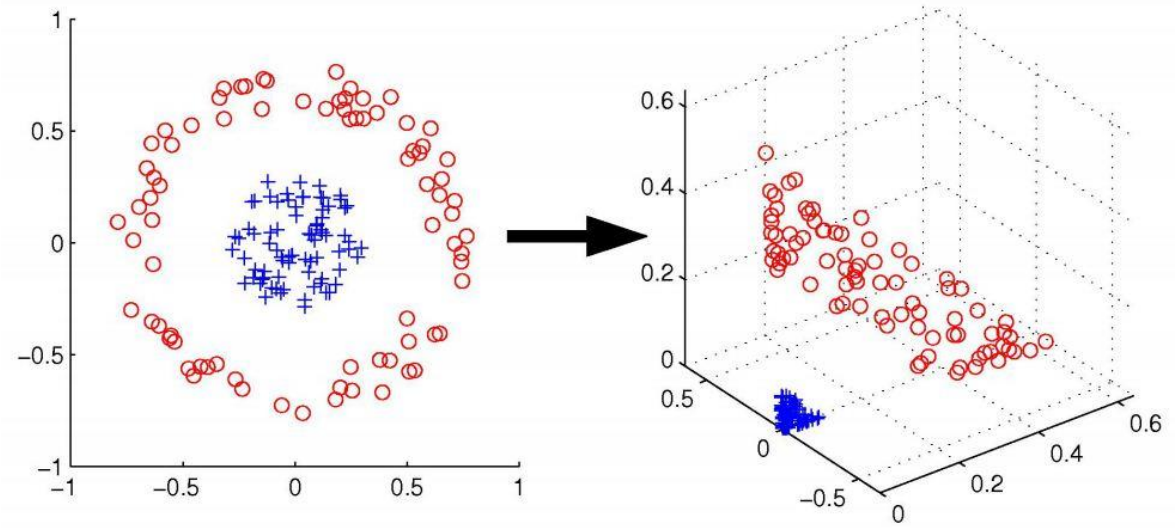
f hipotez fonksiyonunun karmaşıklığı, hipotez uzayının kapasitesi sınırlandırılarak kontrol edilebilir. Yani, girdi ve çıktı uzayları arasında X 'i Y 'e haritalayan tüm olası fonksiyonlarının uzayı Y^X 'in bir alt kümesi, hipotez uzayı $H \subset Y^X$ olacak şekilde seçilmelidir. Bu alt küme, $f : X \to Y$ bağlantısının iyi bir tahmini olan hipotez fonksiyonunu kapsayacak şekilde; aynı zamanda gereksiz bir karmaşıklığa sahip ve ayrıca, gözlenen veriye iyi uyum sağlayan ancak genelleştirilebilme potansiyeli oldukça düşük olan fonksiyonları kapsamayacak şekilde olmalıdır.

2.2. Özellik Uzayında Doğrudan Görüntüleme

Gözlem uzayından alınan eğitim veri kümesinin karmaşıklığı, kullanılacak herhangi bir öğrenme algoritmasının performansını etkileyebilir. Bu nedenle, çoğu zaman, verilen veri kümesi için uygun tahmin fonksiyonu bu algoritmalar tarafından bulunamayabilir. Bu gibi durumlarda verinin bir şekilde düzenlenmesi ve öğrenmenin mümkün kılınması gerekmektedir. Verinin doğrusal yapıda olmaması veya istenilen biçimde bulunmaması veriyi manipüle etme ve/veya yüksek boyutlu uzayda görüntülemeyi gerektirmektedir.

Şekil 2.1'de iki-sınıflı bir sınıflandırma verisi, verinin doğrusal yapıda olmamasına örnek gösterilebilir. Bu veri kümesi $\mathbb{R}^2 : x_1^2 + x_2^2 = 1$ ile dairesel karar sınırlarına sahiptir. Bu nedenle, bu veri $\mathbb{R}^3$ özellik uzayına, $\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ ile karesel haritalama kullanılarak doğrudan aktarılır ve doğrusal olarak ayrılabilir olması sağlanır. Böylece doğrusal bir algoritma, dönüştürme yoluyla doğrusal olmayan bir veride kullanılabilir ve sadece doğrusal yöntemler kullanılarak karesel (yani doğrusal olmayan) fonksiyonlar ile çalışılabilir. Bu stratejiye *önişleme (preprocessing)* adı verilir ve makine öğrenmesinin ilk

yıllarından beri popüler konumdadır [7]. Ancak, buradaki sorun, öğrenmeye devam etmeden önce, haritalama fonksiyonu ϕ 'nin önceden seçilmesidir. Bu örnekte karesel fonksiyonlar kullanılarak doğrusal yöntemlerin kullanılabileceği gösterilmiştir. Fakat bu örnekte olduğu gibi üç katsayılı bir karesel fonksiyon yerine, ϕ yüksek dereceli bir polinom olarak seçilirse, daha karmaşık karar fonksiyonlarına yaklaşım sağlanması mümkündür. Ancak, bu durum, özellik uzayının boyutunun yükselmesine ve serbest değişkenlerin sayısının artmasına neden olacaktır [8]. Elimizde bulunan 70x70 piksel örüntülerden oluşan bir girdi uzayını, yüksek boyutlu bir özellik uzayına haritalamak istediğimizi düşünelim. Doğrusal olmayışı 5'inci dereceden tek terimliler olarak seçersek eğer, elimizdeki girdi uzayının boyutu 4900 satırdan oluşacağından, özellik uzayının boyutu $\binom{5+4900-1}{5} \approx 2.3\times10^{17}$ olacaktır. Böyle bir özellik uzayında çalışmak imkânsızdır. Bu nedenle, bazı durumlarda haritalamanın kendisinin kullanılması verimsiz olabilir.



Şekil 2.1. Özellik uzayına haritalama

Verinin özellik uzayına aktarılmasında önemli olan, yeni bir uzayda haritalanan verinin gösterimi ve bu gösterime denk gelen fonksiyonun tanımlanmasıdır. Doğrusal bir algoritma ile doğrusal olmayan ilişkileri öğrenmek için, doğrusal olmayan özellikler kümesinin seçilmesi ve eğitim verisinin yeni bir gösterimle tekrar yazılması gerekmektedir [9].

$\phi(\bullet) \in \mathrm{H}$, eğitim kümesindeki girdiler $\vec{x}_i \in X$ üzerinde tanımlı doğrusal olmayan bir haritalama fonksiyonu olsun. Kullanılan eğitim veri kümesi aşağıdaki gibi gösterilirse,

$$S = \{\vec{x}_i, y_i\}_{i=1}^{n}, \qquad \vec{x}_i \in X \ , \ y_i \in Y$$

yüksek boyutlu bir özellik uzayında doğrudan haritalanan bu verinin yeni gösterimi ise,

$$\{\phi(\vec{x}_i), y_i\}_{i=1}^{n}, \qquad \phi(\vec{x}_i) \in H \ , \ y_i \in Y$$

biçiminde olacaktır. Ancak, her bir girdi örneğinin teker teker özellik uzayında haritalanması sırasında bazı problemlerle karşılaşılabilir. Bu problemlerden en sık rastlananı hesaplamaya dayalıdır çünkü ϕ haritalama fonksiyonu veri kümesindeki her bir elemanı sonsuz boyutlu özellik uzayına taşıyabilir.

2.3. Çekirdek Tabanlı Sonlu Özellik Uzayı

Doğrudan haritalama ile sonsuz boyutlu özellik uzayına gitme problemini engellemek amacıyla, dolaylı haritalama yapılması uygundur. $\phi : X \rightarrow \mathcal{F}$ haritalama fonksiyonu ile her bir eğitim örneği x_i 'nin teker teker $\phi(x_i)$ özelliklere haritalanması yerine, çekirdek yöntemleri kullanılarak, veri, ikili ölçümler kümesi olarak yeniden ifade edilebilir.

$$K : X \times X \rightarrow \mathbb{R} \tag{2.1}$$

Eşitlik (2.1)'de verilen K çekirdek fonksiyonunun, X sonsuz uzayında tanımlanması mümkündür. Bu nedenle bu çekirdeğin tanım kümesi, S eğitim kümesindeki gözlemler ile sınırlandırılabilir ve böylece sonlu bir çekirdek elde edilebilir.

$$K_S : \vec{x}_i \times \vec{x}_j \rightarrow \mathbb{R}, \qquad \forall i : 1 \leq i \leq n$$

$$(\vec{x}_i, \vec{x}_j) \rightarrow K_S(\vec{x}_i, \vec{x}_j)$$

Yukarıda görüldüğü gibi bu çekirdek $\vec{x}_i$ ve $\vec{x}_j$ gibi iki veri noktasını, benzerliklerini karakterize eden gerçel bir sayıya götürür [10]. Bu nedenle, sonlu çekirdekler, elemanları $k_{ij} = K_S(\vec{x}_i, \vec{x}_j) \in \mathbb{R}$ olacak biçimde $n \times n$ boyutlu kare bir matris olarak gösterilebilir. Gram matris olarak da isimlendirilen bu matrise, *Çekirdek Matrisi (Kernel Matrix)* adı verilir.

$$\begin{pmatrix} k_{11} & k_{12} & \dots & k_{1n} \\ k_{21} & k_{22} & \dots & k_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & k_{n2} & \dots & k_{nn} \end{pmatrix} \tag{2.2}$$

Çekirdek gösterimi başta sezgisel olmayan bir yöntem gibi görülebilir ancak ϕ haritalama fonksiyonunun doğrudan kullanımı üzerinde birçok faydası vardır. İlerleyen bölümlerde anlatılacağı üzere, bu iki yöntem aslında birbirlerinin aynısıdır ve aynı zamanda, her pozitif tanımlı çekirdek fonksiyonu için uygun bir özellik uzayı (Eşitlik (2.40) ve Eşitlik (2.44)) ve dolaylı bir haritalama (Eşitlik (2.31)) vardır. Dikkat etmek gerekirse, bu tür karşılaştırma fonksiyonları, sadece pozitif tanımlı çekirdekleri kullanmak için sınırlandırılmıştır. Böylece karesel programlama problemine ait amaç fonksiyonunu basitleştiren çekirdek hilesi kullanılabilir ve ana sonuca ulaşılabilir. Bu hile, pozitif tanımlılığın gerek ve yeter koşul olduğu Mercer Teoremi (Teorem 2.3) ile mümkündür [11]. Ayrıca, verinin yapısına bağlı olarak, gözlemlerin tekil gösterimlerini bulmak, ikili karşılaştırmalarını hesaplamaktan daha maliyetli ve zaman alıcı olabilir. Çoğu zaman ise, ikili karşılaştırmaları bulmak, tek tek her nitelik için vektör kullanmaktan daha faydalıdır [10].

Girdiler bir iç çarpım uzayında tanımlanmalıdır. Burada, bir iç çarpım benzerlik ölçütünün en basit halidir. Böylece $\vec{x}_i$ ve $\vec{x}_j$ girdilerinin bir iç çarpımı alınırsa, aşağıdaki doğrusal karşılaştırma fonksiyonunu elde edilir [10].

$$K\left(\vec{x}_i, \vec{x}_j\right) = \langle \vec{x}_i, \vec{x}_j \rangle_X \tag{2.3}$$

Eğer X reel vektör uzayında tanımlıysa, bu iç çarpım, nokta çarpımı olarak hesaplanır.

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j) \tag{2.4}$$

$\|\vec{x}_i\|$ ve $\|\vec{x}_j\|$ sırasıyla, $\vec{x}_i$ ve $\vec{x}_j$ vektörlerinin normları olmak üzere, nokta çarpımı, geometrik olarak, $(\vec{x}_i \cdot \vec{x}_j) = \|\vec{x}_i\|\|\vec{x}_j\|\cos\theta$ formülü ile bu vektörler arasındaki açının kosinüsünü verir. Burada dikkat edilmesi gereken nokta, bu vektörlerin birim (normalleştirilmiş) vektörler olmasıdır, yani, $\|\vec{x}_i\| = \sqrt{\langle \vec{x}_i, \vec{x}_i \rangle} = \sqrt{(\vec{x}_i \cdot \vec{x}_i)} = 1$'dir.

Eğer iç çarpımlar, X girdi uzayında iyi tanımlanmamış ise, ilk olarak yapılması gereken işlem, gözlenen girdilere ϕ gibi bir haritalamanın uygulanması ve bu girdilerin bir iç çarpım uzayına yansıtılmasıdır. Daha sonra aşağıdaki gibi bir karşılaştırma fonksiyonu kurulabilir.

$$K(\vec{x}_i, \vec{x}_j) = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle_H = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j) = \phi(\vec{x}_i)^T \phi(\vec{x}_j) \tag{2.5}$$

Bu haritalamanın yapılmasının sebebi, iç çarpım uzayında bulunmayan bir girdinin veya örüntünün herhangi bir *nesne (object)* olabilmesidir. Bir iç çarpımı veya nokta çarpımını, benzerlik ölçütü olarak kullanabilmek için, ilk olarak, bu girdilerin veya örüntülerin, H gibi bir iç çarpım uzayında vektör olarak gösterilmesi gerekmektedir. H uzayı burada özellik uzayıdır. Bir iç çarpım uzayının simetrik olma özelliğinden dolayı, burada da çekirdek fonksiyonu simetrik özelliğe sahiptir.

$$K(\vec{x}_i, \vec{x}_j) = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle = \langle \phi(\vec{x}_j), \phi(\vec{x}_i) \rangle = K(\vec{x}_j, \vec{x}_i) \tag{2.6}$$

Aynı zamanda bu fonksiyon Cauchy-Schwarz eşitsizliğini sağlar.

$$\begin{aligned} K(\vec{x}_i, \vec{x}_j)^2 = \langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle^2 &\le \|\phi(\vec{x}_i)\|^2 \|\phi(\vec{x}_j)\|^2 \\ &\le \langle \phi(\vec{x}_i), \phi(\vec{x}_i) \rangle \langle \phi(\vec{x}_j), \phi(\vec{x}_j) \rangle \\ &\le K(\vec{x}_i, \vec{x}_i) K(\vec{x}_j, \vec{x}_j) \end{aligned} \tag{2.7}$$

Ancak, bu iki koşul, her çekirdek fonksiyonu için bir özellik uzayının varlığını garanti etmez [12]. Bu nedenle, Mercer [11] çalışmasında, $K(\vec{x}_i, \vec{x}_j)$ gibi bir simetrik fonksiyonun, çekirdek olması için gerek ve yeter koşullar verilmiştir. Bu konu sonraki alt bölümlerde detaylı olarak verilecektir.

2.4. Hilbert Uzayları

Hilbert uzayları, içerisinde uzaklıkların[1] ve açıların[2] iyi tanımlandığı, tam iç çarpım uzaylarıdır. Bir Hilbert uzayı, tüm $h, g \in H$ fonksiyonları için tanımlanmış $\langle h, g \rangle$ iç çarpıma sahip bir fonksiyon uzayıdır. Doğrusal uzaylarda, bir iç çarpım yardımıyla

[1] Her iç çarpım uzayı bir normlu uzaydır ve her normlu uzay bir metrik uzaydır.
$d(\vec{x}_i, \vec{x}_j) = \|\vec{x}_i - \vec{x}_j\| = \sqrt{\langle (\vec{x}_i - \vec{x}_j), (\vec{x}_i - \vec{x}_j) \rangle}$

[2] Burada açılardan kasıt $(\vec{x}_i \cdot \vec{x}_j) = \|\vec{x}_i\| \|\vec{x}_j\| \cos\theta$ 'dır.

$\|h\|_H = \langle h,h\rangle^{1/2}{}_H$ gibi bir norm da tanımlanabilir. Burada "tam" ifadesinden kasıt, bu fonksiyon uzayında bulunan her Cauchy dizisinin yakınsak olmasıdır.

TANIM 2.1. (CAUCHY YAKINSAKLIK KRİTERİ) *$h_i \in H$ olacak biçimde her $\{h_i\}_{i=1}^{\infty}$ dizisinin yakınsak olması için gerek ve yeter koşul, bu dizilerin Cauchy kriterini sağlamasıdır. Her $\varepsilon > 0$ için öyle bir $N(\varepsilon)$ sayısı vardır ki, her $n,m > N(\varepsilon)$ için $\|h_n - h_m\|_H < \varepsilon$ sağlanır ve uzay içinde $\lim_{i\to\infty} s_i \in H$ şeklindeki bir limite yakınsar.*

N , H gibi bir Hilbert uzayının kapalı bir alt uzayı olsun (Sadece ve sadece $N = \bar{N}$ olduğu zaman, N kapalıdır denir). O halde bu alt uzayın dik tümleyeni $N^{\perp}$'i kapalı doğrusal bir alt uzay olarak tanımlayabiliriz (N bir alt uzay ise, $N^{\perp}$ da kapalı bir alt uzaydır) [13].

$$N^{\perp} = \{l \in H \mid \langle l,g\rangle = 0, \forall g \in \bar{N}\}$$

Burada, $\langle g,g\rangle = 0$ olduğu tek örnek, $g = 0$ olmasıdır. Bu durumda, $\bar{N} \cap N^{\perp} = \{0\}$'dır. Bu durum bir çelişki kurularak ispatlanabilir. Sıfır fonksiyonunun $\bar{N} \cap N^{\perp}$ kesişim uzayında olduğu aşikârdır. Çünkü her fonksiyon uzayı sıfır fonksiyonu kapsar. Ancak, $\bar{N} \cap N^{\perp}$ kesişim uzayının sıfır olmayan bir g fonksiyonuna sahip olduğunu düşünelim. Bu g fonksiyonu, $\bar{N}$'nin elemanıdır, aynı zamanda $N^{\perp}$'ün de bir elemanı olacaktır. $\bar{N} \perp N^{\perp}$ olduğundan $g \perp g$ olup, $\langle g,g\rangle = 0$'dır. $\langle g,g\rangle = 0$ olması $\|g\|^2 = 0$ anlamına gelmektedir. Sadece sıfır fonksiyonunun normu sıfır olduğundan, $g = 0$ olacaktır. Bu nedenle $\bar{N} \cap N^{\perp}$ kesişim uzayı sıfır fonksiyonundan başka bir fonksiyona sahip olamaz.

N'nin kapalılığı $\bar{N}$ ile dik tümleyeni $N^{\perp}$'ün dik toplamı, H Hilbert uzayına eşit olacaktır. Burada dikkat edilmesi gereken nokta, hem $\bar{N}$ hem de $N^{\perp}$'ün kendilerinin de birer Hilbert uzayı olmasıdır.

$$H = \bar{N} \oplus N^{\perp} = \{g + l \mid g \in \bar{N} \; ve \; l \in N^{\perp}\} \tag{2.8}$$

Ancak, bu iki alt uzayın birleşimi, H Hilbert uzayını kapsamak zorunda değildir.

$$\bar{N} \cup N^{\perp} \subseteq H \tag{2.9}$$

Bu nedenle H uzayında tanımlı herhangi bir h fonksiyonu, diğer iki fonksiyonun toplamı şeklinde yazılabilir.

$$h = g + l, \quad g \in \bar{N} \; ve \; l \in N^{\perp} \tag{2.10}$$

Böylelikle, her H Hilbert uzayı, iki farklı (sıfır vektörü hariç) kapalı doğrusal alt uzaya Eşitlik (2.10)'daki gibi parçalanabilir. Ancak, bu parçalama ikiden fazla birbirlerine dik alt uzaylar için de yapılabilir.

Zorn Ön savı [14] ve Gram-Schmidt dikleştirme yöntemi gereği, her Hilbert uzayı bir ortonormal (birimdik) tabana sahiptir. Ancak, bir Hilbert uzayı sadece sayılabilir bir ortonormal tabana sahip ise iki ve ikiden fazla alt uzaya ayrılabilir.

Sonlu bir Hilbert uzayı, $\{h_1, h_2, ..., h_i \in H\}$ şeklinde bir kümeye sahip olsun. Bu kümenin ortonormal olabilmesi için aşağıdaki ilk iki özelliği sağlaması gerekmektedir:

1. Normalleştirme: $\|h_i\| = 1, \quad \forall i \quad için.$

   Kümede bulunan her elemanın birim vektör olması gerekmektedir.
2. Diklik: $\langle h_i, h_j \rangle = 0, \quad i \neq j \quad için.$

   Kümede bulunan her elemanın birbirine dik olması gerekmektedir.

Bu iki koşulu sağlayan vektörler sistemine, ortonormal sistem veya ortonormal küme denir. Bu tür sistemler her zaman doğrusal bağımsızdır. Böylelikle, H uzayında bulunan her fonksiyon, bu sabit elemanların doğrusal birleşimi şeklinde yazılabilir. Bu da tamlık özelliğidir.

3. Tamlık: $\forall h \in H, \exists \{\alpha_1, \alpha_2, ..., \alpha_i \in \mathbb{R}\} \quad öyle \; ki \quad h = \sum_{i=1}^{\infty} \alpha_i h_i$. Bu şekilde oluşacak fonksiyonlar kümesine, H uzayının *sonlu doğrusal germesi (finite linear span)* nedir ve H uzayının bir alt uzayıdır. Eğer bu sonlu doğrusal germe H uzayında yoğunsa, $\{h_1, h_2, ..., h_i \in H\}$ kümesine tam ortonormal küme denir. Tanım gereği, eğer bu sonlu doğrusal germenin kapalılığı H uzayının tümüyse, bu sonlu doğrusal germe H uzayında *yoğundur (dense)* [15].

Yukarıdaki koşulları sağlayan *en büyük (maximal)* ortonormal kümeye, H Hilbert uzayının ortonormal tabanı denir.

Hilbert uzayı sonlu boyutlu fonksiyon uzayı ise, sonlu sayılabilir bir ortonormal tabana sahiptir ve kapalılığı uzayın tümü olan, sayılabilir yoğun alt kümeye sahiptir. Bu durumda

bu Hilbert uzayı kolaylıkla iki ve ikiden fazla alt uzaya parçalanabilir. Böylece, uzaydaki her fonksiyon ve bu fonksiyonlar üzerinde tanımlı her doğrusal işleç (operatör) matris formunda gösterilebilir.

2.5. Doğrusal Fonksiyoneller

Değişkenleri de fonksiyon (örneğin hipotez fonksiyonu $f : X \to Y$) olan fonksiyonlara, *fonksiyonel* denir. $\mathcal{F}$ ile gösterilen bir fonksiyonel, $\mathcal{F}: H(X \to Y) \to \mathbb{R}$ gibi bir dönüşümdür. H Hilbert uzayından alınan elemanlar, $\mathbb{R}$ uzayında gerçel sayılara dönüştürülür. *Doğrusal fonksiyoneller* ise, toplama ve skaler ile çarpma işlemlerinin korunduğu doğrusal vektör uzayları üzerinde tanımlanmıştır.

$$\mathcal{F}(\alpha_1 h_1 + \alpha_2 h_2) = \alpha_1 \mathcal{F}(h_1) + \alpha_2 \mathcal{F}(h_2), \quad \forall h_1, h_2 \in H \; ve \; \forall \alpha_1, \alpha_2 \in H$$

Toplanıp ölçeklendirilebilen fonksiyoneller kümesinin kendisi de τ gibi bir vektör uzayı kurabilir.

$$\mathcal{F}_1(\alpha_1 h) + \mathcal{F}_2(\alpha_2 h) = (\alpha_1 \mathcal{F}_1 + \alpha_2 \mathcal{F}_2)(h), \quad \forall \mathcal{F}_1, \mathcal{F}_2 \in \tau, \quad \forall h \in H, \quad \forall \alpha_1, \alpha_2 \in H$$

TANIM 2.2. (DEĞERLENDİRME FONKSİYONELİ) $\varepsilon_{\vec{x}}[f]: H(X) \to Y$ *ile gösterilen bir değerlendirme fonksiyoneli, uzayda bulunan her* $f \in H$ *hipotez fonksiyonunu, tanım kümesinde bulunan* $\vec{x} \in X$ *gibi sabit bir noktada hesaplayan (değerlendiren) fonksiyondur.*

$$\varepsilon_{\vec{x}}[f] = f(\vec{x}) \tag{2.11}$$

Hipotez uzayındaki noktasal yakınsama değerlendirme fonksiyonunun sürekliliğini sağlar.

$$f_n(\vec{x}) \to f(\vec{x}), \quad \forall \vec{x} \; için \Rightarrow \varepsilon_{\vec{x}}[f_n] \to \varepsilon_{\vec{x}}[f], \quad \forall \vec{x} \; için \tag{2.12}$$

$\mathcal{F}$ fonksiyonelinin çekirdek uzayı (çek) ve görüntü uzayı (gör) aşağıdaki şekilde tanımlanır:

$$Çek_{\mathcal{F}} \equiv \{h \in H \mid \mathcal{F}(h) = 0\}$$

$$Gör_{\mathcal{F}} \equiv \{\mathcal{F}(h) \mid h \in H\}$$

Bu fonksiyonelin çekirdek ve görüntü uzayı, sırasıyla, H tanım kümesinin ve $\mathbb{R}$ değer kümesinin alt uzaylarıdır. Rank-Nullity Teoremi'ne [16] göre, tanım kümesinin boyutu, çekirdek ve görüntü uzaylarının boyutlarının toplamına eşittir.

$$Boyut(H) = Boyut(Çek_{\mathcal{F}}) + Boyut(Gör_{\mathcal{F}})$$

Doğrusal bir fonksiyonel,

$$\left|\mathcal{F}(h)\right|_{\mathbb{R}} \leq \alpha \left\|h\right\|_H, \quad \forall h \in H \ \ için$$

olacak şekilde sabit bir $\alpha > 0$ reel sayısı varsa bu doğrusal fonksiyonel sınırlıdır denir. Bu eşitsizliği sağlayan en küçük α sabitine bu doğrusal işlecin normu adı verilir. Ek olarak, bir doğrusal işleç, sadece ve sadece sürekli olduğunda sınırlıdır. Bir Hilbert uzayında bir fonksiyonlar dizisinin, sabit bir fonksiyona yakınsadığını yani $h_i \to h$ iken $\left\|h_i - h\right\|_H \to 0$ olduğunu varsayalım. Bu durumda, $\mathcal{F}$ sınırlı doğrusal fonksiyonel için süreklilik ölçütü aşağıdaki gibidir:

$$\begin{aligned} &\forall \varepsilon > 0, \exists N(\varepsilon) \in \mathbb{N}, öyle\,ki\ \forall i > N(\varepsilon) \\ &\left|\mathcal{F}(h_i) - \mathcal{F}(h)\right| = \left|\mathcal{F}(h_i - h)\right| \leq \alpha \left\|h_i - h\right\|_H \to 0 \end{aligned} \tag{2.13}$$

$\{h_1, h_2, ..., h_i \in H\}$, H gibi bir Hilbert uzayı için ortonormal taban olsun. $h \in H$ olacak biçimde herhangi bir vektör, bu tabanın elemanlarının doğrusal kombinasyonu şeklinde şu şekilde yazılabilir:

$$h = \sum_{i=1}^{\infty} \alpha_i h_i = \sum_{i=1}^{\infty} \langle h, h_i \rangle h_i$$

Burada eşitliğin ikinci kısmı aşağıdaki gibi elde edilir.

$$\langle h, h_j \rangle = \left\langle \sum_{i=1}^{\infty} \alpha_i h_i, h_j \right\rangle \underset{a}{=} \sum_{i=1}^{\infty} \alpha_i \left\langle h_i, h_j \right\rangle \underset{b}{=} \alpha_j$$

Burada, (a) eşitliği bir iç çarpımın doğrusallığından ve sürekliliğinden, (b) eşitliği ise bu tabanın dik olması özelliğinden gelmektedir. Bu nedenle, sonsuz boyutlu bir Hilbert uzayında tanımlı herhangi bir *doğrusal* ve *sürekli* bir fonksiyonel, ortonormal bir taban kullanılarak, doğrusal fonksiyonellerin doğrusal bir biçimi şeklinde yazılabilir.

$$\mathcal{F}(h) = \sum_{i=1}^{\infty} \alpha_i \mathcal{F}(h_i) = \sum_{i=1}^{\infty} \langle h, h_i \rangle \mathcal{F}(h_i) = \sum_{i=1}^{\infty} \langle h, h_i \mathcal{F}(h_i) \rangle \tag{2.14}$$

TANIM 2.3. (İZ DÜŞÜM İŞLECİ) *G ve L, H gibi doğrusal bir vektör uzayının alt uzayları olsun. Burada L, G 'nin tümleyen alt uzayı ise, H uzayından $h \in H$ olacak şekilde bir vektör seçildiğinde, bu h vektörü, $h = g + l$, $g \in G$, $l \in L$ olacak biçimde parçalanabilir. $H = G \oplus L$ 'e G ve L alt uzaylarının dik toplamı denir. Bu parçalama işlemi, sadece ve sadece $G \cap L = \{0\}$ olduğunda tektir. Bir izdüşüm $P : H \to L$ ile gösterilen ve H uzayında bulunan noktaları $Ph = g$ ile H 'den, G alt uzayı boyunca, L alt uzayına haritalayan doğrusal bir işleçtir. Burada, P'nin görüntü uzayı L, çekirdek uzayı ise G 'dir. Böylelikle, $P^2 = P$ sağlanır. Yani izdüşümün iki kere uygulanması ile elde edilen sonuç, bir kere uygulanmasıyla elde edilen sonuç ile aynıdır. P idempotenttir. $(I - P)$ işleci ise, H uzayında bulunan noktaları, L boyunca, G'e haritalayan tümleyen bir izdüşümdür. Eğer G ve L kapalı alt uzayları, yani izdüşüm işlecinin görüntü ve çekirdek uzayları birbirlerine dik ise, bu izdüşüm diktir denir. Bu durumda, P işleci, öz-eşleniktir ($\langle Px, y \rangle = \langle x, Py \rangle$, $\forall x, y \in H$ için). Öz-eşlenik işleçler aynı zamanda normal işleçlerdir. Dik izdüşüm işleci zorunlu olarak sınırlıdır.*

P izdüşüm işlecinin üzerinde tanımlandığı H uzayı eğer sonlu-boyutlu ise, yani $boyut(H) = n$ ise, P işleci de sonlu boyutludur ve girdileri, L alt uzayının taban vektörlerinin bir fonksiyonu olan $n \times n$ boyutlu bir matris oluşturur. Bu matrise, izdüşüm matrisi adı verilir. Şekil 4.2'de $\vec{x}$ matrisinin $\vec{w}$ ağırlık vektörüne dik izdüşümü görülebilir. Bu durumda, izdüşüm matrisi şu şekilde verilir:

$$P_{\vec{w}} = \frac{\vec{w}}{\|\vec{w}\|} \frac{\vec{w}^T}{\|\vec{w}\|}$$

Burada, orijinden geçen (böylelikle eşik/eğilim değeri $b = 0$'dır.) $\mathfrak{H}$ ayırıcı hiperdüzleme paralel, $\vec{w}$'a dik herhangi bir vektör sıfıra haritalanır. O halde bir dik izdüşüm,

$$P_{\vec{w}} \vec{x} = \left(\frac{\vec{w}}{\|\vec{w}\|} \frac{\vec{w}^T}{\|\vec{w}\|} \right) \vec{x}$$

şeklindedir, ki bu da Eşitlik (4.6) ile verilen vektör izdüşümüne eşittir.

L kapalı alt uzayının $\{a_1, a_2, ..., a_t\}$ gibi ortonormal bir tabana sahip olduğu varsayıldığında, izdüşüm matrisi, sütunları, ortonormal tabanı kuran vektörlerden oluşan A_p matrisinin karesidir ve şu şekilde hesaplanır:

$$P_L = A_p A_p^T = \left[a_1 \mid a_2 \mid \ldots \mid a_t\right]\left[a_1 \mid a_2 \mid \ldots \mid a_t\right]^T$$

Eğer tabandaki vektörler ortonormal değilse, izdüşüm matrisi, $P_L = A_p A_p^T$, normalleştirilerek elde edilir.

$$P_L = A_p (A_p A_p^T)^{-1} A_p^T$$

Bu matris kullanılarak doğrusal bağlanımdaki normal denklemler elde edilir.

2.6. Hilbert Uzayının Duali

H Hilbert uzayı üzerinde $\mathcal{F}: H \to \mathbb{R}$ gibi tanımlanan tüm doğrusal fonksiyonellerin oluşturduğu uzaya, H uzayının dual uzayı adı verilir ve genellikle H^* ile gösterilir. Bu Hilbert uzayında tanımlı iç çarpım kullanılarak, doğrusal (sınırlı) bir fonksiyonel şu şekilde tanımlanabilir:

$$\mathcal{F}_g(\cdot) = \langle g, \cdot \rangle_H \in H^*$$

Eşitlik (2.1)'de verilen uzaydaki $K_{\vec{x}} = K(\vec{x}, \cdot) \in H$ gibi bir çekirdek fonksiyonu türünden tanımlanmış fonksiyonel ise, $\vec{x} \in X$ girdi vektörü için aşağıdaki gibi yazılabilir.

$$\mathcal{F}_{K_{\vec{x}}}(\cdot) = \langle K_{\vec{x}}, \cdot \rangle_H \in H^*$$

Dolayısıyla $g \in H$ (veya $K(\vec{x}, \cdot) \in H$) gibi her elemanın, dual uzay H^* 'da karşılık gelen doğrusal sınırlı bir fonksiyoneli vardır.

$$g \to \mathcal{F}_g(\cdot) = \langle g, \cdot \rangle_H \in H^*$$

H Hilbert uzayı üzerinde tanımlı bütün doğrusal sınırlı fonksiyonellerin H^* dual uzayı da bir Hilbert uzayıdır ve orijinal uzayın ortonormal tabanının bir fonksiyonu olan dual bir tabana sahiptir [13]. H ve H^* uzayları izomorfiktir, yani, H uzayında bulunan her eleman, H^* dual uzayında karşılık gelen bir elemana sahiptir. Bu ifadenin tersi de doğrudur. $g \in H$ gibi bir taban vektörünün karşılık geldiği fonksiyonelin çekirdek uzayı aşağıdaki gibidir ve bu uzay, H uzayında, g 'ye dik tüm vektörlerden (sıfır vektörü dahil) oluşur.

$$Çek_{\mathcal{F}_g} \equiv \{h \in H \mid \mathcal{F}_g(h) = \langle h, g \rangle_H = 0\} \tag{2.15}$$

Bu g vektörü, kendisi dışında diğer tüm taban vektörlerine dik olduğu için, oluşacak çekirdek uzayının boyutu, H uzayının boyutundan bir küçüktür. Burada, $Boyut(H)=n$ olduğundan, $Boyut(Çek_{\mathcal{F}})=n-1$ 'dir. Bu nedenle, bu çekirdek uzayına dik bir uzayın boyutu ise, Rank-Nullity Teoremi'ne göre 1'dir, yani, $Boy((Çek_{\mathcal{F}})^{\perp})=1$ 'dir.

TEOREM 2.1. (RIESZ GÖSTERİM TEOREMİ) *Bir Hilbert uzayının dual uzayından elde edilen en önemli sonuç Riesz Gösterim Teoremi'dir. Bu teorem, Hilbert uzayı H ile onun sürekli dual uzayı H^* arasındaki önemli ilişkiyi gösterir. H Hilbert uzayında tanımlı sınırlı ve sürekli her doğrusal fonksiyonel $\mathcal{F}$, $r_{\mathcal{F}} \in H$ gibi sabit, tek ve sıfır olmayan bir vektör ile bir iç çarpım olarak yazılabilir. Burada $r_{\mathcal{F}}$ 'ye $\mathcal{F}$ 'nin göstergesi denir* [17].

$$\exists r_{\mathcal{F}} \in H : \mathcal{F}(h) = \langle r_{\mathcal{F}}, h \rangle_H \in H^*, \quad \forall h \in H \tag{2.16}$$

Değerlendirme fonksiyoneli için Riesz Gösterim Teoremi ise şu şekildedir:

$$\forall \vec{x} \in X, \exists r_{\varepsilon_{\vec{x}}} \in H : f(\vec{x}) = \varepsilon_{\vec{x}}[f] = \langle r_{\varepsilon_{\vec{x}}}, f \rangle_H \in H^*, \quad \forall f \in H \tag{2.17}$$

2.7. Fonksiyonel Uzayın Arka Planı

Fonksiyon uzaylarını ve Doğuran Çekirdekli Hilbert Uzayları'nı (DÇHU) tanımlamadan önce, fonksiyonel analizden birkaç terimin tanımı bu bölümde verilecektir [18].

TANIM 2.4. (FONKSİYON UZAYI) *$\mathcal{F}$ ile gösterilen bir fonksiyon uzayı, elemanları fonksiyon olan bir uzaydır, yani $f : \mathbb{R}^d \to \mathbb{R}$.*

TANIM 2.5. (İÇ ÇARPIM) *$\mathcal{F}$ bir fonksiyon uzayı olsun. $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ ile gösterilen bir iç çarpım, her $f, g \in \mathcal{F}$ ve $\alpha \in \mathbb{R}$ için, aşağıdaki özellikleri sağlayan bir fonksiyondur.*

1. *Simetrik:* $\langle f, g \rangle = \langle g, f \rangle$
2. *Doğrusal:* $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$

3. *Pozitif-tanımlı: Bütün* $f \in \mathcal{F}$ *için* $\langle f, f \rangle \geq 0$ *ve* $\langle f, f \rangle = 0$ *sadece ve sadece* $f = 0$

TANIM 2.6. (NORM) $\mathcal{F}$ *bir fonksiyon uzayı olsun.* $\|\cdot\| : \mathcal{F} \to \mathbb{R}$ *ile gösterilen bir norm her* $f, g \in \mathcal{F}$ *ve* $\alpha \in \mathbb{R}$ *için, aşağıdaki özellikleri sağlayan ve negatif olmayan bir fonksiyondur.*

1. $\|f\| \geq 0$ *ve* $\|f\| = 0$ *sadece ve sadece* $f = 0$
2. $\|f + g\| \leq \|f\| + \|g\|$
3. $\|\alpha f\| = |\alpha| \|f\|$

Bir norm, bir iç çarpım ile $\|f\| = \sqrt{\langle f, f \rangle}$ *şeklinde tanımlanabilir.*

TANIM 2.7. (KARESİ İNTEGRALLENEBİLİR FONKSİYONLAR) *Hilbert uzaylarına en iyi örnek sonsuz boyutlu* $L^2(\mathrm{Z})$ *verilebilir.* $L^2(\mathrm{Z})$*, ölçüm uzayında tanımlı, gerçel değerler alan, karesi integrallenebilir, Lebesque ölçülebilir fonksiyonlar ailesidir ve aşağıdaki gibi gösterilebilir.*

$$\int_{\mathrm{Z}} \left| f(\vec{z}) \right|^2 d\vec{z} < +\infty \tag{2.18}$$

TANIM 2.8. (σ**-CEBRİ)** Z *bir küme olsun. Bir* σ*-cebiri, aşağıdaki koşulları sağlayan* Z *kümesinin alt kümelerinin boş kümeden farklı, tümleyen ve sayılabilir birleşim altında kapalı,* Σ *gibi bir sınıfıdır.*

1. $\varnothing, \mathrm{Z} \in \Sigma$
2. $S \in \Sigma \Rightarrow S^c \in \Sigma$ *(*S^c*,* S *kümesinin tümleyenidir.)*
3. $S_n \in \Sigma, n = 1, 2, \ldots, \Rightarrow \bigcup_{i=1}^{\infty} S_n \in \Sigma$

TANIM 2.9. (ÖLÇÜLEBİLİR UZAY) Z *bir küme ve* Σ*,* Z *'nin alt kümelerinin bir* σ*-cebiri olmak üzere* (Z, Σ) *ikilisi ölçülebilir uzaydır.*

TANIM 2.10. (ÖLÇÜM VE ÖLÇÜM UZAYI) *Z bir küme ve Σ, Z 'nin alt kümelerinin bir σ-cebiri olmak üzere, $\mu:\Sigma\to\overline{\mathbb{R}}^{+}$ fonksiyonu, aşağıdaki özellikleri sağladığı takdirde bir ölçümdür denir.*

1. $\mu(\varnothing)=0$
2. $S_j\in\Sigma,\ j=1,2,...,$ *ayrık kümeler ise,* μ *sayılabilir toplamsaldır.*

$$\mu\left(\bigcup_{j=1}^{\infty}S_j\right)=\sum_{j=1}^{\infty}\mu(S_j)$$

(Z,Σ,μ) *üçlüsüne ölçüm uzayı adı verilir.*

(Z,Σ,μ) ölçüm uzayı üzerinde tanımlı $L^2(\mathrm{Z})$ karesi integrallenebilir fonksiyonlardan oluşan sonsuz boyutlu uzayda, Z'nin tanım kümesi $[a,b]$ gibi kapalı aralığa veya (a,b) gibi açık aralığa sahip olması farklılık yaratmaz. Her iki durumda $\mu(\mathrm{Z})=b-a$ ile aynı Lebesque ölçüsüne sahiptir çünkü açık bir kümenin kapalılığı sıfır ölçüsüne sahiptir.

Daha genel olarak, $\mathrm{Z}=\mathbb{R}^n$ sonlu boyutlu reel uzayının kapalı veya açık herhangi bir alt kümesi Lebesque ölçülebilirdir ki bu durumda $L^2(\mathbb{R}^n)$ uzayı sonsuz boyutludur (Eğer σ-cebiri Σ sonsuz sayıda elemana sahipse, $L^2(\mathrm{Z})$ uzayı da sonsuz boyutludur). Böylelikle karesi integrallenebilir fonksiyonlar aşağıda verilen Lebesque integrali kullanılarak bir iç çarpım uzayı oluştururlar.

$$\langle f,g\rangle_{L^2}=\int_{\mathrm{Z}}f(\vec{z})g(\vec{z})\,d\mu(\vec{z}),\quad f,g:\mathrm{Z}\to\mathbb{R} \tag{2.19}$$

$f,g\in L^2(\mathrm{Z},\Sigma,\mu)$ için Cauchy-Schwarz eşitsizliği nedeniyle bu integral mevcuttur. Ek olarak uzayı tamlayan (complete) bir norm ve uzaklık ölçütü aşağıdaki şekilde tanımlanır. Bu norma aynı zamanda L^2-norm adı verilir.

$$\|f\|_{L^2}=\sqrt{\langle f,f\rangle_{L^2}}=\left(\int_{\mathrm{Z}}f(\vec{z})^2\,d\mu(\vec{z})\right)^{1/2} \tag{2.20}$$

$$\|f-g\|_{L^2}=\sqrt{\int_{\mathrm{Z}}\left(f(\vec{z})-g(\vec{z})\right)^2\,d\mu(\vec{z})} \tag{2.21}$$

$L^2(\mathrm{Z})$, Z'de tanımlı karesi integrallenebilir tüm fonksiyonları kapsar.

$$L^2(\mathrm{Z}) = \left\{ f : \mathrm{Z} \to \mathbb{R} \mid \| f \|_{L^2} = \sqrt{\langle f, f \rangle_{L^2}} = \left(\int_{\mathrm{Z}} f(\vec{z})^2 \, d\mu(\vec{z}) \right)^{1/2} < +\infty \right\} \tag{2.22}$$

$L^2(\mathrm{Z})$ fonksiyon uzayı bir Hilbert uzayıdır çünkü toplama altında kapalı bir iç çarpım uzayıdır.

$$f, g \in L^2(\mathrm{Z}) \Rightarrow f + g \in L^2(\mathrm{Z})$$

$L^2(\mathrm{Z})$ fonksiyon uzayı Riesz-Fischer Teoremi gereği tamdır [19]. Başka bir deyişle, bu fonksiyon uzayında bulunan her Cauchy dizisi yakınsaktır. $\{h_1, h_2, ..., h_i \in H\}$, karesi integrallenebilir fonksiyonların Cauchy dizisi ise, bu fonksiyon uzayının aşağıdaki koşulu sağlaması gerekmektedir.

$$\lim_{i,j \to \infty} \| h_i - h_j \|_{L^2} = \lim_{i,j \to \infty} \left(\int_{\mathrm{Z}} \left(h_i(\vec{z}) - h_j(\vec{z}) \right)^2 d\mu(\vec{z}) \right)^{1/2} = 0$$

Bu durumda Cauchy dizisinin ortalama limiti olan $h \in H$ gibi karesi integrallenebilir bir fonksiyon vardır.

$$\lim_{i \to \infty} \left(\int_{\mathrm{Z}} \left(h_i(\vec{z}) - h(\vec{z}) \right)^2 d\mu(\vec{z}) \right)^{1/2} = 0$$

Riesz Gösterim Teoremi'nden (Tanım 2.4), L^2 uzayında tanımlı her sınırlı, gerçel-değerli, doğrusal fonksiyonelin aşağıdaki formda yazılabileceği kolaylıkla görülebilir.

$$\exists r_{\mathcal{F}} \in L^2(\mathrm{Z}) : \mathcal{F}(g) = \langle r_{\mathcal{F}}, g \rangle_{L^2} = \int_{\mathrm{Z}} r_{\mathcal{F}}(\vec{z}) g(\vec{z}) \, d\mu(\vec{z}), \quad \forall g \in L^2(\mathrm{Z}) \tag{2.23}$$

$L^2(\mathrm{Z})$ fonksiyon uzayı aşağıdaki gibi genelleştirilebilir.

$$L^p(\mathrm{Z}) = \left\{ f : \mathrm{Z} \to \mathbb{R} \mid \| f \|^p = \left(\int_{\mathrm{Z}} \left| f(\vec{z}) \right|^p d\mu(\vec{z}) \right)^{1/p} < +\infty \right\} \tag{2.24}$$

Burada dikkat edilmesi gereken en önemli nokta sadece $p = 2$ olduğunda oluşan uzayın bir Hilbert uzayı olmasıdır. L^2'nin bir Hilbert uzayı olması sebebiyle, bu uzay sayılabilir bir ortonormal tabana sahiptir ve böylelikle parçalanabilirdir. Bunlara ek olarak, tanım kümesi kompakt dayanağa sahip olduğu sürece, sürekli fonksiyonlar da, L^2 uzayında

yoğundur. Yani, L^2 uzayındaki herhangi bir fonksiyon, bir sürekli fonksiyon tarafından doğru bir şekilde tahmin edilebilir. Temelde, L^2 uzayı, Eşitlik (2.20) ile verilen norm ile C^0 sürekli fonksiyonlar uzayının Cauchy tamlamasıdır [20].

Herhangi bir X tanım kümesi üzerinde, k 'nıncı dereceden türevlenebilen, gerçel değerli sürekli fonksiyonların oluşturduğu uzay $C^k(\mathbb{R}^X)$ ile gösterilsin. C^0 sürekli fonksiyonlar uzayıdır. C^1 türevi de sürekli olan sürekli fonksiyonlar uzayıdır. C^2 iki kez türevlenebilen fonksiyonlar uzayıdır. C^∞ ise sonsuz defa türevlenebilen *düzgün (smooth)* fonksiyonlar uzayıdır. L^2 ve C^0 uzayları arasındaki en büyük fark, C^0 uzayı Cauchy tam değildir, böylelikle, bir Hilbert uzayı değildir ($C^0[0,1]$ bir Banach uzayıdır. Bir Banach uzayı normlu bir tam vektör uzayıdır. [0,1] kapalı aralığı, $a<b$ olacak şekilde herhangi bir $[a,b]$ kapalı aralığına homeomorfiktir yani topolojik olarak aynıdırlar. Böylelikle $C^0[0,1]$ ve $C^0[a,b]$ de aynıdırlar. $C^0[0,1]$, herhangi bir iç çarpımdan oluşmayan supremum norm ile tamdır. Bu supremum norma herhangi bir iç çarpıma karşılık gelmediği için bir Hilbert uzayı değildir). Önceden de bahsedildiği gibi, L^2 uzayı, C^0 sürekli fonksiyonlar uzayının Cauchy tamlamasıdır, diğer bir ifadeyle X kümesi üzerinde tanımlı sürekli fonksiyonlar, $L^2(X)$ uzayında yoğundur.

2.8. Normlu Uzaylar

Bu bölümde, μ ölçüsünün sayma ölçüsü olduğu, böylelikle integral yerine toplamanın kullanıldığı L^p uzaylarının özel bir durumunu ele alınacaktır. Aslında, sayılabilir sonsuz uzunluğa sahip $\vec{z}$ gibi bir vektör ile gösterilen doğal sayılardan gerçel doğruya ($\mathbb{N}\to\mathbb{R}$) giden bir fonksiyona ulaşılır. O halde bu fonksiyona ait l^p -norm denilen ve $\|\cdot\|_p$ ile gösterilen norm aşağıdaki gibi tanımlanır.

$$\|\vec{z}\|_{l^p}=\left(\sum_{i=1}^{\infty}|z_i|^p\right)^{1/p}$$

Yukarıdaki serinin yakınsaması $\vec{z}$ vektörüne bağlıdır. Bu nedenle, l^p uzayı, sonlu ℓ^p -norma sahip, sonsuz uzunluklu $\vec{z}$ vektörlerinin oluşturduğu bir kümedir.

$$l^p(Z)=\left\{\vec{z}\in Z:\|\vec{z}\|_{l^p}<\infty\right\}$$

l^p uzayının büyüklüğü p'e bağlıdır ve burada $1 \leq p \leq \infty$'dur. Örneğin, l^∞, bütün sınırlı dizilerin oluşturduğu bir uzaydır ve bütün diğer l^p uzaylarının üst kümesidir. l^1, tüm mutlak toplanabilir yakınsak oluşturduğu bir uzaydır. l^2, tüm karesi toplanabilir yakınsak dizilerin oluşturduğu bir uzaydır. l^0 ise tüm sıfır dizilerinin (sıfıra yakınsar) oluşturduğu bir uzaydır. Tüm bu l^p dizileri arasında sadece ve sadece l^2 karesi toplanabilir diziler uzayı bir Hilbert uzayıdır çünkü aşağıdaki şekilde yazılabilen l^2-normu, diğer normlar aksine bir iç çarpımdan doğmaktadır. Bu norma aynı zamanda Öklid normu da denilmektedir (İki vektör arasındaki uzaklık bulurken oluşan norma da Öklid uzaklığı adı verilir).

$$\|\vec{z}\|_{l^2}^2 = \sum_{i=1}^{\infty} |z_i|^2 < +\infty$$

l^2 karesi toplanabilir diziler uzayı aynı zamanda bir DÇHU'dur.

2.9.Kompakt ve Öz-eşlenik İşleçler – Spektral Teorem

Doğrusal cebir, sonlu boyutlu uzaylar üzerinde çalışan işleçler için gerekli tüm bilgiyi sağlamaktadır. Sonsuz boyutlu uzaylar içinde durum daha da karmaşık bir hal almaktadır. Ancak, sonlu boyutlu uzaylar için üretilmiş teorilerin geçerli olduğu sonsuz boyutlu uzaylar üzerinde işlem gören işleçler de vardır. Bu tür işleçlere *kompakt işleçler* denir.

TANIM 2.11. (KOMPAKT İŞLEÇLER) *Bir Hilbert uzayından, diğer bir Hilbert uzayına tanımlı sınırlı (sürekli) doğrusal bir işleç* $T : L^2(\mathbb{R}^X) \to L^2(\mathbb{R}^X)$ *verilsin. Bu işleç,* $L^2(\mathbb{R}^X)$ *tanım kümesinin sınırlı herhangi bir altkümesinin elemanlarına uygulansın. Eğer,* $L^2(\mathbb{R}^X)$ *'in sınırlı her* S *altkümesi için, oluşan görüntü uzayının kapalılığı,*

$$\overline{\{(Tf) : f \in S\}} \subset L^2(\mathbb{R}^X)$$

kompakt (tam ve tümel sınırlı) veya oluşan bu görüntü uzayı prekompakt (tümel sınırlı) ise T *işleci kompakttır denir. Her kompakt doğrusal işleç sınırlıdır, dolayısıyla süreklidir.* T *kompakt ise, öz-eşleniği de kompakttır.*

TANIM 2.12. (ÖZ-EŞLENİK İŞLEÇLER) *Eğer T gibi bir doğrusal işleç T^* ile gösterilen adjointine eşit ise ($T = T^*$), bu doğrusal işleci öz-eşleniktir denir. Yani,*

$$\langle Th, g \rangle = \langle h, T^* g \rangle$$

'dir. Bir öz-eşlenik işlecin tüm öz değerleri gerçeldir. Sonlu boyutlu bir uzayda, bir öz-eşlenik işleç, aynı zamanda eşlenik simetriktir.

$\mathcal{F} : h \to \langle g, Th \rangle, \quad \forall h, g \in H$ gibi sınırlı (sürekli) doğrusal bir fonksiyonel tanımlayan her T işlecinin adjointinin varlığı ve tekliği Riesz Gösterim Teoremi'nden gelir.

$$\exists r_{\mathcal{F}} \in H : \mathcal{F}(h) = \langle g, Th \rangle_H = \langle r_{\mathcal{F}}, h \rangle_H \in H^*, \quad \forall h \in H$$

Bu durumda, $T * g = r_{\mathcal{F}}$'dir.

Kompakt ve öz-eşlenik işleçlerini tanımladıktan sonra, sıra, bu işleçlerin görüntü uzaylarının tabanlarının var olduğunu göstermeye geldi. Bu gösterim Spektral Teoremi ile yapılır.

TEOREM 2.2. (SPEKTRAL AYRIŞIM TEOREMİ) $T : H_D \to H_R$ *gibi gösterilen her kompakt, öz-eşlenik işleç, $f \in H$ gibi Hilbert uzayında tanımlı bir fonksiyona uygulandığında, bu fonksiyon aşağıdaki gibi bir spektral ayrışıma sahip olur.*

$$Tf = \sum_{i=1}^{\infty} \alpha_i P_{H_i}[f] \in H \tag{2.25}$$

Burada α_i'ler karmaşık sayı ve $P_{H_i}[f]$ f'nin H_i üzerine dik izdüşümü olacak şekilde her H_i, H_D'nin kapalı bir alt uzayıdır.

Birbirlerine dik ve birbirlerini tümleyen tüm bu alt uzayların (tanım kümesinin çekirdek uzayı veya H_0 sıfır öz uzayı hariç) dik toplamı işlecin görüntü uzayına eşittir.

$$H_R = H_1 \oplus H_2 \oplus H_3 \oplus \cdots$$

T işleci kimi zaman *özayrışım (eigendecomposition)* denilen aşağıdaki ayrışımı da yapabilir.

$$T\varsigma_i = \upsilon_i \varsigma_i \tag{2.26}$$

Burada, ς_i 'ye bu işleç için özfonksiyon, υ_i 'e bu işlecin özdeğeri denir. T işlecinin özfonksiyonları görüntü uzayının tam, sayılabilir ortonormal bir tabanını oluştururlar. Böylelikle, her H_i alt uzayı, aynı özdeğerlerden oluşmuş özfonksiyonların bir tabanına sahip olur. Bu durumda Eşitlik (2.25)'teki ayrışım şu şekilde yazılabilir:

$$Tf = \sum_{j=1}^{\infty} \upsilon_j P_{\varsigma_j}[f] \in H \tag{2.27}$$

Bu durumda, $P_{\varsigma_j}[f]$, f 'nin ς_i özfonksiyonu üzerine dik izdüşümüdür. Farklı alt uzaylar ilgili özfonksiyonları birbirine dik olan farklı özdeğerlere sahiptir.

$$H_i \neq H_j \Rightarrow \upsilon_i \neq \upsilon_j \Rightarrow \left\langle \varsigma_i, \varsigma_j \right\rangle_H = 0$$

Ancak bunun tersi doğru değildir; Birbirine dik iki özfonksiyon aynı özdeğere sahip olabilir ve aynı alt uzay için taban vektörleri olabilir. İşlecin tanım kümesi sonlu n-boyutlu bir uzay ise, bu uzaya ait n tane özfonksiyon ve ilgili özdeğer vardır. İşleç pozitif tanımlı ise, özdeğerler de pozitif gerçel sayılardır ve tümüyle yakınsaktır.

Örneğin, $f \in L^2(\mathbb{R}^X)$ gibi tek bir fonksiyon ve bu uzayın Υ gibi sınırlı bir alt uzayı alınsın. Bu sınırlı alt uzay bir birimlik kapalı bir top (an unit closed ball) olabilir.

$$\Upsilon = \left\{ f, g \in L^2(\mathbb{R}^X) : \| f - g \|_{L^2} \leq 1 \right\}$$

T kompakt operatörün, Υ sınırlı alt uzayındaki tüm elemanlara uygulanması sonucu, kapalılığı kompakt olan bir görüntü uzayı oluşur, yani sonlu boyutludur. T işlecinin, Υ 'daki herhangi bir fonksiyona uygulanması sonucu oluşacak yeni fonksiyon, Eşitlik (2.25) veya Eşitlik (2.26)'daki gibi dik taban vektörlerinin sonlu doğrusal kombinasyonu şeklinde yazılabilir.

2.10. İntegral İşleci

Bu bölümde bir fonksiyona yapılacak dönüşüme ilişkin temel bilgiler verilecektir. Aslında amaç, bir fonksiyonun, manipüle edilmesi zor olan bir uzaydan, manipüle edilmesi daha kolay olan basit fonksiyonların bir toplamı şeklinde gösterilebilecek bir uzaya dönüşümünü sağlamaktır. Eğer varsa, ilgili bir ters dönüşüm işleci ile bu fonksiyon daha sonra orijinal

uzaya geri döndürülebilir. İlk olarak bu dönüşüm işleci ve bu işlece bağlı çekirdeğe ait tanımlamalar yapılacaktır.

TANIM 2.13. (İNTEGRAL İŞLECİ) *K gibi bir çekirdek fonksiyonuna sahip olunsun. μ Lebesque ölçüsü olmak üzere, bir fonksiyon uzayından diğerine aşağıda verilen dönüşüm tüm $f \in L^2(X)$ 'ler için hemen hemen her yerde geçerli ise, $T_K : L^2(X) \to L^2(X)$ doğrusal işleci bir integraldir.*

$$\left(T_K f\right)(\cdot) = \int_X K\left(\cdot, \vec{x}\right) f\left(\vec{x}\right) d\mu\left(\vec{x}\right) \tag{2.28}$$

Görüntü uzayı sonlu boyutlu ise, T_K integral dönüşümü, girdi fonksiyonu f 'nin gösterimini, çıktı fonksiyonu $\left(T_K f\right)$ olarak değiştirir. $\left(T_K f\right)$, dik taban fonksiyonlarının sonlu bir kümesinin doğrusal bir birleşimi olarak yazılır.

$$\left(T_K f\right) = \sum_{i=1}^{n} \alpha_i f_i \text{ öyle ki, } \left\langle f_i, f_j \right\rangle = 0, \quad \forall i, j < b \tag{2.29}$$

TANIM 2.14. (POZİTİF TANIMLILIK) *Sıfırdan farklı herhangi bir $c = \left[c_1, c_2, ..., c_n\right]^T \in \mathbb{R}^n$ katsayı vektörü için, bir gerçel simetrik $n \times n$ boyutlu K matrisinin ilgili karesel formu pozitif ise, bu K matrisi pozitif tanımlıdır denir. Yani,*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\vec{x}_i, \vec{x}_j) > 0$$

'dir. Çalışmanın amacına uygun olarak, burada, K çekirdeği, iki değişkenin gerçel değerli bir fonksiyonudur.

$$K : X \times X \to \mathbb{R}, \quad (\vec{x}_i, \vec{x}_j) \to K(\vec{x}_i, \vec{x}_j)$$

James Mercer, integral işleçleri ile ilgilendiği için, sonlu-boyutlu kesikli karesel form yerine, sonsuz boyutlu karesel bir form kullanılırsa, aşağıda verilen pozitif çekirdek tanımı elde edilir.

TANIM 2.15. (POZİTİF ÇEKİRDEK) X , $\mathbb{R}^d$ 'nin kompakt bir alt kümesi olsun. K *gibi simetrik bir çekirdek, aşağıdaki koşulu sağlıyorsa, bu çekirdek* X *üzerinde pozitif tanımlıdır.*

$$\iint_X K(\vec{x},\vec{y}) f(\vec{x}) f(\vec{y}) d\mu(\vec{x}) d\mu(\vec{y}) > 0, \quad \forall f \in L^2(X)$$

Bir sonlu çekirdek $X \times X$ uzayındaki vektörlerin olası tüm sonlu kümeleri üzerinde pozitif tanımlıysa, bu çekirdek pozitiftir. Ek olarak, tanım kümesinde bulunan tüm fonksiyonlar pozitif ise ($f > 0$), integral işleci de $T_K f > 0$ pozitiftir. Bu durumun tersi de doğrudur.

TANIM 2.16. (SÜREKLİ ÇEKİRDEK) $K \in C^0(X \times X)$ *gibi bir çekirdek, aşağıdaki koşulu sağladığında,* $(\vec{b},\vec{c}) \in X \times X$ *gibi bir noktada süreklidir denir.*

$$\begin{aligned} &\forall \varepsilon > 0, \exists \delta > 0, \\ &\forall \vec{x},\vec{s} \in X, \vec{b} - \delta < \vec{x} < \vec{b} + \delta, \vec{c} - \delta < \vec{x} < \vec{c} + \delta \\ &\Rightarrow K(\vec{b},\vec{c}) - \varepsilon < K(\vec{x},\vec{s}) < K(\vec{b},\vec{c}) + \varepsilon \end{aligned} \tag{2.30}$$

Eğer K çekirdeği simetrik ise, T_K integral işleci öz-eşlenik olmalıdır. Bunun için $\langle (T_K f), g \rangle_{L^2} = \langle f, (T_K g) \rangle_{L^2}$ olduğu ispatlanması gerekmektedir. $f, g \in H$ gibi iki hipotez fonksiyonu olsun.

$$\begin{aligned} \langle (T_K f), g \rangle_{L^2} &\underset{a}{=} \int_X T_K f(\vec{y}) g(\vec{y}) d\mu(\vec{y}) \\ &\underset{b}{=} \int_X g(\vec{y}) \left(\int_X K(\vec{y},\vec{x}) f(\vec{x}) d\mu(\vec{x}) \right) d\mu(\vec{y}) \\ &= \int_X \int_X g(\vec{y}) K(\vec{y},\vec{x}) f(\vec{x}) d\mu(\vec{x}) d\mu(\vec{y}) \\ &\underset{c}{=} \int_X f(\vec{x}) \left(\int_X K(\vec{x},\vec{y}) g(\vec{y}) d\mu(\vec{y}) \right) d\mu(\vec{x}) \\ &= \int_X f(\vec{x}) T_K g(\vec{x}) d\mu(\vec{x}) \\ &= \langle f, (T_K g) \rangle_{L^2} \end{aligned}$$

Burada, (a) eşitliği Eşitlik (2.19)'dan gelmektedir. (b) eşitliği ise Tanım 2.13.'teki İntegral İşleci tanımından oluşturulmuştur. (c) eşitliğine Fubini Teoremi uygulanmıştır. Daha ileri gidilip, K çekirdeğinin sürekli olduğu varsayılırsa, karesi integrallenebilir özelliğinden,

$$\int_X \int_X \left|K\left(\vec{x},\vec{y}\right)\right|^2 d\mu\left(\vec{x}\right) d\mu(\vec{y}) < \infty$$

elde edilir. Bu özellik gereği bu integral işleci aynı zamanda bir Hilbert-Schmidt işlecidir. Bu durumda, $X \times X$ tanım kümesinin herhangi sınırlı bir alt uzayı için, T_K integral işleci altında oluşan görüntü uzayının $L^2(X)$ uzayında prekompakt olduğu kolaylıkla görülebilir. Bu durumda Eşitlik (2.28) ile verilen T_K integral işleci kompakttır.

Özet olarak, K çekirdeği simetrik, pozitif ve karesi integrallenebilir olduğu zaman, T_K integral işleci pozitif, öz-eşlenik ve kompakttır. O halde Teorem 2.2. ile verilen Spektral Ayrışım Teoremi gereği, T_K integral işleci negatif olmayan özdeğerlere sahiptir, böylelikle bu özdeğerlere karşılık gelen özfonksiyonlar ($\varsigma_1, \varsigma_2, \varsigma_3, \ldots$), tüm bu özfonksiyonların normalleştirildiği $\|\varsigma_i\|_{L^2} = 1$ varsayımı altında, $L^2(X)$ görüntü uzayının tam, sayılabilir ortonormal bir tabanını oluştururlar.

2.11. Çekirdeklerin Karakterize Edilmesi

Çekirdek fonksiyonlarını karakterize eden en önemli teorem Mercer Teoremi'dir [11]. $K(\vec{x},\vec{s})$, X üzerinde tanımlı simetrik bir fonksiyon olsun. O halde, bir öğrenme problemi için $K = \left(K(\vec{x}_i,\vec{x}_j)\right)_{i,j=1}^{n}$ gibi bir matris yazılabilir. Λ , K matrisinin özdeğerlerini kapsayan köşegen matris olmak üzere, K simetrik olduğundan, $K = V\Lambda V'$ şeklinde parçalanabilir. Burada, V , sütunları, K matrisinin özdeğerlerine karşılık gelen özfonksiyonlardan oluşan dik bir matristir. Bu parçalamaya özdeğer parçalanması da denir [21]. Bütün özdeğerlerin negatif olmadığı varsayılırsa, özellik haritalaması şu şekilde yapılabilir.

$$\phi : \vec{x} \to \left(\sqrt{\upsilon_1}\varsigma_1\left(\vec{x}\right), \sqrt{\upsilon_2}\varsigma_2\left(\vec{x}\right), \ldots\right) \in l^2$$

O halde, özellik uzayında iki fonksiyonunun L^2 iç çarpımı,

$$\langle \phi(\vec{x}), \phi(\vec{s}) \rangle = \sum_{i=1}^{\infty} \upsilon_i \varsigma_i\left(\vec{x}\right)\varsigma_i\left(\vec{s}\right) = K(\vec{x},\vec{s}) \tag{2.31}$$

elde edilir. Buradan görüldüğü üzere, $K(\vec{x},\vec{s})$, ϕ özellik haritalamasına karşılık gelen bir çekirdek fonksiyonudur. Bu durumda, özellik uzayında bir iç çarpımın bir çekirdek fonksiyonu olarak yazılabileceği söylenebilir. Mercer Teoremi ise bu durumun tersini ispatlar ve pozitif, sürekli ve simetrik bir çekirdeğin, sonsuz boyuta dolaylı bir şekilde haritalanan girdi vektörlerinin bir iç çarpımına ayrışabileceğini gösterir.

TEOREM 2.3. (MERCER TEOREMİ) *Eşitlik (2.26)'daki gibi bir öz-ayrışıma sahip pozitif, öz-eşlenik ve kompakt T_K integral işleci tanımlayan, $X \times X$ gibi kompakt bir tanım kümesi üzerinde bulunan $K \in L^2(X \times X)$ gibi pozitif (Tanım 2.15), simetrik (Eşitlik (2.6)), ve sürekli (Tanım 2.16) çekirdek fonksiyonları için aşağıdaki beş koşulun sağlanması gerekmektedir.*

1. $\{\upsilon_1, \upsilon_2, ...\} \in l^1$ *: T_K integral işleci kompakt olduğu için, özdeğerler dizisi mutlak yakınsak olmalıdır.*
2. $\upsilon_i > 0, \forall i$ *: Bütün özdeğerler tam pozitif olmalıdır.*
3. $\varsigma_i \in L^\infty(X)$*: Tüm özfonksiyonların $\varsigma_i : X \to \mathbb{R}$ her biri sınırlandırılmış olmalıdır.*
4. $\sup_i \|\varsigma_i\|_{L^\infty} < \infty$*: Tüm özfonksiyonların oluşturduğu küme de sınırlı olmalıdır.*
5. $\forall \vec{s}, \vec{x} \in X : K(\vec{s},\vec{x}) = \sum_{i=1}^{\infty} \upsilon_i \varsigma_i(\vec{s}) \varsigma_i(\vec{x}) = \langle \phi(\vec{s}), \phi(\vec{x}) \rangle_{L^2}, i, j = 1, 2, ..., n$ *: Burada, (5) tüm $(\vec{x},\vec{y}) \in X \times X$ noktaları için mutlak ve düzgün yakınsak olmalıdır* [22].

$$\forall \vec{s}, \vec{x} \in X : \lim_{n\to\infty} \sup_{\vec{s},\vec{x}} \left| K(\vec{s},\vec{x}) - \sum_{i=1}^{\infty} \upsilon_i \varsigma_i(\vec{s}) \varsigma_i(\vec{x}) \right| = 0, i, j = 1, 2, ..., n$$

İSPAT T_K integral işleci kompakt olduğu için, ortonormal bir tabanın varlığını garanti eden Spektral Ayrışım Teoremi özdeğerler ve özfonksiyonlar cinsinden yazılabilir.

$$(T_K \varsigma_i)(\vec{s}) = \int_X K(\vec{s},\vec{t}) \varsigma_i(\vec{t}) \, d\mu(\vec{t}) = \upsilon_i \varsigma_i(\vec{s}) \tag{2.32}$$

Özfonksiyonlar, $L^2(X)$ için ortonormal bir taban kurduğundan,

$$\|\varsigma_i\|_{L^2} = \sqrt{\langle \varsigma_i, \varsigma_i \rangle} = \int_X \varsigma_i(\vec{x})^2 \, d\mu(\vec{x}) = 1 \tag{2.33}$$

eşitliği yazılabilir. Mercer Teoremi'nin birinci koşulu, beşinci koşuldan ve $K \in L^{\infty}(X \times X)$ gibi bir çekirdek fonksiyonunun sınırlılık (kompakt bir tanım kümesi üzerinde süreklilik ile aynı anlama gelir) özelliğinden gelir. $\vec{s} = \vec{x}$ olarak kabul edilir ve teoremin beşinci koşulundaki çekirdek genişlemesinin her iki tarafının integrali alınırsa,

$$\begin{aligned}\int_X \left(\sum_{i=1}^{\infty} \upsilon_i \varsigma_i(\vec{x}) \varsigma_i(\vec{x}) \right) d\mu(\vec{x}) &= \sum_{i=1}^{\infty} \upsilon_i \underbrace{\int_X \varsigma_i(\vec{x})^2 d\mu(\vec{x})}_{1} = \sum_{i=1}^{\infty} \upsilon_i \\ &= \int_X K(\vec{x},\vec{x}) d\mu(\vec{x}) < \infty\end{aligned} \tag{2.34}$$

elde edilir. $f > 0 \Rightarrow T_K f > 0$ 'dan hareketle, Mercer Teoremi'nin ikinci koşulu, T_K integral işlecinin pozitif tanımlı olma özelliğinden gelmektedir.

Bu teoremin üçüncü ve dördüncü koşulları ise, kompakt bir tanım kümesi üzerinde bulunan çekirdeğin ve özfonksiyonların süreklilik özelliğinden gelmektedir. Eğer $\upsilon_i \neq 0$ ise, ilgili özfonksiyonlar X üzerinde süreklidir.

Eşitlik (2.32)'den, $\varsigma_i(\vec{s}) = \frac{1}{\upsilon_i}(T_K \varsigma_i)(\vec{s}) = \frac{1}{\upsilon_i}\left(\int_X K(\vec{s},\vec{t}) \varsigma_i(\vec{t}) d\mu(\vec{t}) \right)$ olmak üzere,

$$\begin{aligned}&\forall \varepsilon > 0, \exists \delta > 0 : |\vec{x} - \vec{y}| < \delta \Rightarrow \\ &\qquad |\varsigma_i(\vec{x}) - \varsigma_i(\vec{y})| = \frac{1}{|\upsilon_i|} \left| \int_X \left(K(\vec{x},\vec{s}) - K(\vec{y},\vec{s}) \right) \varsigma_i(\vec{s}) d\mu(\vec{s}) \right| \\ &\qquad \leq \frac{1}{|\upsilon_i|} \int_X |K(\vec{x},\vec{s}) - K(\vec{y},\vec{s})| |\varsigma_i(\vec{s})| d\mu(\vec{s}) \\ &\qquad \leq \frac{\sup_i \|\varsigma_i\|_{L^{\infty}}}{|\upsilon_i|} \int_X |K(\vec{x},\vec{s}) - K(\vec{y},\vec{s})| d\mu(\vec{s}) \\ &\qquad \leq \varepsilon\end{aligned} \tag{2.35}$$

'dir. Burada son eşitsizlik, K çekirdek fonksiyonunun süreklilik özelliğinden gelmektedir, böylelikle, $|K(\vec{x},\vec{s}) - K(\vec{y},\vec{s})|$ farkı keyfi olarak küçük seçilebilir.

2.12. Doğuran Çekirdekli Hilbert Uzayları

Bir DÇHU, DVM algoritması için çalışan bir hipotez (fonksiyon) uzayıdır. Gözlem uzayında bulunan elemanlar, var olan ayırma veya bağlanım probleminin tanımlanması ve çözümü için gerekli olan yapıyı sağlamak üzere, DÇHU'ya haritalanırlar. Bu DÇHU'da herhangi bir gözlem bir özelliğe dönüştürülebilir. Böylece gözlem uzayından alınan herhangi bir küme için genel temsili bir uzay oluşturulur. Özelliklerin aldıkları dolaysız form, dolaylı bir şekilde bir iç çarpım olarak gösterilebilecek iki gözlem arasındaki çekirdekleştirilmiş uzaklık ölçüsüdür. Aslında, DÇHU, kısıtlandırılmış bir Hilbert uzayını, ilgili pozitif tanımlı bir çekirdek fonksiyonu ile birleştirir [23].

TANIM 2.17. (DOĞURAN ÇEKİRDEKLİ HİLBERT UZAYI) $\mathbb{R}^X$ *üzerinde noktasal tanımlı bir Hilbert uzayında, her değerlendirme fonksiyoneli* $\varepsilon_t[f]: H(X) \to \mathbb{R}$ *sürekli ise, bu Hilbert uzayı, DÇHU'dur.*

Hipotez uzayında bulunan fonksiyonların yakınsak serileri noktasal yakınsak olmalıdır. Yani,

$$\|f_n - f\|_H \to 0 \Rightarrow \lim_{n\to\infty} f_n(\vec{x}) - f(\vec{x}) = 0, \quad \forall \vec{x} \in X \tag{2.36}$$

'dir. Bu kural her Hilbert uzayı için sağlanmasa da, tüm DÇHU'lar için geçerlidir. Eşitlik (2.12)'de hipotez uzayındaki noktasal yakınsamanın değerlendirme fonksiyonunun sürekliliğini sağladığı gösterilmişti. Sürekli bir fonksiyonel aynı zamanda sınırlı olduğu için, Riesz Gösterim Teoremi'nden sabit bir $\vec{x} \in X$ noktasındaki her fonksiyon değerlendirmesi, Eşitlik (2.17)'yi sağlayan $r_{\varepsilon_{\vec{x}}}$ gibi sabit bir gösterge fonksiyonuna sahiptir.

Bunlara ek olarak, bir DÇHU'daki norm yakınmasının noktasal olduğu kolaylıkla görülebilir [24].

$$\|f_n - f\|_H \to 0 \Rightarrow \lim_{n\to\infty} f_n(\vec{x}) = \lim_{n\to\infty} \varepsilon_{\vec{x}}[f_n] = \lim_{n\to\infty} \varepsilon_{\vec{x}}[f] = f(\vec{x}), \quad \forall \vec{x} \in X \tag{2.37}$$

Eşitlik (2.37)'deki ikinci eşitlik, değerlendirme fonksiyonelinin süreklilik özelliğinden ve f_n'nin f'e norm halinde yakınsadığı varsayımından gelir.

TANIM 2.18. (DOĞURAN ÇEKİRDEK) *Tüm* $\vec{x} \in X$ *noktaları için aşağıdaki koşulları sağlayan,* $L^2(X \times X)$ *gibi bir Hilbert uzayına ait* K *çekirdek fonksiyonu doğurandır.*

1. $\forall \vec{x} \in X, K_{\vec{x}} = K(\vec{x}, \cdot) \in H$: $\vec{x} \in X$ *gibi bir noktada sabit olan bir çekirdek, bir Hilbert uzayı üzerinde fonksiyondur.*
2. $\forall f \in H$ *ve* $\forall \vec{x} \in X$ *olmak üzere doğurganlık özelliği şu şekilde sağlanır:*

$$\langle f, K_{\vec{x}} \rangle = f(\vec{x})$$

ve özellikle, herhangi bir $\vec{s}, \vec{x} \in X$ *için,*

$$\langle K_{\vec{s}}, K_{\vec{x}} \rangle = \langle K(\vec{s}, \cdot), K(\vec{x}, \cdot) \rangle = K(\vec{s}, \vec{x})$$

elde edilir.

Doğuran çekirdek tanımı kafalarda birkaç soru yaratabilir. Bu çekirdeğin, bir DÇHU ile ilişkisi nedir? Bu çekirdek var mıdır? Varsa, tek midir? Bu çekirdek hangi özelliklere sahiptir? Bu soruların cevabı aşağıdaki bölümde verilecektir.

Doğuran çekirdeğin tek olduğu, Tanım (2.18)'den hemen görülebilir. Eğer bir doğuran çekirdek varsa, bu çekirdek tektir.

Doğuran bir çekirdeğin varlığını göstermek için ise, bir DÇHU'da değerlendirme fonksiyonelinin kendisinin de bir iç çarpım olarak yazılabildiğini söyleyen Riesz Gösterim Teoremi kullanılır. Bu durumda tanım gereği, $\varepsilon_{\vec{x}}[f]$ değerlendirme fonksiyoneli için $K_{\vec{x}}$ tek bir göstergedir.

$$\forall \vec{x} \in X, \exists K_{\vec{x}} \in H_K : f(\vec{x}) = \varepsilon_{\vec{x}}[f] = \langle f, K_{\vec{x}} \rangle_{H_K} \leq \|K_{\vec{x}}\|_{H_K} \|f\|_{H_K}, \quad \forall f \in H \qquad (2.38)$$

Dikkat edilirse, burada, gösterge fonksiyonu $r_{\varepsilon_{\vec{x}}}$, $K_{\vec{x}} = K(\vec{x}, \cdot)$ gibi bir çekirdek fonksiyonu formuna sahiptir. Bu durumda, Doğuran Çekirdekli bir Hilbert Uzayı'ndaki her fonksiyon, fonksiyonun kendisi ve ilgili doğuran çekirdek fonksiyonunun bir iç çarpımı olarak yazılabileceği söylenebilir.

TEOREM 2.4. (MOORE-ARONSZAJN TEOREMİ) *Buraya kadar anlatılan kısımda, verilen bir DÇHU'da nasıl pozitif tanımlı doğuran bir çekirdek fonksiyonu tanımladığı gösterildi. Moore-Aronszajn Teoremi ise bu durumun tam tersini tanıtlar. Yani* $X \times X$ *üzerinde pozitif tanımlı simetrik her doğuran çekirdek fonksiyonu* $K(\cdot,\cdot)$ *için ilgili tek bir DÇHU vardır. Kısacası, DÇHU'lar kümesi ile bütün pozitif çekirdek fonksiyonları kümesi arasında tam eşleşme (bijection) vardır.*

İSPAT H_K ile gösterilen bir DÇHU'da, Riesz Gösterim Teorisi gereği, bu uzay üstünde tanımlı tüm değerlendirme fonksiyonelleri için H_K 'da $K_{\vec{x}}$ gibi tek bir gösterge fonksiyonu mevcuttur. Bu durumda (pozitif ve tek) doğuran çekirdek aşağıdaki gibi gösterilir.

$$K(\vec{x},\vec{s}) = \langle K_{\vec{x}}, K_{\vec{s}} \rangle_{H_K}, \quad \forall \vec{x},\vec{s} \in X \tag{2.39}$$

Tam tersi olarak, verilen pozitif bir K çekirdeği için, her $\vec{x}_i \in X$ için $\{K_{\vec{x}_1}, K_{\vec{x}_2}, K_{\vec{x}_3} ...\}$ gibi bir fonksiyonlar kümesi tanımlanabilir. Bu durumda, DÇHU'daki elemanlar, bu küme tarafından gerilen uzayın tümleyeninde bulunan noktasal tanımlı fonksiyonlar olarak tanımlanabilir.

$$H_K = \left\{ f \in \mathbb{R}^X \mid f = \sum_{i=1}^{\infty} \alpha_i K_{\vec{x}_i}, \|f\|_{H_K} < \infty, \forall \alpha_i \in \mathbb{R} \right\} \tag{2.40}$$

Bu uzayda doğurganlık özelliği sağlanır.

$$\langle K_{\vec{s}}, f \rangle_{H_K} = \left\langle K_{\vec{s}}, \sum_j \beta_j K_{\vec{t}_j} \right\rangle_{H_K} = \sum_j \beta_j \langle K_{\vec{s}}, K_{\vec{t}_j} \rangle_{H_K} = \sum_j \beta_j K(\vec{s},\vec{t}_j) = f(\vec{s}) \tag{2.41}$$

O halde, $K_{\vec{s}}$, $\varepsilon_{\vec{s}}[\cdot]$ değerlendirme fonksiyonelinin bir göstergesidir. Bu uzaydaki değerlendirme fonksiyonelleri sınırlı ve süreklidir.

$$\begin{aligned} |\varepsilon_{\vec{x}}[f]| &= |f(\vec{x})| = |\langle f, K_{\vec{x}} \rangle| \\ &\le \|K_{\vec{x}}\|_H \|f\|_H \\ &= \langle K_{\vec{x}}, K_{\vec{x}} \rangle_H^{1/2} \|f\|_H \\ &= K(\vec{x},\vec{x})^{1/2} \|f\|_H \end{aligned} \tag{2.42}$$

Burada, ilk satırdaki eşitlik, çekirdeğin doğurganlık özelliğinden, ikinci satırdaki eşitsizlik Cauchy-Schwarz Eşitsizliği'nden, üçüncü ve dördüncü satırdaki eşitlikler ise Hilbert uzayında iyi tanımlanmış iç çarpım özelliğinden gelmektedir. Çekirdek pozitif olduğunda bu uzaydaki tüm normlar aşağıdaki gibi tanımlanan simetrik ve doğrusal olan bir iç çarpımdan oluşur. $f(\cdot)=\sum_{i=1}^{\infty}\alpha_i K_{\vec{x}_i}$ ve $g(\cdot)=\sum_{j=1}^{\infty}\beta_j K_{\vec{x}_j}$ olmak üzere,

$$\begin{aligned}\langle f,g\rangle_{H_K} &= \left\langle \sum_{i=1}^{\infty}\alpha_i K_{\vec{x}_i}, \sum_{j=1}^{\infty}\beta_j K_{\vec{x}_j}\right\rangle_{H_K} \\ &= \sum_i\sum_j \alpha_i\beta_j \left\langle K_{\vec{x}_i}, K_{\vec{x}_j}\right\rangle \\ &= \sum_i\sum_j \alpha_i\beta_j K\left(\vec{x}_i,\vec{x}_j\right) \\ &= \sum_i \alpha_i g\left(\vec{x}_i\right) = \sum_j \beta_j f\left(\vec{x}_j\right)\end{aligned} \tag{2.43}$$

elde edilir. DÇHU'nun bir Hilbert uzayı olduğunu göstermek için bu uzayın tam olduğunu göstermek gerekmektedir. K çekirdek fonksiyonu tarafından gerilen bu uzayda bulunan tüm Cauchy dizilerinin limit fonksiyonları, eğer uzayın içinde bulunmuyorlarsa, bu uzaya eklenmelidir. Bu limit fonksiyonları aynı zamanda noktasal olarak iyi tanımlanmış olmalıdır. Ancak bölümün başında da bahsedildiği gibi, bir DÇHU'da norm yakınsaması, noktasal yakınsama (ve özellikle Cauchy yakınsaması) anlamına gelmektedir, bu durumda tüm limit noktaları iyi tanımlanmıştır.

Alternatif olarak, DÇHU, Mercer parçalanması (Mercer Teoremi'nin beşinci koşulu) kullanılarak da inşa edilebilir. İntegral işlecinin sıfırdan farklı özdeğerlere karşılık gelen özfonksiyonlar tarafından gerilen uzay aşağıdaki gibi tanımlanabilir.

$$H_K=\left\{ f\in\mathbb{R}^X \mid f=\sum_{i=1}^{\infty}\alpha_i\varsigma_i, \|f\|_{H_K}<\infty, \forall\alpha_i\in\mathbb{R}, \varsigma_i\in L^{\infty}(X)\right\} \tag{2.44}$$

Böylelikle, H_K uzayının boyutu, integral işlecinin sıfırdan farklı özdeğerlerin sayısına eşittir. Daha sonra, bu uzay üstündeki norm,

$$\|f\|_{H_K}^2=\langle f,f\rangle_{H_K}=\left\langle \sum_{i=1}^{\infty}\alpha_i\varsigma_i, \sum_{i=1}^{\infty}\alpha_i\varsigma_i\right\rangle_{H_K}=\sum_{i=1}^{\infty}\frac{\alpha_i^2}{\upsilon_i} \tag{2.45}$$

şeklinde yazılabilir Benzer şekilde, uzayda tanımlı iki fonksiyonun bir iç çarpımı,

$$\langle f,g\rangle_{H_K}=\left\langle \sum_{i=1}^{\infty}\alpha_i\varsigma_i,\sum_{i=1}^{\infty}\beta_i\varsigma_i\right\rangle_{H_K}=\sum_{i=1}^{\infty}\frac{\alpha_i\beta_i}{\upsilon_i} \tag{2.46}$$

şeklindedir. Mercer Teoremi bir kez daha kullanılarak, $K_{\vec{x}}$ 'in, $\varepsilon_{\vec{x}}[\cdot]$ değerlendirme fonksiyonelinin bir göstergesi olduğu ve bir DÇHU yarattığı gösterilebilir.

$$\begin{aligned}\left\langle K_{\vec{x}}(\cdot),f(\cdot)\right\rangle_{H_K}&=\left\langle K_{\vec{x}}(\cdot),\sum_{i=1}^{\infty}\alpha_i\varsigma_i(\cdot)\right\rangle_{H_K}\\&=\left\langle \sum_{i=1}^{\infty}\upsilon_i\varsigma_i(\cdot)\varsigma_i(\vec{x}),\sum_{i=1}^{\infty}\alpha_i\varsigma_i(\cdot)\right\rangle_{H_K}\\&=\sum_{i=1}^{\infty}\frac{\alpha_i\upsilon_i\varsigma_i(\vec{x})}{\upsilon_i}=\sum_{i=1}^{\infty}\alpha_i\varsigma_i(\vec{x})=f(\vec{x})\end{aligned} \tag{2.47}$$

λ , düzenleme veya düzgünlük parametresi, $\|f\|_{H_K}^2$ cezalandırma fonksiyoneli ve $\ell(f,\{\vec{x}_i,y_i\})$ kayıp fonksiyonu olmak üzere hipotez uzayındaki bütün fonksiyonlar üzerinden, aşağıdaki gibi verilen *düzenlenmiş risk fonksiyonelini (regularized risk function)* minimize etmek yerine,

$$f^*=\min_{f\in H_K}\left\{\sum_{i=1}^{n}\ell(f,\{\vec{x}_i,y_i\})+\lambda\|f\|_{H_K}^2\right\} \tag{2.48}$$

$\{\alpha_1,\alpha_2,\alpha_3...\}$ ile verilen açılım katsayılarının bütün dizileri üzerindeki aşağıdaki fonksiyonel minimize edilmeye çalışılır.

$$f^*=\min_{\{\alpha_1,\alpha_2,\alpha_3...\}}\left\{\sum_{i=1}^{n}\ell\left(\sum_{j=1}^{\infty}\alpha_j\varsigma_j(\cdot),\{\vec{x}_i,y_i\}\right)+\lambda\sum_{j}\frac{\alpha_j^1}{\upsilon_j}\right\} \tag{2.49}$$

Bu fonksiyonel Eşitlik (2.40) ve Eşitlik (2.44) kullanılarak elde edilir [25]. Açılım katsayılarının sayısı, sıfırdan farklı özdeğerlerin sayısına eşittir. Bu sayı aynı zamanda Eşitlik (2.44) ile elde edilen DÇHU'nun boyutunu verir. Bu değer sonsuz olduğunda, yukarıdaki optimizasyon probleminin uygun çözümü elde edilemeyebilir.

Moore-Aronszajn Teoremi her pozitif tanımlı fonksiyonunun bir doğuran çekirdek olduğunu söyler. Önceki bölümlerde her doğuran çekirdeğin bir çekirdek olduğu ve her çekirdeğin pozitif tanımlı olduğu söylenmişti. Yani, her üç ifade de tamamıyla aynıdır. Ek olarak $\mathbb{R}^{X\times X}$ ile gösterilen, $X\times X$ üzerinde tanımlı bütün pozitif tanımlı fonksiyonların

oluşturduğu küme ile $\mathbb{R}^X$ 'in alt uzaylarından oluşan ve H_K ile gösterilen bütün DÇHU'ların oluşturduğu küme arasında tam eşleme (bijection) olduğu gösterildi.

2.13. Uzayın ve Fonksiyonun Düzenlenmesi

Bir önceki bölümde hipotez uzaylarının DÇHU ailesi tanıtıldı. Bu bölümde bu uzay üstüne konması gereken kısıtlar ve bu kısıtların neden gerekli olduğundan bahsedilecektir. Öğrenme algoritmasının seçtiği hipotezin üç temel kıstası sağlaması gerekmektedir. Öğrenme algoritmasının genelleştirilebilme özelliğinin yanı sıra, "iyi" bir algoritma aynı zamanda *durağan (stable)* olmalıdır. Elde edilecek f hipotez/tahmin fonksiyonu sürekli bazda eğitim veri kümesine bağlı olmalıdır. Yani n sonsuza giderken, eğitim verilerinden birinin değiştirilmesinin, çözüm üzerinde oldukça az etkiye sahip olması gerekir. Durağanlık herhangi bir matematiksel problem için olduğu gibi, bir öğrenme algoritması için de gereklidir.

TANIM 2.19. (İYİ KONUMLANMIŞ OPTİMİZASYON (WELL-POSED OPTIMIZATION)) *ψ optimizasyonunun $f^* : X \to Y$ gibi çözümü aşağıdaki kriterlere sahipse, bu optimizasyon iyi konumlanmıştır denir.*

1. *Vardır (existing): Eğer hipotez uzayı çok küçükse, bu problemin çözümü olmayabilir.*

$$\exists \hat{f}^* \in H : f^* = \arg\inf_{f \in H} \psi$$

2. *Tektir (unique): Eğer hipotez uzayı çok büyük veya eğitim kümesi çok küçükse, bu problemin çözümü tek olmayabilir.*

$$\exists \hat{f}_1^*, \hat{f}_2^* \in H : \hat{f}_1^*, \hat{f}_2^* = \arg\inf_{f \in H} \psi \Rightarrow \hat{f}_1^* = \hat{f}_2^*$$

3. *Durağandır (stable): f^* sürekli bazda eğitim veri kümesine bağlı olmalıdır. Böylelikle eğitim kümesindeki küçük bozulmalar elde edilecek çözümü etkilemez. Özellikle, eğitim örneklerinin sayısı artarsa.*

Kullanılan öğrenme algoritması ile elde edilen tahmin fonksiyonu genelleştirilebilir ve iyi konumlanmış olmalıdır. Tanım (2.19)'de verilen üçüncü kıstas bu nedenle özellikle önemlidir. Çünkü bu kıstas bir hipotez uzayının genelleştirilebilme özelliğiyle yakından ilgilidir. Durağan bir dönüşümün eğitim verisine aşırı uyum sağlaması olasılığı oldukça düşüktür, yani deneysel riskin, minimize edilmeye çalışılan beklenen riskten çok küçük çıkması beklenmez. Eğer bir problem kötü konumlanmış ise, bu problemi iyi konumlandırmak için literatürde önerilen birkaç yöntem vardır. Düzenleme teknikleri denilen bu yöntemlerin altındaki temel amaç aşırı uyumu engellemek için hipotez uzayı üzerine kısıtlar koymaktır [26].

DRM prensibine göre, H uzayının kompakt ve ℓ kayıp fonksiyonunun (ve böylelikle deneysel risk $R_{deney}[f]$'nin) sürekli olduğu varsayımı altında, optimizasyon probleminin çözümünün varlığını garanti eder. Genelde bu koşulların hiç biri sağlanmaz. Ancak, DRM, çözümün tekliğini (minimum deneysel riske sahip tüm fonksiyonlar aynı sınıftadır ancak bu sınıftaki sadece bir fonksiyon eğitim veri üzerinde iyi bir şekilde genelleştirilebilir) ve durağanlığını (eğitim verisindeki tek bir örneğin silinmesi, bir önceki tahmin fonksiyonundan tamamen farklı yeni bir tahmin fonksiyonunun elde edilmesine neden olur) garanti etmez. Bu nedenlerden dolayı DRM yöntemi, yapılacak öğrenme işlemini kötü konumlandırır.

Bu nedenle, önsel bir bilgi kullanılarak, en küçük deneysel riske sahip fonksiyonlar sınıfı içinden elde edilecek hangi çözümün, tahmin için en uygun olduğunun araştırılması gerekmektedir. Bu araştırma, mesela, hipotez uzayının kapasitesini kısıtlayarak yapılabilir. Bu kısıtlama için çoğunlukla iki düzenleme yöntemi kullanılır, böylelikle çözümün tekliği ve durağanlığı korunmuş olur. Hipotez uzayının nasıl kısıtlanacağının sorusunun cevabı Occam'ın Usturası teorisinde gizlidir. Occam'ın Usturası teorisi temel olarak "her şeyin birbirine eşit olduğu bir ortamda, en basit açıklama doğruya en yatkın olandır" felsefesi üzerinde şekillenir. Yani, tüm diğer değişkenler sabit kabul edildiğinde, en basit çözüm genellikle en iyi olandır. Özetle, düzenleme yöntemleri, uzayda bulunan ve genelleştirilebilmesi mümkün olmayan karmaşık fonksiyonları elimine ederek hipotez uzayının kapasitesini kısıtlar ve öğrenme işlemi için iyi konumlanmış çözümü elde etmeye çalışır, böylelikle bu öğrenme problemi için tek ve durağan bir çözüm elde eder.

Bu bölümde anlatılacak iki düzenleme yöntemi vardır. Bunlardan ilki Ivanov Düzenlemesi'dir. Bu yöntemde doğrudan olarak hipotez uzayının kapasitesi kısıtlandırılır. İkinci yöntem ise Tikhonov Düzenlemesi'dir. Bu yöntemde ise hipotez uzayının

kapasitesini düzenleyen bir parametre dolaysız bir şekilde optimize edilir. Her iki yöntemde birbirinin aynısıdır ve hipotez uzayını düzenlemek için bir fonksiyonun "düzgünlük" ölçüsünü kullanır. İlgili çekirdek fonksiyonu benzerlik ölçütü olarak kullanılırken, burada norm fonksiyoneli düzgünlük ölçütü olarak kullanılır.

TANIM 2.20. (LIPSCHITZ SÜREKLİLİĞİ (LIPSCHITZ CONTUNITY)) $f : X \to Y$ *haritalaması aşağıdaki koşulu sağladığı takdirde Lipschitz süreklidir denir.*

$$\left| f\left(\vec{x}_1\right) - f\left(\vec{x}_2\right) \right| \leq M \left| \vec{x}_1 - \vec{x}_2 \right|$$

Tüm $\vec{x}_1, \vec{x}_2 \in X$ *için yukarıdaki eşitsizliği sağlayan en küçük* $M \geq 0$ *değerine, fonksiyonun Lipschitz sabiti denir. Her Lipschitz sürekli haritalama, aynı zamanda düzgün süreklidir. Bu süreklilik türü, basit süreklilikten daha güçlüdür. Lipschitz Süreklilik kavramı, metrik uzaylarda iyi tanımlanmıştır.*

DÇHU'daki tüm fonksiyonlar aynı zamanda Lipschitz süreklidir. Tanım kümesinden $\vec{x}_1, \vec{x}_2 \in X$ gibi iki nokta alındığında, Riesz Gösterim Teoremi'nden,

$$\begin{aligned} \left| f\left(\vec{x}_1\right) - f\left(\vec{x}_2\right) \right| &= \left| \left\langle f, K_{\vec{x}_1} \right\rangle_{H_K} - \left\langle f, K_{\vec{x}_2} \right\rangle_{H_K} \right| \\ &= \left| \left\langle f, K_{\vec{x}_1} - K_{\vec{x}_2} \right\rangle_{H_K} \right| \\ &\leq \| f \|_{H_K} \left\| K_{\vec{x}_1} - K_{\vec{x}_2} \right\|_{H_K} \\ &= \| f \|_{H_K} \, d(K_{\vec{x}_1}, K_{\vec{x}_2}) \end{aligned} \tag{2.50}$$

elde edilir. Burada, Lipschitz sabiti, f fonksiyonunun normu ile verilmiştir. Ayrıca tanım kümesindeki iki veri noktası arasındaki uzaklık, bu noktaların çekirdekleştirilmiş pozisyonları arasındaki uzaklığa eşittir. Lipschitz sabiti (bu durumda fonksiyonunun normudur) küçüldüğünde, tanım kümesindeki benzer noktalar için (burada benzerlik çekirdek fonksiyonu ile ifade edilir), görüntü uzayında bulunan fonksiyonlardaki değişim daha azdır. Bu tanım Eşitlik (2.48) tanımlanan düzenlenmiş risk fonksiyonelinde ve aşağıdaki bölümde tanıtılacak olan Ivanov Düzenlenmesi'nde neden norm kullanıldığını doğrular niteliktedir.

2.13.1. Ivanov Düzenlemesi

Hipotez uzayını doğrudan kısıtlamak için en çok kullanılan düzenleme yöntemlerinden ilki Ivanov Düzenlemesi'dir [27]. Bu düzenleme yöntemine göre, $f \in H_\tau$ hipotez uzayındaki tüm fonksiyonlar, H_K DÇHU'nun τ -sınırlı alt kümesinde bulunurlar.

$$\hat{f}^* = \arg\inf_{f \in H_K} R_{deney}[f], \quad kısıt: \|H\|_{H_K} \le \tau \tag{2.51}$$

İki farklı hipotez uzayında bulunan fonksiyonları düşünelim. Bu uzayın biri, diğerine göre, daha az karmaşık olsun, yani bu uzayda bulunan fonksiyonlar daha düzgündür (daha az karmaşıktır). Bu durumda,

$$H_{\tau_i} = \left\{ f : f \in H_K \; ve \, \| f \|_{H_K}^2 \le \tau_i \right\}, \quad i \in \{1, 2\}, \tau_1 \ll \tau_2$$

'dır. Eğitim kümesindeki küçük dalgalanmalar, daha karmaşık yapıya sahip H_{τ_2} hipotez uzayında bulunan tahmin fonksiyonlarının daha fazla kararsız kalmasına neden olur. Ancak, daha düzgün fonksiyonlar sınıfına sahip hipotez uzayı H_{τ_1}, H_{τ_2} 'e göre daha durağandır. Ancak, bu düzenleme yönteminde τ değerinin seçimi için özel bir yöntem bulunmadığından, genellikle başka bir düzenleme yöntemi olan Tikhonov Düzenlemesi kullanılır.

2.13.2. Tikhonov Düzenlemesi

Tikhonov Düzenlemesi 1984 yılında Andrey Tychonoff tarafından geliştirilmiştir [28]. Bu düzenleme yönteminin, Ivanov Düzenlemesi'nden farkı, hipotez uzayını doğrudan bir sabit ile sınırlandırmak yerine, optimizasyon probleminde bulunan amaç fonksiyonundaki hipotez uzayının karmaşıklığını ve kararsızlığını cezalandırmasıdır.

$$\hat{f}^* = \min_{f \in H_K, \lambda} \left\{ R_{deney}[f] + \lambda \| f \|_{H_K}^2 \right\} \tag{2.52}$$

Burada, λ, düzenleme veya düzgünlük parametresidir. Bu parametre, çözümün optimal genelleştirilebilme performansını, durağanlığını ve tekliğini sağlamak için iyi bir şekilde optimize edilmelidir [29]. Aşağıdaki teoremde görüleceği üzere, hipotez uzayı sonsuz boyutlu bir Hilbert fonksiyon uzayı olsa bile, Tikhonov optimizasyonun çözümü sonlu taban genişlemesi formuna sahiptir.

TEOREM 2.4. (GÖSTERGE TEOREMİ (REPRESENTER THEOREM)) *Bir kayıp fonksiyonunun toplamını ve düzenleme terimini optimize eden Tikhonov düzenleme yönteminin amaç fonksiyonunun aşağıdaki gibi verildiği önceki bölümlerde bahsedilmişti.*

$$f^* = \min_{f \in H_K} \left\{ \sum_{i=1}^{n} \ell(f, \{\vec{x}_i, y_i\}) + \lambda \| f \|_{H_K}^2 \right\}$$

Burada, ℓ *, noktasal tanımlı kayıp fonksiyonu ve* λ *monoton artan düzenleme terimi olmak üzere, bu amaç fonksiyonunun çözümü vardır ve bu çözüm, eğitim verisi üzerinde tanımlı fonksiyonların bir sonlu kümesinin doğrusal bileşeni olarak yazılabilir.*

$$f^* = \sum_{j=1}^{n} \alpha_j K_{\vec{x}_j}$$

Burada, $K_{\vec{x}_j}$ *, tüm* $f \in H$ *için* $\varepsilon_{\vec{x}_j}[f] = f\left(\vec{x}_j\right)$ *sınırlı değerlendirme fonksiyonelinin bir göstergesidir* [30].

3. İSTATİSTİKSEL ÖĞRENME TEOREMİ

İstatistiksel Öğrenme Teoremi günümüzün birçok makine öğrenmesi algoritmalarına teorik alt yapı oluşturur ve yapay zekâ konusunun en iyi şekilde geliştirilmiş dallarından biridir. 1960'larda Rusya'da geliştirilmeye başlayan teorem DVM'nin tanınmasıyla birlikte büyük bir popülariteye kavuşmuştur. Teorem yeni geliştirilen öğrenme algoritmalarına temel teşkil etmekle birlikte, deneysel veriden gerekli bilgileri çekmek ve geçerli sonuçları elde etmek için de kullanılmaktadır. Bu bilgileri elde etmek ve sonuçları yorumlayabilmek için ise bir öğrenme algoritmasından elde edilecek en iyi hipotez fonksiyonunun bulunmasıdır.

Bir öğrenme problemi için optimal tahmin fonksiyonunu ararken uygulanması gereken en uygun yaklaşım, $S = \{\vec{x}_i, y_i\}_{i=1}^{n}$ eğitim kümesi için kabul edilebilir tahmin fonksiyonlarının doğruluklarını ve ne kadar "iyi" olduklarını ölçen bir ölçüt üzerinde optimizasyon problemi tanımlamaktır. Hipotez uzayında bulunan her hipotez/tahmin fonksiyonuna $f \in H$ bu tür bir ölçüt veya daha doğru ifadeyle $\ell(f, \{\vec{x}, y\})$ şeklinde bir kayıp fonksiyonu uygulandığında, aşağıdaki gibi gösterilen ve kayıp fonksiyonları kümesi olarak isimlendirilen fonksiyonlar uzayı elde edilir.

$$L(H, \cdot) = \left\{ \ell(f, \cdot) : f \in H \right\}$$

Yukarıdaki, bu uzaydan alınan bir kayıp fonksiyonu yardımıyla, elde edilen hipotez fonksiyonu tüm gözlem uzayı üzerinden değerlendirilir. Ancak, eğitim verisindeki tüm gözlemler $P(\vec{x}, y)$ gibi bir dağılımdan geldiği için, kullanılacak kayıp fonksiyonu bu dağılım fonksiyonuyla birlikte hesaplanmalıdır.

TANIM 3.1. (BEKLENEN RİSK) *Beklenen risk, verinin türetildiği dağılıma göre, $X \times Y$ gözlem uzayı üzerinde sabit bir fonksiyonun ürettiği ortalama kayıp veya hatadır ve Eşitlik (3.1)'deki gibi tanımlanır.*

$$R_X(f) = E[\ell(f, \{\vec{x}, y\})] = \int_X \int_Y \ell(f, \{\vec{x}, y\}) dP(\vec{x}, y) = \int_X \int_Y \ell(f, \{\vec{x}, y\}) P(\vec{x}, y) d\vec{x} dy \quad (3.1)$$

Bu durumda kullanılan öğrenme algoritması, sabit bir kayıp fonksiyonu ℓ için, f^* tahmin fonksiyonunu, tüm ölçülebilir fonksiyonlar üzerinden Eşitlik (3.1)'deki beklenen riski minimize ederek bulmaya çalışır.

$$f^* = \arg\inf_{f \in H} R_X(f) \tag{3.2}$$

Ancak burada eğitim veri kümesinin çekildiği olasılık dağılımı $P(\vec{x}, y)$ bilinmediğinden ve tüm girdi uzayındaki çıktılar elde edilemediğinden, f^* tahmin fonksiyonunu bulabilmek teknik olarak neredeyse imkânsızdır çünkü bu beklenen riski hesaplamak mümkün değildir. İşte tam bu noktada İstatistiksel Öğrenme Teoremi devreye girer ve bu problemin analizinin yapılması ve çözümlerinin elde edilmesi için bu çözümlerin başarılı bir şekilde işe yaradıklarının garantisini veren iskelet bir yapı oluşturur.

3.1. Deneysel Risk Minimizasyonu

Beklenen riski elde etmek mümkün olmadığı için, en uygun yol bu riske yakınsamaktır. Tümevarım prensibi denilen bu yöntem ile $P(\vec{x}, y) = P(x)\ P(\vec{y}$ bilgisi kullanılarak eğitim verisi üzerinden bu beklenen risk tahmin edilmeye çalışılır.

$$R_{deney}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f, \{\vec{x}_i, y_i\}\right) \tag{3.3}$$

Deneysel risk, beklenen riskin yansız bir tahmin edicisidir, yani, $E\left[R_{deney}(f)\right] = R_X(f)$ 'dir. Yani,

$$\begin{aligned}
E\left[R_{deney}(f)\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} \ell\left(f, \{\vec{x}_i, y_i\}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[\ell\left(f, \{\vec{x}_i, y_i\}\right)\right], \quad \{\vec{x}_i, y_i\}\text{'ler bağımsız.} \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[\ell\left(f, \{\vec{x}, y\}\right)\right], \quad \{\vec{x}_i, y_i\}\text{'ler aynı biçimli dağılır.} \\
&= \frac{1}{n} nE\left[\ell\left(f, \{\vec{x}, y\}\right)\right] \\
&= E\left[\ell\left(f, \{\vec{x}, y\}\right)\right] = R_X(f)
\end{aligned}$$

dir. Böylelikle, DRM prensibi, Eşitlik (3.1) ile verilen beklenen risk yerine, Eşitlik (3.3)'teki deneysel riski minimize etmeye çalışır. Sonuç olarak, beklenen riskin de minimize edildiği varsayılır. Bu durumda, deneysel riskin hesaplandığı eğitim kümesinin türetildiği $P(\vec{x})$ girdi dağılımıyla aynı dağılımdan gelen test örneklerinin çıktıları doğru bir şekilde tahmin edilebilir.

$$f^*_{deney} = \arg\inf_{f \in H} R_{deney}(f) \tag{3.4}$$

Deneysel riskin doğrudan minimize edilmesi birkaç probleme neden olur. Bu problemin ilki, problemin çözümünün kötü-konumlanmış olmasıyla ilgilidir çünkü deneysel riski minimize eden birçok hatta sonsuz sayıda fonksiyon bulunabilir. İkinci olarak, bu durum aşırı uyuma neden olabilir, yani, deneysel risk sıfıra yakın bir değer bulunsa bile, asıl hesaplanmaya çalışılan beklenen risk çok büyük olabilir. Çünkü model karmaşıklığı arttıkça, deneysel risk sıfıra yaklaşacaktır. Ancak, sıfır bir deneysel riske sahip model test kümesi üzerinde iyi sonuç vermeyebilir. Karmaşık bir model ile eğitim verisinin altında yatan daha karmaşık yapılara uyum sağlanabilir ancak yeni veri kümesi üzerindeki tahmin hatası yükselecektir. Bu yüzden, deneysel riskin minimum tutulması, beklenen riskin veya genelleştirme hatasının küçük olmasını garanti edemez [25].

Uygulamalarda DRM ile elde edilen f^*_{deney} hipotez fonksiyonu ile öğrenme problemine en iyi uyumu sağlayan f^* hipotez fonksiyonunun birbirine yakınsaması beklenir. Yani, beklenen risk ile deneysel risk arasındaki fark oldukça küçük olmalıdır.

$$R_{deney}(f) \approx R_X(f)$$

Ancak bu yakınsama 4 faktöre bağlıdır:

1. Eğitim veri kümesinin büyüklüğü: Elde edilen herhangi bir hipotez fonksiyonu için, eğitim veri kümesi büyüdükçe $R_{deney}(f) \approx R_X(f)$ olması beklenir.
2. Veri kümesinin geldiği dağılımın bilinmesi: Eğitim verisindeki tüm gözlemlerin türetildiği dağılımın bilinmesiyle, bu dağılıma ait en iyi tahmin elde edilebilir.
3. Kullanılan Kayıp Fonksiyonu
4. Hipotez uzayı H : En iyi tahmin fonksiyonunun araştırıldığı hipotez uzayı ne kadar büyük ve bu uzayda bulunan fonksiyonlar ne kadar karmaşık ise, iyi bir yaklaşım elde etmek o kadar zordur.

Bu bölümün devamında DRM ile elde edilen f^*_{deney} hipotez fonksiyonu ile öğrenme problemine en iyi uyumu sağlayan f^* hipotez fonksiyonunun hangi koşullar altında birbirlerine eşit olabileceği tartışılacaktır.

Anlaşılacağı üzere, DRM'nin performansı, hipotez uzayındaki en uygun fonksiyonu bulmak için çalışan ilgili keşif algoritmasına bağlıdır. Bu fonksiyonu bulabilmek için

hipotez uzayındaki her elemanı arayan iterasyon yöntemini kullanır. Böylelikle Eşitlik (3.4) ile verilen infimumu bulmak mümkündür. Ancak, bu prensip ile çözülen bir optimizasyon probleminde, öğrenme algoritmasının başlangıç pozisyonundan çok uzak olmayan bir yerel minimumu bulması ve bu noktaya takılması mümkündür. Bu prensibin diğer bir sıkıntısı ise, prensibin veri setinin büyüklüğü ile ilişkisindedir. Herhangi bir veri bulunmadığında, deneysel risk (veya bu durumda eğitim kaybı) sıfır olacaktır ve tahmin fonksiyonu eğitim kümesindeki tüm elemanları doğru bir şekilde sınıflandırıncaya kadar sıfır olarak kalacaktır. Veri sayısı artmaya başladığında, deneysel risk de artmaya başlayacaktır, böylelikle eğitim kümesindeki sayısı artan elemanları doğru bir şekilde sınıflandıramayacaktır. Bu durumda deneysel riskin eğitim kümesindeki veri sayısına bağlı monoton artan bir fonksiyon olduğu söylenebilir.

Deneysel risk ile beklenen risk arasındaki fark, öğrenme makinesinin genelleştirilebilme özelliğiyle yakından ilgilidir. Bir öğrenme makinesinin genelleştirilebilme özelliğe ise, o algoritmanın bağımsız bir test verisi üstündeki tahmin kapasitesine bağlıdır. Yani, deneysel risk minimize edilerek, küçük beklenen riske sahip bir fonksiyon seçilme olasılığı yüksektir. Bu fonksiyonun test (beklenen) ve eğitim (deneysel) performanslarının yüksek ilişkili olması, doğru tahmin fonksiyonunun elde edildiği anlamına gelir. Deneysel risk ile beklenen risk arasındaki fark, *genelleştirme hatası (generalization error)* olarak bilinir.

Fakat düşük deneysel riske ve genelleştirilemeyen bir tahmin fonksiyonuna sahip bir öğrenme algoritması kullanışlı değildir. Benzer şekilde, genelleştirilebilen ancak yüksek deneysel riske sahip bir tahmin fonksiyonunun DRM yaklaşımı ile tanımlanması mümkün değildir.

Sadece DRM prensibiyle seçilen tahmin fonksiyonlarının genelleştirilebilme potansiyelleri oldukça düşüktür. Yukarıda da bahsedildiği gibi, deneysel riski minimize eden birçok fonksiyon bulunabilir ancak bunlardan sadece biri en büyük genelleştirilebilme potansiyeline sahiptir. Bütün olası fonksiyonlardan oluşan bir hipotez uzayı düşünüldüğünde, sıfır deneysel riske sahip ve genelleştirilebilme potansiyeli olmayan karmaşık bir fonksiyon ile herhangi bir eğitim kümesine kolaylıkla uyum sağlanabilir. İşte bu nedenle, hipotez uzayının kapasitesi sınırlandırılmazsa, öğrenme işlemi doğru bir şekilde gerçekleştirilemez. Bölüm 2.13'de hipotez uzayındaki eğitim verisine en iyi uyumu sağlayan, genelleştirilebilir ve küçük deneysel riske sahip hedef fonksiyonunun bulunmadığı gereksiz bölgeleri ekarte eden iki düzenleme yöntemi (Ivanov/Tikhanov

Düzenlemeleri) tanıtıldı. Eğitim verisinin sayısı arttıkça, verinin geldiği dağılım ile ilgili daha güçlü varsayımlar yapılarak, hipotez uzayı daha iyi bir şekilde düzenlenebilir.

En küçük kareler gibi bazı geleneksel istatistiksel yöntemler Eşitlik (3.5) ile verilen ve büyük sayılar yasası gereği doğru olan prensibi temel alırlar. Bu prensip YSA ve Bulanık Mantık gibi yöntemler tarafından da kullanılmaktadır.

$$\lim_{n\to\infty} R_{deney}\left(f\right) = R_X\left(f\right) \tag{3.5}$$

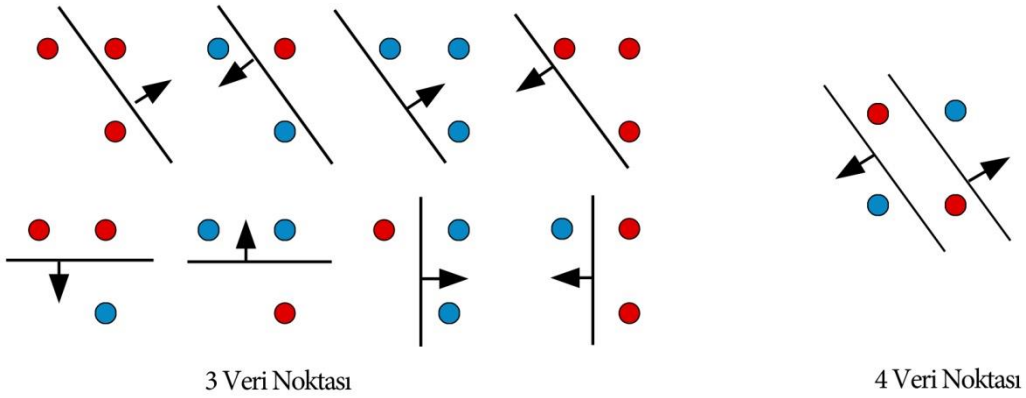
Bu prensibe göre eğitim verilerinin sayısı n sonsuza doğru artarken, deneysel risk beklenen riske yakınlaşır. Ancak DRM prensibi deneysel riski minimize eden fonksiyonu garanti altına almaz. Düzgün yakınsama prensibine göre n değeri artarken deneysel risk beklenen riske daha çok yakınlaşır. Fakat DRM öğrenme algoritmasının uygunluğu ve yakınsama prensibi, yapısal prosedürlerin nasıl oluşturulacağı hakkında bilgi vermez. Gerekli modellerin oluşturulması için yapısal prosedürlerin tanıtılması gerekmektedir. YRM ve VC teorisinin geliştirilmesiyle bu kavramlar oluşturulmuş ve DVM'nin temeli atılmıştır.

3.2. Yapısal Risk Minimizasyonu ve Vapnik-Chervonenkis Teoremi

VC teorisi ve YRM tümevarımsal prensibi istatistiksel öğrenme teorisinde önemli bir yere sahiptir. YRM, küçük örneklem kümelerinde öğrenme işlemi ile ilgili uygulamalarda yararlı olan yeni bir tümevarımsal prensiptir [4]. Olasılıkta *düzgün yakınsama (uniform convergence)*, öğrenme makineleri için beklenen risk ile deneysel risk arasındaki sapma değerine bir sınır getirir [31]. İstatistiksel öğrenme teorisinde bu genelleştirme sınırları iki yolla dar tutulabilir. İlki hipotez uzayının kapasitesini sınırlandırmaktır. Böylelikle sonsuz boyutlu bile olsa bu uzayda bulunan fonksiyonların sayısı kabaca hesaplanabilir. İkinci ise fonksiyonların durağanlığını sınırlandırmak (Ivanov Düzenlemesi) veya eğitim kümesindeki dalgalanmaya karşı tahmin fonksiyonunun duyarlılığı azaltmaktır (Tikhonov Düzenlemesi). Bu yüzden YRM prensibi ve VC teorisindeki ana fikir; çok sayıda tahmin fonksiyonu arasından beklenen risk veya genelleme hatasını minimum yapacak doğru oranda karmaşıklığa sahip modelin seçilmesidir. Böylece, seçilen optimal öğrenme makinesi verilen eğitim kümesi için en uygun kapasiteye sahip olacaktır. İstatistiksel öğrenme teorisinde, uygun fonksiyonun seçildiği fonksiyonlar kümesinin karmaşıklığı (kapasitesi) VC boyutu kullanılarak tanımlanır.

TANIM 3.2. (AYRIŞTIRMA) *n veri noktasına sahip bir veri kümesine sahip olunsun. Bu n adet veri kümesi 2^n farklı yolla pozitif ve negatif olarak etiketlenebilir. Bu durumda, 2^n farklı öğrenme problemi bu n veri noktası tarafından tanımlanabilir. Bu problemlerin herhangi biri için pozitif örnekleri negatiflerden ayırabilen $h \in H$ gibi bir hipotez bulunursa, H fonksiyon kümesinin, n adet veri bulunan bu veri kümesini ayrıştırabildiği söylenir. Böylelikle n veri noktasına sahip herhangi bir öğrenme problemi, H 'den çekilen bir hipotez fonksiyonuyla hatasız bir şekilde öğrenilebilir* [32].

TANIM 3.3. (VC-BOYUTU) *H hipotez uzayı tarafından hatasız bir şekilde ayrıştırılabilen veri noktalarının maksimum sayısına VC-boyutu denir ve $VC(H)$ ile gösterilir. VC-boyutu H hipotez uzayının kapasitesini ölçer. Kapasite ise fonksiyon karmaşıklığının bir ölçüsüdür ve uzayda bulunan fonksiyonların gücü, zenginliği ve esnekliği ile ilgili bilgi verir* [33].



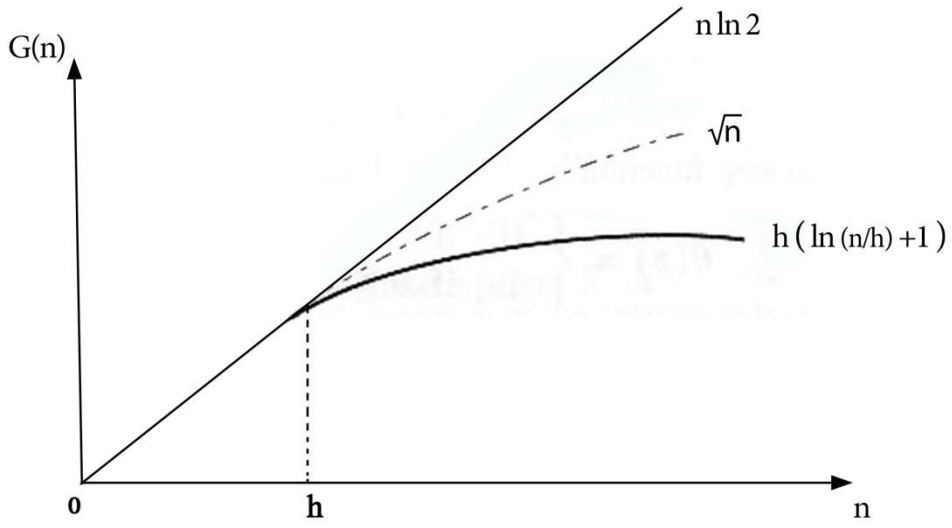
Şekil 3.1. İki boyutlu uzayda 3 ve 4 veri noktası için VC-boyutu

Şekil 3.1'de $n = 3$ *ve* $n = 4$ veri örneği sayıları ve $d = 2$ boyut değeri için mümkün olan sınıflandırma durumları gösterilmiştir. $n = 3$ için, 3 nokta olası $2^3 = 8$ farklı şekilde 2 ayrı sınıfa parçalanabilir. $\mathbb{R}^2$ uzayında görüntülen bu üç nokta, doğrusal bir ayırıcı düzlem ile 8 farklı şekilde sınıflandırılabilir. Fakat $n = 4$ için olası $2^4 = 8$ durumun tamamında doğrusal olarak sınıflandırma yapılamaz. Sonuç olarak, her iki durum için de VC boyutu $d + 1$ yani 3'tür. İki boyutlu uzaydaki doğrusal olmayan sınıflandırıcılar için VC boyutu daha büyük olabilir. Genel olarak, $\mathbb{R}^d$, d boyutlu özellik uzayındaki doğrusal ayırıcı

hiperdüzlem ile yapılan sınıflandırma işleminde, VC boyutu $h = d + 1$ (d tane özellik ve b sabiti) olacağı söylenebilir.

Sınıflandırma ve regresyon uygulamalarında VC boyutunun değeri hesaplanabilir veya tahmin edilebilir. Fakat bazı uygulamalarda VC boyutunun hesaplanması hiç kolay değildir. Çünkü bilinmeyen bir değer veya sonsuz olabilmektedir. Bu durumda iyi bir sınıflandırma için sonsuz sayıda eğitim örneğine ihtiyaç vardır.

Öğrenme makinesinin genelleşebilme yeteneği üzerindeki dağılımdan bağımsız sınırlar elde edilmesi ve hipotez uzayının karmaşıklığı (kapasitesi) büyüme fonksiyonu kullanılarak hesaplanabilir. Vapnik ve Chervonenkis (1971) [34], n sayıda örnek için büyüme fonksiyonun ya doğrusal ya da logaritmik fonksiyonu olması gerektiğini savunmaktadır. Şekil 3.2'deki gibi $n = h$ noktasında yükselme fonksiyonu eğrisi düşüşe geçer. Bu noktaya VC boyutu denir. Eğer VC boyutu değeri sonsuz değilse büyüme fonksiyonu yeterli örnek için doğrusal olarak artmaz [35].



Şekil 3.2. Büyüme fonksiyonu davranış grafiği

YRM prensibi genelleştirme hatası sınırlarını minimum tutmaya çalışır. Bu işlemi, yalnızca sonlu sayıdaki eğitim örneklerini kullanarak ve çok sayıdaki tahmin fonksiyonları arasından en iyisini seçerek gerçekleştirir. Bu şekilde, gerçek hata oranını yansıtan beklenen risk fonksiyonunu doğrudan minimum tutmaya çalışmak yerine YRM ve VC teorisi kullanılarak hipotez uzayındaki en uygun hipotez seçilir.

YRM prensibi, az sayıda örneklem içeren eğitim kümesi için oldukça yararlıdır. Ancak, DRM prensibi bu gibi durumlarda pek uygun değildir. Çünkü küçük deneysel risk beklenen riskin küçük olmasını garantilemez. YRM prensibini temel alan istatistiksel öğrenme teorisi aynı anda hem deneysel riski hem de VC boyutunu minimum tutarak öğrenme makinesinin genelleştirilebilme yeteneğini kontrol edebilir. Böylece, herhangi bir dağılıma sahip fonksiyon için YRM prensibi en iyi çözüme yakınsama garantisi verir [36, 37]. Ayrıca genelleme hatasının üst sınırının analiz edilmesi, hızlı bir yakınsama ve dağılımdan bağımsız DRM öğrenme yöntemi ile elde edilen tahmin fonksiyonlarının VC boyutu değerinin de sonlu bir değer olduğunu işaret eder. İstatistiksel öğrenme teoremi, öğrenme makinelerinin genelleştirilebilme yeteneğini Eşitlik (3.6)'daki genelleştirilmiş risk fonksiyonelini minimum tutarak kontrol eder. Bu, istatistiksel öğrenme açısından önemli bir sonuçtur.

$$R_X(f) \le R_{deney}(f) + \sqrt{\frac{h\left(\log\left(2n/h\right)+1\right)-\log\left(\eta/4\right)}{n}} \tag{3.6}$$

Burada, $R_X(f)$ beklenen gerçek riski, $R_{deney}(f)$ deneysel riski, h VC boyutu değerini, n eğitim örneklerinin sayısını ve η, $0 \le \eta \le 1$ aralığında güven düzeyini belirleyen gerçel bir değerdir. Bu işlemde fonksiyon kümesi mümkün olduğu kadar basit tutulurken, eğitim hatası da minimum tutulur [35].

TANIM 3.3. (YAPISAL RİSK MİNİMİZASYONU) $H = \{f : f(\vec{x}) = (\vec{w} \cdot \vec{x}) + b\}$ *fonksiyonlar uzayının tümü,* $H_1 \subset H_2 \subset ... \subset H_k$ *olacak biçimde iç içe alt uzaylara parçalansın. Öyle ki bu alt uzaylar,* $V(H_1) \le V(H_2) \le ... \le V(H_k)$ *şeklinde VC boyutuyla ölçülen artan kapasiteye sahip olsun. Örneğin, sınıflandırma görevi için, aşağıdaki gibi doğrusal fonksiyonlar dizisi alınabilir.*

$$\begin{aligned} H_1 &= \{f : f(\vec{x}) = \mathrm{sgn}[b + w_1 x_1]\} \\ H_2 &= \{f : f(\vec{x}) = \mathrm{sgn}[b + w_1 x_1 + w_2 x_2]\} \\ &\vdots \\ H_k &= \{f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})]\} \end{aligned}$$

Burada, k, bir gözlem vektörünün büyüklüğüdür. VC-boyutu doğrusal olarak artar ve bağımsız parametre sayısına eşittir; $V(H_1) = 2$, $V(H_2) = 3$ *,...,* $V(H_k) = k + 1$. *Alternatif*

olarak, doğrusal olmayan sınıflandırma fonksiyonlar ailesinin aşağıdaki gibi bir dizisi alınabilir:

$$H_1 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})] \}$$
$$H_2 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x}) + (\vec{w} \cdot \vec{x})^2] \}$$
$$\vdots$$

Ayrıca, sınırlı ağırlık vektörlerine sahip doğrusal sınıflandırma fonksiyonlarının bir dizisi olabilir:

$$H_1 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})] \; öyle\, ki \; 2/\|\vec{w}\| \le R_1 \}$$
$$H_2 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})] \; öyle\, ki \; R_1 \le 2/\|\vec{w}\| \le R_2 \}$$
$$\vdots$$

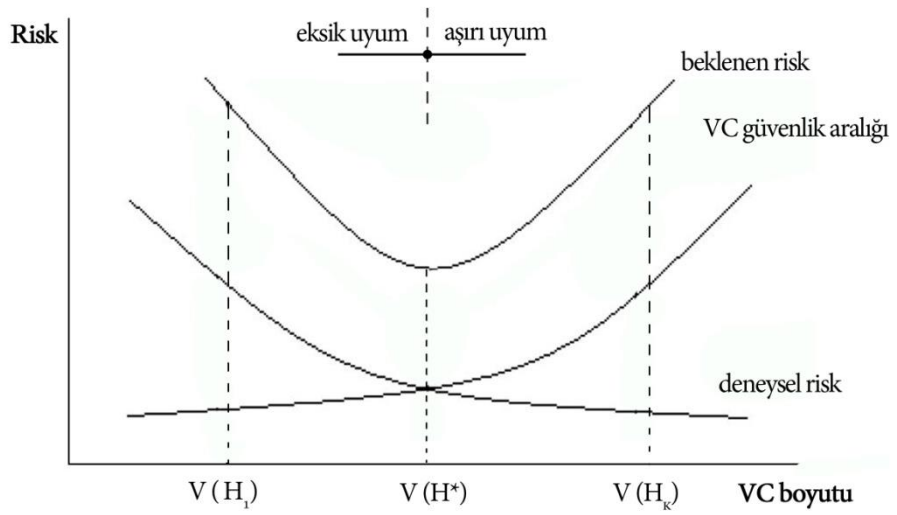
Veya, bu fonksiyonlar dizisinin geometrik sınır cinsinden yeniden formüle edilmiş durumları olabilir:

$$H_1 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})] \; öyle\, ki \; \gamma^* \ge g_1 \}$$
$$H_2 = \{ f : f(\vec{x}) = \mathrm{sgn}[b + (\vec{w} \cdot \vec{x})] \; öyle\, ki \; \gamma^* \ge g_2 \ge g_1 \}$$
$$\vdots$$

Kullanılacak iç içe modellerin seçimi, yapılacak sınıflandırma/bağlanım görevi ile ilgili önceden düşünülerek yapılır. Örneğin, verinin doğrusal olmayan bir dağılımdan geldiği düşünülüyorsa, artan dereceli polinomiyal sınıflandırma/bağlanım fonksiyonları kullanılabilir. Ancak, bunun kararı, VC-boyutunun dağılımdan bağımsız olma özelliğini sağlamak için, eğitim kümesi türetilmeden önce verilmesi gerekmektedir. Böylelikle, verilen bir eğitim kümesi için, YRM prensibi, Eşitlik (3.6) ile verilen risk fonksiyonelinin sınırını minimize eden alt kümeyi seçer.

Seçilecek fonksiyonlar dizisi ne olursa olsun, birinci alt uzayda (H_1) bulunan fonksiyonlar en basit yapıda olanıdır ve büyük bir deneysel risk ile küçük bir VC terimine sahiptir. Ancak, bu fonksiyonlar tüm eğitim kümesini doğru bir şekilde sınıflandıramayabilirler çünkü basit modeller yeterli temsil gücüne sahip değildirler, fakat sadece birkaç tane serbest parametreye sahip olduklarından veriye karşı duyarlı değildirler ve kararlıdırlar. k. hipotez uzayı (H_k) en karmaşık yapıya sahiptir ve bu özelliğiyle de küçük eğitim hatası ile yüksek bir VC değeriyle sonuçlanır. Uygulamalarda, güçlü yakınsama kapasitesine sahip daha karmaşık modeller daha fazla parametreye sahiptir, fakat bu durumda da aşırı

uyum problemi ortaya çıkar. Genelleşme hatasının üst sınırını minimize eden optimal fonksiyon en basit ve en karmaşık olan fonksiyonlar sınıfının arasında bir karmaşıklığa sahiptir. Beklenen riskin Eşitlik (3.6) ile gösterilen üst sınır fonksiyonu ve bu fonksiyonun eksik uyum ile aşırı uyum durumları arasındaki ilişkileri Şekil 3.2'de verilmiştir.



Şekil 3.3. Beklenen risk fonksiyonu yapısı

Parametrelerin doğrusal olduğu makine öğrenmesi yöntemlerinde, VC boyutu serbest parametrelerin sayısına eşdeğerdir. Diğer bir ifadeyle öğrenme makinelerinin karmaşıklığı bağımsız parametrelerin sayısıyla ifade edilir. Azalmayan VC boyutlu hiyerarşik hipotez yapısı $H_1 \subset H_2 \subset ... \subset H_k$, azalmayan sayılı bağımsız parametrelere göre kurulmuş modelleri ifade eder. Her bir model kendisinden önceki daha az karmaşık yapıdaki modeli kapsar. Aslında hiyerarşik yapı tüm bu modellere uygulanabilecek genel bir tasarımdır. Bir taraftan hipotez uzayı H_1'den H_k'ye kadar daha yüksek dereceden bileşenler kullanarak artarken, öğrenme makinesinin karmaşıklığı da ek ayarlanabilir parametrelerin kullanılmasıyla artar. Ancak diğer taraftan da, bu ilave parametrelerinin de aynı veri kümesi vasıtasıyla en uygun değerlerinin bulunması gerekir ki, bu da beklenen riskin minimuma ulaşmasını yavaşlatır. Deneysel riskin beklenen riske kararlı olarak yakınsamasını sağlamak için, bağımsız parametre sayısındaki artışın veri kümesindeki artışla desteklenmesi gerekmektedir. Bu yüzden VC boyutu kavramına göre kurulan bu tip modellerin hedefi, eğitim hatası ve *model karmaşıklığını (sample complexity)* dengelemek amacıyla uygun sayıda bağımsız parametre içeren alt modellerden birisinin bulunmasıdır.

4. DESTEK VEKTÖR MAKİNELERİ

DVM, yüksek boyutlu ve küçük sayıda eğitim verisinden öğrenebilen yeni nesil bir öğrenme metodu olarak Vladamir Vapnik tarafından 1995 yılında önerilmiştir [38]. DVM'nin dayandığı teori olan "İstatistiksel Öğrenme Teorimi", Vladamir Vapnik ve Alexey Chervonenkis tarafından 1960da önerilmiş ve 1970'li yıllarda geliştirilmiştir. DVM, genelleştirebilme yeteneğinin yüksek olması, teorik yapısının kuvvetli ve uygulamalarda yüksek performans göstermesi nedeniyle son yıllarda örüntü tanımlamada, bağlanım analizinde, yüz tanımlamada, resim ve metin sınıflandırmada, veri madenciliğinde, kalite kontrol yöntemlerinde, finans, ekonomi, genetik, biyoloji ve diğer biyoenformatik uygulamalarda sıklıkla kullanılmaya başlanmıştır.

4.1. İkili Sınıflandırma İçin Destek Vektör Makineleri

$H = \mathbb{R}^d$ Hilbert uzayında tanımlı n tane girdi vektörü ve bu girdi vektörlerine karşılık gelen ve sadece iki değer alabilen çıktılar aşağıdaki gibi gösterilebilir.

$$S = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_n, y_n)\} \subseteq \mathbb{R}^d \times \{+1, -1\} = H \times Y$$

Bu eğitim kümesindeki verilerin $P(\vec{x}, y) = P(\vec{x}) \cdot P(y \mid \vec{x})$ gibi bilinmeyen bir bileşik olasılık dağılımından çekildiği ve tüm veri noktalarının bağımsız ve aynı biçimli dağıldığı varsayılır. Bu öğrenme algoritmasıyla yapılmak istenen gürültü varlığında, $\vec{x}_t$ için doğru çıktıyı tahmin edebilecek $f(\vec{x})$ gibi bir tahmin fonksiyonunun bulunmasıdır.

$$y_t = \max_{y \in \{+1, -1\}} P(y \mid \vec{x}_t)$$

Optimal tahmin fonksiyonu $f(\vec{x})$ için yapılan arama, Bölüm 3'te detaylı bir şekilde anlatılan YRM prensibi kullanılarak sınırlı fonksiyonel uzayda gerçekleştirilir. Elde edilecek tahmin fonksiyonunun eğitim veri kümesine ne kadar iyi uyum sağladığı veya yeni gözlemleri ne kadar iyi tahmin ettiği kayıp fonksiyonu yardımıyla bulunabilir.

Tanım 4.1. (KAYIP FONKSİYONU) Bir kayıp fonksiyonu, iki değişkene ait negatif olmayan bir fonksiyon olarak tanımlanır:

$$\ell(f, z) = \ell(f, \{\vec{x}, y\})$$
$$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$$

Literatürde var olan kayıp fonksiyonları ya $\left| f(\vec{x}) - y \right|$'de ya da $-yf(\vec{x})$'de azalmayan fonksiyonlardır.

Sınıflandırma problemleri için kullanılan bazı kayıp fonksiyonları aşağıda verilmiştir [6]:

(i) *Gösterge Kayıp Fonksiyonu (Indicator Loss Function):* İkili sınıflandırma problemleri için kullanılan en sezgisel kayıp fonksiyonudur. Ancak bu fonksiyon konveks ve diferansiyellenebilir ($x = 0$'da türev tanımsızdır.) değildir ve konveks olmayan fonksiyonları, minimizasyon problemlerinde kullanmak oldukça zordur [39]. *θ adım fonksiyonu (step function)* olmak üzere, gösterge kayıp fonksiyonu şu şekilde tanımlanır:

$$\ell(f, \{\vec{x}, y\}) = \theta(-f, \{\vec{x}, y\}) = \begin{cases} 1, & -yf(\vec{x}) \le 0 \\ 0, & -yf(\vec{x}) > 0 \end{cases}$$

(ii) *Hinge Kayıp Fonksiyonu (Hinge Loss Function):* Gösterge kayıp fonksiyonunun aksine, Hinge kayıp fonksiyonu konvekstir ve bu durum, seyrek çözümlemelerde (sparse solutions) kullanışlı algoritmaların olmasını sağlar. DVM için kullanılan bir kayıp fonksiyonudur. Bu kayıp fonksiyonu konveks olsa da, diferansiyellenebilir değildir. Ancak *gevşek değişkenler (slack variables)* kullanılarak bu problemin üstesinden gelinebilir [39]. Hinge kayıp fonksiyonu aşağıdaki gibi tanımlanır.

$$\ell(f, \{\vec{x}, y\}) = (1 - yf(\vec{x}))_+ = \begin{cases} 1 - yf(\vec{x}), & yf(\vec{x}) < 1 \\ 0, & yf(\vec{x}) \ge 1 \end{cases}$$

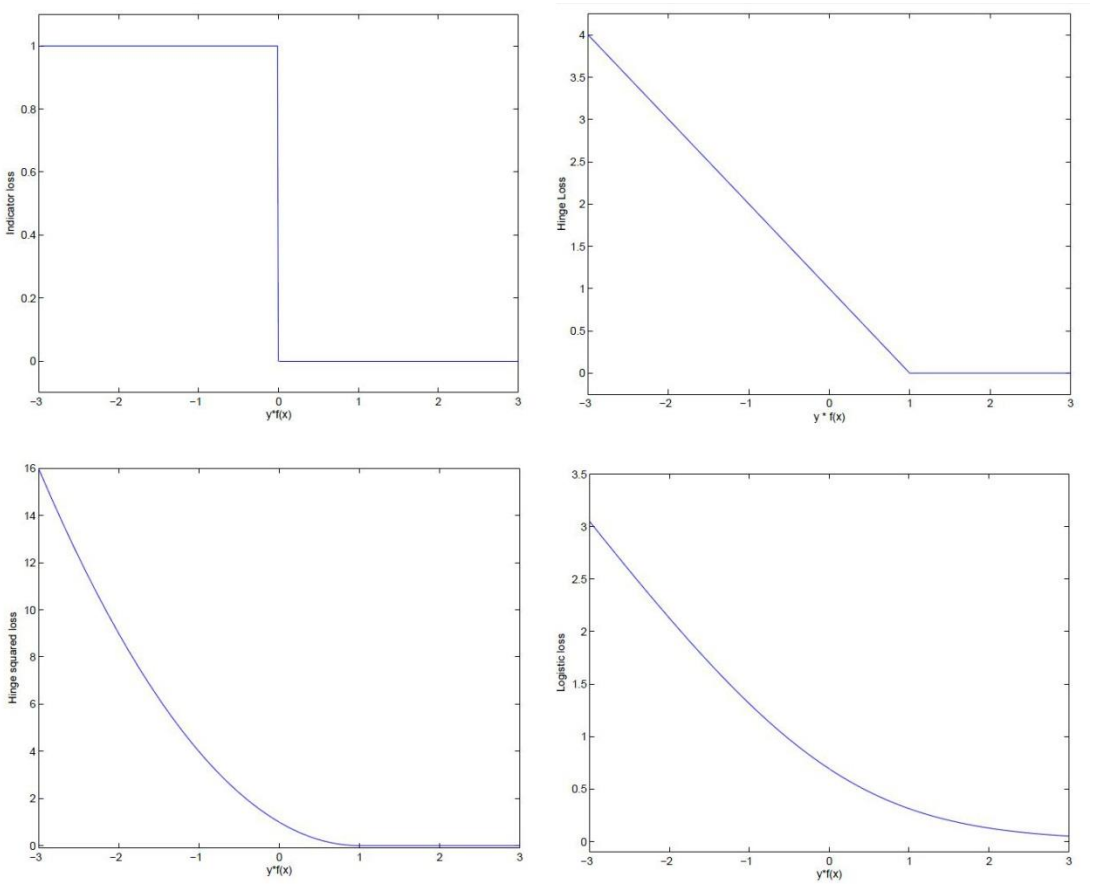
(iii) *Karesel Hinge Kayıp Fonksiyonu (Quadratic Hinge Loss Function):* Hinge kayıp fonksiyonuna benzer ancak hinge kayıp fonksiyonundaki sapmanın karesi alınarak oluşturulan bir kayıp fonksiyonudur.

$$\ell(f, \{\vec{x}, y\}) = [(1 - yf(\vec{x}))_+]^2 = \begin{cases} (1 - yf(\vec{x}))^2, & yf(\vec{x}) < 1 \\ 0, & yf(\vec{x}) \ge 1 \end{cases}$$

(iv) *Lojistik Kayıp Fonksiyonu (Logistic Loss Function):* Konveks bir kayıp fonksiyondur ve bu kayıp fonksiyon ile ilişkili aşağıdaki gibi bir olasılıksal model yazılabilir.

$$\ell(f,\{\vec{x},y\}) = \ln(1+e^{-yf(\vec{x})})$$

Bu dört kayıp fonksiyonu Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Sınıflandırma problemi için kullanılan dört kayıp fonksiyonu

Kullanılabilecek tüm kayıp fonksiyonları eğitim kümesinden hesaplanır. $R_X[f]$ beklenen risk ve $dP(\vec{x},y)$ olasılık dağılımı olmak üzere, kayıp fonksiyonu kullanılarak hesaplanabilecek beklenen risk aşağıdaki gibi ifade edilebilir.

$$R_X[f] = E[\ell(f,\{\vec{x},y\})] = \int_{X\times Y} \ell(f,\{\vec{x},y\})dP(\vec{x},y) = \int_X\int_Y \ell(f,\{\vec{x},y\})dP(y\,|\,\vec{x})dP(\vec{x}) \quad (4.1)$$

Burada, beklenen risk, yanlış sınıflandırma olasılığıdır. Sınıflandırma (veya örüntü tanımlama) problemlerinde amaç, Eşitlik (4.1)'de verilen beklenen riskin, yani bilinmeyen bir olasılık dağılımına göre alınan eğitim veri kümesi üzerinde yanlış sınıflandırma olasılığının minimize edilmesidir. Ancak burada eğitim veri kümesinin çekildiği olasılık dağılımı $P(\vec{x}, y)$ bilinmediğinden, bu beklenen riski hesaplamak neredeyse imkânsızdır. Bu nedenle Eşitlik 4.2.'de verilen deneysel risk hesaplanır.

$$R_{deney}[f] = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f\left(\vec{x}_i\right), \{\vec{x}_i, y_i\}\right) \tag{4.2}$$

Elde edilen deneysel risk ile beklenen risk arasındaki fark küçük ise, bu algoritmanın genelleştirilebilir olduğu söylenir. Eğer bu öğrenme algoritması genelleştirilebilir ise, yeni gelecek test örneklerin çıktılarının tahmininde doğru sonuçlar vereceği anlamına gelir. Ancak genelleştirilebilir ve yüksek deneysel riske sahip bir algoritmanın kullanışlı olduğunu söylenemez [6].

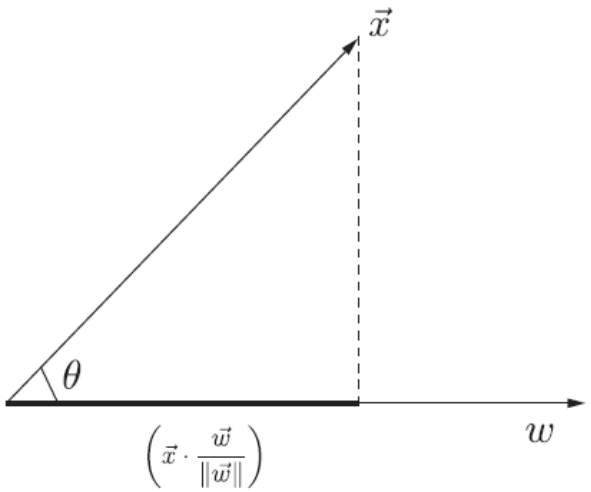
4.1.1. Nokta Çarpımın Geometrisi

Gerçel değerli, $\mathbb{R}^d$ Hilbert uzayında bir doğrusal fonksiyon tanımlansın. Bu fonksiyon $\vec{w} \in \mathbb{R}^d$ ağırlık vektörü ve $b \in \mathbb{R}$ eşik (eğilim) değeri ile parametrize edilsin. Toplamda $d+1$ tane serbest değişkene sahip bu doğrusal fonksiyonu Eşitlik (4.3)'teki gibi yazılabilir.

$$f(\vec{x}) = (\vec{w} \cdot \vec{x}) + b \tag{4.3}$$

$\vec{x}$ vektörünün $\vec{w}$ vektörü üzerine dik izdüşümünün uzunluğuna sayısal izdüşüm denir. $\hat{w} = \vec{w} / \|w\|$ birim vektör ve θ, $\vec{w}$ ile $\vec{x}$ vektörleri arasında açı olmak üzere, sayısal izdüşüm aşağıdaki gibi bir nokta çarpımıyla gösterilir.

$$\left(\hat{w} \cdot \vec{x}\right) = \|\vec{x}\| \cos\theta = \left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x} \right) \tag{4.4}$$



Şekil 4.2. Dik izdüşüm olarak bir iç çarpım

Eşitlik 4.3'te bulunan nokta çarpımı,

$$(\vec{w}\cdot\vec{x}) = \|\vec{w}\|\|\vec{x}\|\cos\theta \tag{4.5}$$

şeklinde yazılabilir. Böylelikle $\vec{x}_1$ ve $\vec{x}_2$ gibi iki girdi vektörüne ait, $\vec{w}$ gibi sabit bir ağırlık vektörü ile olan nokta çarpımları karşılaştırılarak, bu iki vektör arasındaki benzerlik ölçülebilir. Bu nokta çarpımları ile aynı yöne ve farklı büyüklüklere sahip iki vektör ayırt edilebilir. θ_1 ve θ_2, sırasıyla, $\vec{x}_1$ ile $\vec{w}$ ve $\vec{x}_2$ ile $\vec{w}$ arasındaki açılar olmak üzere,

$$\|\vec{x}_1\| = \|\vec{x}_2\| \; ve \; \theta_1 = \theta_2 \Rightarrow (\vec{w}\cdot\vec{x}_1) = (\vec{w}\cdot\vec{x}_2) \Rightarrow \vec{x}_1 = \vec{x}_2$$
$$\|\vec{x}_1\| \neq \|\vec{x}_2\| \; ve \; \theta_1 = \theta_2 \Rightarrow (\vec{w}\cdot\vec{x}_1) \neq (\vec{w}\cdot\vec{x}_2) \Rightarrow \vec{x}_1 \neq \vec{x}_2$$

dir. Benzer şekilde, aynı büyüklüklere sahip ancak farklı yönlere uzanan vektörler de nokta çarpımı ile aşağıdaki gibi ayırt edilebilir.

$$\theta_1 = \theta_2 \; ve \; \|\vec{x}_1\| = \|\vec{x}_2\| \Rightarrow (\vec{w}\cdot\vec{x}_1) = (\vec{w}\cdot\vec{x}_2) \Rightarrow \vec{x}_1 = \vec{x}_2$$
$$\theta_1 \neq \theta_2 \; ve \; \|\vec{x}_1\| = \|\vec{x}_2\| \Rightarrow (\vec{w}\cdot\vec{x}_1) \neq (\vec{w}\cdot\vec{x}_2) \Rightarrow \vec{x}_1 \neq \vec{x}_2$$

Ancak, hem büyüklükleri hem de yönleri farklı vektörlerin eşitliği hakkında genel bir çıkarsama yapmak imkânsızdır. Bu, nokta çarpımının doğasında olan bir zayıflıktır.

$$(\vec{w}\cdot\vec{x}_1) = (\vec{w}\cdot\vec{x}_2) \not\Rightarrow \vec{x}_1 = \vec{x}_2$$

Eşitlik (4.4) skaler olduğu için, herhangi bir yöne sahip değildir. Bu nedenle, vektör izdüşümünün elde edilmesi için, $(\hat{w}\cdot\vec{x})$ skaler değeri, $\vec{w}$ vektörünün yönü olan $\hat{w}$ ile çarpılır.

$$\left(\hat{w}\cdot\vec{x}\right)\hat{w}=\left(\|\vec{x}\|\cos\theta\right)\frac{\vec{w}}{\|\vec{w}\|}=\frac{\vec{w}}{\|\vec{w}\|}\left(\frac{\vec{w}}{\|\vec{w}\|}\cdot\vec{x}\right)^{T} \tag{4.6}$$

4.1.2. Hipotez Uzayını Düzenlemek ve Ayırıcı Düzlemler

Bir ikili sınıflandırma probleminde ilk yapılacak iş, girdi uzayını iki yarım uzaya parçalamaktır. Bu yarım uzaylardan biri pozitif değerler alan sınıfı ($y_i=+1$), diğeri negatif değerler alan sınıfı ($y_i=-1$) temsil eder. O halde bu iki sınıfı birbirinden ayıran bir hiperdüzlem aşağıdaki gibi tanımlanabilir.

$$\mathfrak{H}=\{\vec{x}\in\mathbb{R}^{d}\mid f(\vec{x})=(\vec{w}\cdot\vec{x})+b=0\} \tag{4.7}$$

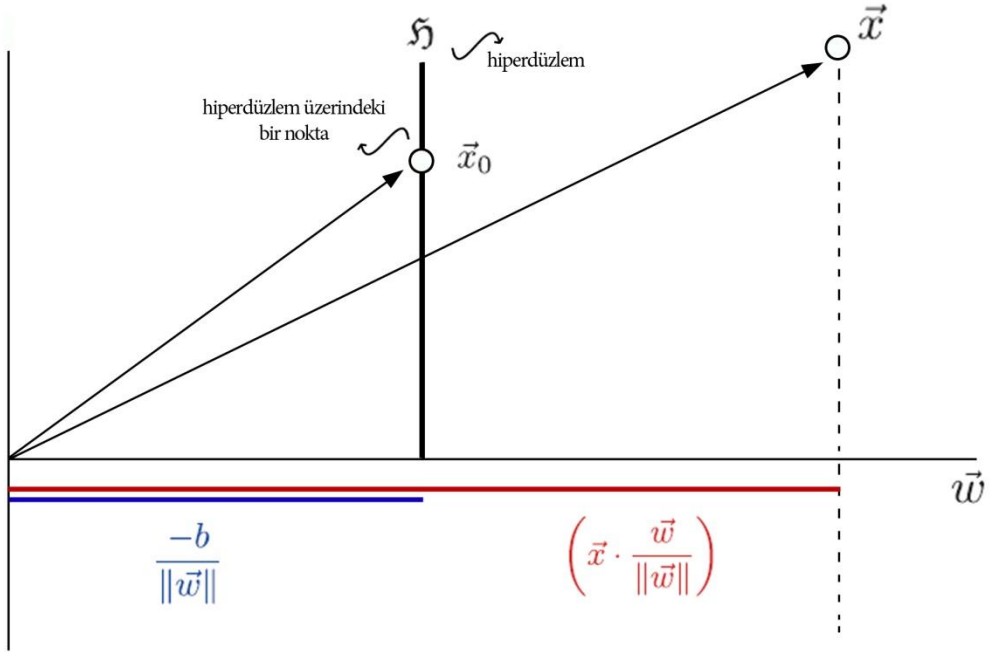
Bu tanımlama ile pozitif ve negatif sınıflar iki ayrı alt uzay oluşturacak biçimde sırasıyla $\{\vec{x}\mid f(\vec{x})>0\}$ ve $\{\vec{x}\mid f(\vec{x})<0\}$ olarak tanımlanır. Yani, $y_i=+1$ çıktısı alan tüm girdi örnekleri için $f(\vec{x})>0$; $y_i=-1$ çıktısı alan tüm girdi örnekleri için $f(\vec{x})<0$ geçerlidir. Bu nedenle, $(\vec{x}_t,y_t)$ gibi yeni gelecek test örnekleri için sınıflandırma bu kural kullanılarak yapılır. $f(\vec{x}_t)>0$ olduğu zaman $y_t=+1$, $f(\vec{x}_t)<0$ olduğu zaman $y_t=-1$ olacaktır. O halde, ortaya çıkan sınıflandırma kuralı ve karar (tahmin) fonksiyonu, "sgn" işaret fonksiyonunu göstermek üzere, aşağıdaki gibi tanımlanır.

$$y_t=g(\vec{x}_t)=sgn[f(\vec{x}_t)]=sgn[(\vec{w}\cdot\vec{x}_t)+b] \tag{4.8}$$

Eşitlik (4.8)'de anlaşılabileceği üzere, bu kuralı kullanarak, iki sınıf arasında sonsuz sayıda hiperdüzlem çizilebilir. Peki, bu durumda kullanılan eğitim veri kümesi için elde edilen ayırıcı problemi çözebilen ancak yeni gelecek test örnekleri için farklı performanslar sergileyecek bu hiperdüzlemlerden en iyisi nasıl seçilebilir? Örneğin, hiperdüzlem, $y_t=-1$ çıktısı alan sınıfın üyelerin olduğu bölgeye yakın çizilsin. Sezgisel olarak düşünüldüğünde, böyle bir ayırma probleminde, $y_t=+1$ olarak sınıflanması gereken test örnekleri üzerinde hata yapma olasılığı düşük olacaktır. Ancak, $y_t=-1$ olarak sınıflandırılması gereken test örnekleri üzerinde hata yapma olasılığı yükselecektir. Bu nedenle, bu ayırıcı

hiperdüzlemin, hem $y_i = +1$ hem de $y_i = -1$ çıktısı alan sınıflardan yeteri kadar uzaklıkta (mesela, iki sınıfın tam ortasına) çizilmesi daha uygun olacaktır.

Hiperdüzlem üzerinde herhangi bir noktanın fonksiyon değerinin sıfır olduğu Eşitlik (4.7)'de verilmiştir. $\vec{x}_0$, $\mathfrak{H}$ hiperdüzlemi üzerinde bulunan bir nokta ise, $f(\vec{x}_0) = (\vec{w} \cdot \vec{x}_0) + b = 0$ olacaktır.



Şekil 4.3. Sayısal izdüşüm

Şekil 4.3'te görüleceği üzere, hiperdüzlem üzerinde bulunan $\vec{x}_0$ noktasının, $\vec{w}$ vektörü üzerine dik izdüşümünün uzunluğu (sayısal izdüşüm), $(\vec{w} \cdot \vec{x}_0) = -b$ olduğundan, $\frac{-b}{\|\vec{w}\|}$'dır. $\frac{-b}{\|\vec{w}\|}$ aynı zamanda, $\mathfrak{H}$ hiperdüzleminin orijinden olan uzaklığını verir. $\vec{x}$ noktasının $\vec{w}$ vektörü üzerine dik izdüşümünün uzunluğu ise $\left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}\right)$'dir. Bu durumda, $\vec{x}$ noktasının $\mathfrak{H}$ hiperdüzlemine olan uzaklığı ise şu şekilde hesaplanır:

$$\left(\frac{\vec{w}}{\|\vec{w}\|}\cdot\vec{x}\right)-\left(\frac{\vec{w}}{\|\vec{w}\|}\cdot\vec{x}_0\right)=\frac{1}{\|\vec{w}\|}[(\vec{w}\cdot\vec{x})-(\vec{w}\cdot\vec{x}_0)]$$
$$=\frac{1}{\|\vec{w}\|}[(\vec{w}\cdot\vec{x})+b]$$
$$=\frac{1}{\|f'(\vec{x})\|}f(\vec{x})$$

TANIM 4.2. (FONKSİYONEL UZAKLIK) $(\vec{x}_i, y_i)$ *gibi bir eğitim örneği ile* $\mathfrak{H}$ *hiperdüzlemi arasındaki fonksiyonel uzaklık (dik ve işaretli) aşağıdaki gibi hesaplanır:*

$$\gamma_i = y_i f(\vec{x}_i) = y_i[(\vec{w}\cdot\vec{x}_i)+b] \tag{4.9}$$

S uzayında bulunan her eğitim örneği ile $\mathfrak{H}$ *hiperdüzlemi arasında hesaplanan tüm fonksiyonel uzaklıkların en küçüğü* γ *olmak üzere,*

$$\gamma = \min_{(\vec{x}_i, y_i)\in S} \gamma_i = \min_{(\vec{x}_i, y_i)\in S} y_i f(\vec{x}_i) = \min_{(\vec{x}_i, y_i)\in S} y_i[(\vec{w}\cdot\vec{x}_i)+b] \tag{4.10}$$

dır.

Eşitlik (4.9)'dan anlaşılacağı üzere, veri noktasına ait çıktı değeri $y_i = +1$ olduğunda, ayırıcı hiperdüzlem ile bu veri noktası arasındaki fonksiyonel uzaklığın maksimum olması için (yani yapılacak tahminin güvenilir ve geçerli olması için), $(\vec{w}\cdot\vec{x}_i)+b$'nin en büyük pozitif bir değer alması gerekir. Benzer şekilde, $y_i = -1$ olduğunda, $(\vec{w}\cdot\vec{x}_i)+b$'nin en büyük negatif bir değer alması gerekir. Ancak, fonksiyonel uzaklık ile ilgili problem tanımın içeriğindedir. Eşitlik (4.8)'de de verildiği üzere, herhangi bir $\vec{x}_i$ eğitim örneği için çıktı $y_i = g(\vec{x}_i) = sgn[f(\vec{x}_i)] = sgn[(\vec{w}\cdot\vec{x}_i)+b]$ eşitliği ile hesaplanır. Burada $\vec{w}$ ve b parametrelerini bir sabit ile çarpmak g fonksiyonunu değiştirmeyecektir. Çünkü g fonksiyonu $(\vec{w}\cdot\vec{x}_t)+b$'nin şiddetine değil, sadece işaretine bağlıdır. Ancak, parametreleri bir sabitle çarpmak, fonksiyonel sınırları değiştirecektir. Bu da yanlış sınıflandırmaya neden olacaktır. Yani, anlamlı hiçbir işlem yapmadan, amaçsız yere fonksiyonel uzaklığı değiştirmek akıllıca değildir. Bu problemin üstesinden gelmek için ise fonksiyonel uzaklık yerine, geometrik uzaklık tercih edilir.

TANIM 4.3. (GEOMETRİK UZAKLIK) $(\vec{x}_i, y_i)$ *gibi bir eğitim örneği ile* $\mathfrak{H}$ *hiperdüzlemi arasındaki geometrik uzaklık (normalleştirilmiş ve Öklid) aşağıdaki gibi hesaplanır:*

$$\gamma_i^* = y_i \left[\left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}_i \right) + \frac{b}{\|\vec{w}\|} \right] \tag{4.11}$$

S uzayında bulunan her eğitim örneği ile $\mathfrak{H}$ *hiperdüzlemi arasında hesaplanan tüm geometrik uzaklıkların en küçüğü* γ^* *olmak üzere*

$$\gamma^* = \min_{(\vec{x}_i, y_i) \in S} \gamma_i^* \tag{4.12}$$

dir.

Ayırıcı düzleme ek olarak, bu düzleme paralel iki hiperdüzlem daha tanımlanır. Bu hiperdüzlemler, her iki bölgedeki en yakın eğitim örneklerini kesen düzlemlerdir. Bu düzlemlere "destek hiperdüzlemleri" denir. Çünkü kestikleri eğitim vektörleri bu hiperdüzlemleri desteklemektedir. Bu eğitim örnekleri de "destek vektörleri" adını alır. Bu durumda, $\vec{x}_+^{sv}$ pozitif çıktı alan sınıfta bulunan destek hiperdüzlemin kestiği eğitim vektörü, $\vec{x}_-^{sv}$ ise negatif çıktı alan sınıfta bulunan destek hiperdüzlemin kestiği eğitim vektörü olmak üzere, bu iki destek hiperdüzlemleri için aşağıdaki eşitlikler yazılabilir.

$$\begin{aligned} (\vec{w} \cdot \vec{x}_+^{sv}) + b &= +\gamma \\ (\vec{w} \cdot \vec{x}_-^{sv}) + b &= -\gamma \end{aligned} \tag{4.13}$$

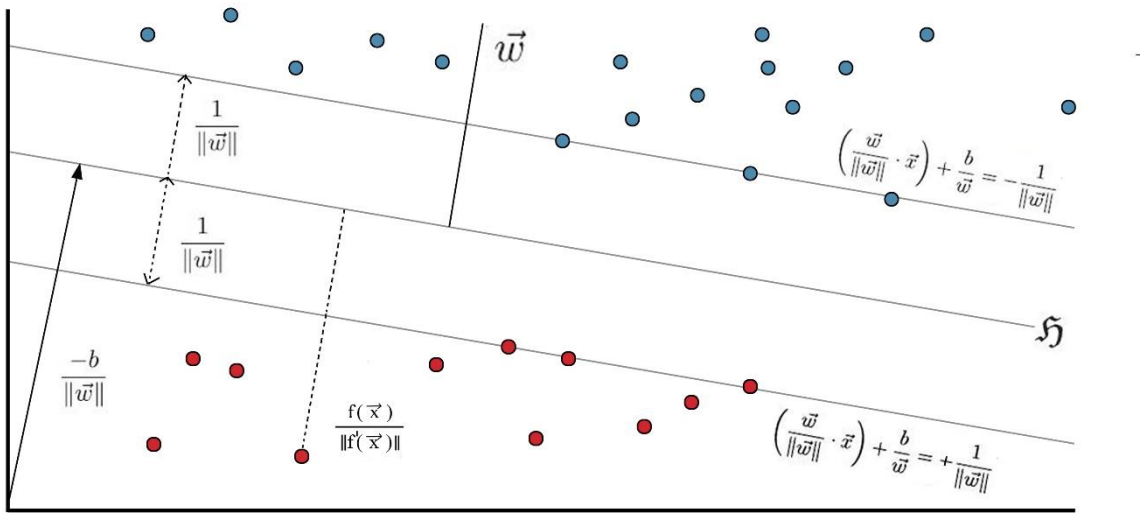
$\vec{w}$ ve b parametreleri $\alpha \in \mathbb{R}$ olacak biçimde herhangi bir sabit ile çarpılsa bile, VC-boyutu değişmeyeceği için, bu eşitlikler her zaman aynı ayırıcı hiperdüzlemi tanımlayacaktır. Yani, hiperdüzlemin pozisyonu veya yönü değişmeyecektir. Ancak, bu durumda, hiperdüzlem tek olmayacaktır. Bu belirsizliği gidermek için, $\gamma = 1$ kabul edilir, böylelikle, bu hiperdüzlem kanonikleştirilir [37]. Bu işlem bir nevi ölçeklendirmedir. İşlem sonunda, her iki sınıfta, destek düzlemlerini en yakın vektörlerin, ayırıcı hiperdüzleme uzaklığı "1" kabul edilir ve basit olarak aşağıdaki gibi gösterilir.

$$\min_{\vec{x}_i \in S} \left| (\vec{w} \cdot \vec{x}_i) + b \right| = 1 \tag{4.14}$$

Bu hiperdüzlem normalleştirildiği takdirde, Eşitlik (4.15) aşağıdaki gibi elde edilir.

$$\min_{\vec{x}_i \in S} \left| \left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}_i \right) + \frac{b}{\|\vec{w}\|} \right| = \frac{1}{\|\vec{w}\|} \tag{4.15}$$

$\left|(\vec{w} \cdot \vec{x}_i) + b\right| = 1$ koşulu sağlayan eğitim veri noktaları "destek vektör" olarak isimnlendirilir. Bu vektörler, $\mathfrak{H}$ ayırıcı hiperdüzlemin her iki kısmında $\mathfrak{H}_+ = \{\vec{x} \,|\, (\vec{w} \cdot \vec{x}_i) + b = +1\}$ ve $\mathfrak{H}_- = \{\vec{x} \,|\, (\vec{w} \cdot \vec{x}_i) + b = -1\}$ destek görevi görürler. Böylelikle $\mathfrak{H}_+$ ve $\mathfrak{H}_-$ destek hiperdüzlemleri arasında oluşan sınırı (margin) tanımlarlar.



Şekil 4.4. Ayırıcı hiperdüzlem ve destek vektörleri

Şekil 4.4'te doğrusal olarak ayrılabilir veri seti için çizilebilecek bir hiperdüzlem örneği görülmektedir. Ayırıcı hiperdüzleme en yakın eğitim vektörleri arasında sabit bir uzaklık vardır. Bunun nedeni Eşitlik (4.14) ile hipotez uzayına sınırlama getirilmesidir. Böylelikle, destek hiperdüzlemlerini kesen ve $\mathfrak{H}$ ayırıcı hiperdüzlemin karşıt taraflarında bu destek vektörleri $\vec{x}_+^{sv}$ ve $\vec{x}_-^{sv}$ arasındaki geometrik uzaklık, bu iki vektörün $\vec{w}$ vektörü üzerine normalleştirilmiş dik izdüşümleri arasındaki fark kullanılarak, aşağıdaki eşitlik ile hesaplanır:

$$\left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}_+^{sv} \right) - \left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}_-^{sv} \right) = \frac{1-b}{\|\vec{w}\|} - \frac{-1-b}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|} \tag{4.16}$$

$\vec{w}$ ağırlık vektörü üzerinde bulunan $\vec{x}_{+}^{sv}$ ve $\vec{x}_{-}^{sv}$ veri noktaları, ayırıcı hiperdüzlemin karşıt taraflarında ve hiperdüzleme en yakın noktalar oldukları için, her birinin hiperdüzleme uzaklıkları eşit ve $\frac{1}{\vec{w}}$ olmalıdır. Ancak, burada dikkat edilmesi gereken nokta, $\gamma = 1$ iken bu geometrik sınırın bulunmasıdır. Daha farklı bir ölçeklendirme ile geometrik sınırı farklı yeni kanonik hiperdüzlemler elde edilebilir.

DVM'nin en son aşaması ise, Eşitlik (4.16) ile verilen sınırın maksimize edilmesidir. Üstesinden gelinmesi en zor olan bu aşamada, oluşacak bir karesel optimizasyon problemi çözülerek, en yüksek genelleştirme potansiyeline sahip ayırıcı bir düzlem bulunur. Optimal ayırıcı düzlem de denilen bu hiperdüzlem karesel programlama probleminin tek çözümüdür. Kısaca, optimal ayırıcı hiperdüzlem, veri kümesini en büyük sınır ile ayırandır [37]. Yani eğitim veri kümesindeki noktaların ayırıcı hiperdüzlemden olabildiğince uzak tutulması anlamındadır.

4.1.3. Sert Sınır Sınıflandırıcıları

Bu bölüme kadar anlatılanlar iki sınıflı doğrusal olarak ayrılabilen veriler için gösterilmiş olsa da, DVM'nin çalışma prensibinin büyük bir kısmını oluşturmaktadır. Bundan sonraki alt bölümlerde eğitim verisinin türü değiştikçe oluşturulması gereken karesel optimizasyon problemleri ayrı ayrı ele alınıp, çözümlemesi yapılacaktır.

Bu veri türlerinden ilki doğrusal olarak ayrılabilen verilerdir. Bu tür veriler için doğrusal karar fonksiyonları çizilebilir. Bulunacak ayırıcı hiperdüzlem ve bu düzlemi destekleyen destek hiperdüzlemleri arasında herhangi bir veri noktasının bulunması istenmemektedir. Bu nedenle bu tür DVM sınıflandırma algoritmasına, *sert sınır sınıflandırıcıları (hard margin classifiers)* denilmektedir. Burada amaç, ayırıcı hiperdüzleme ait sınırı en büyüklemektir. Bu amaçla $\frac{1}{\|\vec{w}\|}$ maksimize edilmelidir. $\frac{1}{\|\vec{w}\|}$ 'i maksimize etmek $\frac{1}{2}\|\vec{w}\|^2$ 'i minimize etmeye eşittir.

Ayırıcı hiperdüzleme en yakın vektörler (yani destek vektörleri) için fonksiyonel sınır, Eşitlik (4.14)'de, geometrik sınır ise Eşitlik (4.15)'de verilmişti. Diğer tüm eğitim örnekleri için aşağıdaki eşitsizlikler benzer şekilde yazılabilir.

$$(\vec{w} \cdot \vec{x}_i) + b \geq +1, \quad y_i = +1 \, için \tag{4.17}$$

$$(\vec{w}\cdot\vec{x}_i)+b\le -1,\quad y_i=-1\,\textit{için} \tag{4.18}$$

Bu iki eşitsizlik, tek bir eşitsizlik olarak,

$$y_i[(\vec{w}\cdot\vec{x}_i)+b]-1\ge 0,\quad i=1,2,...,n \tag{4.19}$$

şeklinde yazılır. Amaç fonksiyonu ve kısıtlar elde edildikten sonra öğrenme problemi,

$$\begin{array}{ll} \min\limits_{\vec{w},b} & \frac{1}{2}\|\vec{w}\|^2 \\ \textit{kısıtlar} & y_i[(\vec{w}\cdot\vec{x}_i)+b]-1\ge 0,\quad i=1,...,n \end{array} \tag{4.20}$$

şeklinde kurulur. Bu problem bir karesel optimizasyon problemidir. Amaç fonksiyonu, ağırlık vektörüne bağlı karesel konveks bir fonksiyondur. Kısıtlar ise doğrusal eşitsizliklerden oluşur. Doğrusal olarak ayrılabilen verilerin sınıflandırılmasında kullanılan bir primal problemdir. Bu primal problem, Lagrange yöntemi kullanılarak çözülebilir. Lagrange yönteminin kullanılmasının iki nedeni vardır: Birincisi Eşitlik (4.19)'daki kısıtlar, Lagrange çarpanlarından oluşan kısıtlarla yer değiştirecektir. Böylelikle çözüme ulaşmak kolaylaşacaktır. İkincisi ise, problemin bu yeni formülünde eğitim verileri, vektörler arası birer nokta çarpımı olarak hesaplanacaktır [40]. Bu, doğrusal olmayan veri türleri için çok önemli bir özelliktir ve bu özellik daha sonraki bölümlerde ayrıntılı olarak anlatılacaktır.

$$\begin{aligned} L_P(\vec{w},b,\alpha) &= \frac{1}{2}\|\vec{w}\|^2-\sum_{i=1}^{n}\alpha_i\left\{y_i[(\vec{w}\cdot\vec{x}_i)+b]-1\right\} \\ &= \frac{1}{2}\|\vec{w}\|^2-\sum_{i=1}^{n}\alpha_i y_i[(\vec{w}\cdot\vec{x}_i)+b]+\sum_{i=1}^{n}\alpha_i \end{aligned} \tag{4.21}$$

Eşitlik (4.21)'deki primal Lagrange fonksiyonunda, $i=1,2,...,n$ olmak üzere α_i 'ler Lagrange çarpanları olarak adlandırılır. Bu Lagrange çarpanları pozitiftir. $L_P(\vec{w},b,\alpha)$ Lagrange fonksiyonu $\vec{w}$ ve b 'ye göre minimize edilir. Aynı zamanda, $\alpha_i\ge 0$ kısıtları altında, bütün α_i 'lere göre $L_P(\vec{w},b,\alpha)$ 'nin türevleri de yok edilmelidir. Burada amaç fonksiyonunun kendisi konveks ve kısıtları sağlayan tüm noktalar bir konveks küme oluşturduğu için, konveks bir karesel optimizasyon problem ini çözmektir. Bu problemi çözmek " *maksimum* L_p " dual problemini çözmek ile eşdeğerdir. Bu durumda $\alpha_i\ge 0$ kısıtları altında $L_P(\vec{w},b,\alpha)$ 'nin tüm gradyanları (yön türevleri) $\vec{w}$ ve b 'e göre sıfırdır. Bu özel dual formülasyona Wolfe dual adı verilir [41] ve aşağıdaki gibi gösterilir:

$$\frac{\partial}{\partial \vec{w}} L_P(\vec{w}, b, \alpha) = 0 \Rightarrow \vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \tag{4.22}$$

$$\frac{\partial}{\partial b} L_P(\vec{w}, b, \alpha) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4.23}$$

Elde edilen bu eşitlikler, Eşitlik (4.21)'de verilen primal Lagrange fonksiyonunda yerine yazılırsa,

$$\begin{aligned}
&\frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{n} \alpha_i y_i [(\vec{w} \cdot \vec{x}_i) + b] + \sum_{i=1}^{n} \alpha_i \\
&\quad = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j - \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0} + \sum_{i=1}^{n} \alpha_i \\
&\quad = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j
\end{aligned}$$

elde edilir. Elde edilen bu eşitliğe dual Lagrange fonksiyonu denir, $L_D(\vec{w}, b, \alpha)$ ile gösterilen dual Lagrange fonksiyonu,

$$L_D(\vec{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \tag{4.24}$$

dır. Bu durumda, Eşitlik (4.20)'de verilen primal optimizasyon probleminin duali aşağıdaki şekilde yazılır.

$$\begin{aligned}
&\underset{\alpha}{maks} \qquad L_D(\vec{w}, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \\
&\qquad\qquad\qquad \alpha_i \geq 0, \quad i = 1, \ldots, n \\
&kısıtlar \qquad\qquad \sum_{i=1}^{n} \alpha_i y_i = 0
\end{aligned} \tag{4.25}$$

Dikkat edilirse, primal problemdeki değişken sayısı $d+1$ (özellik sayısı + 1)'dir. Ancak dual formülasyonda n (eğitim örneklerinin sayısı) tane değişken vardır. Bu nedenle daha az değişkene sahip dual problem tercih edilir.

Dualite teorisine göre, konveks primal problemlerin duali konkavdır. Ek olarak primal problemin tek çözümü, dual problemin tek çözümüne eşittir. Aslında, $L_P(\vec{w}^*) = L_D(\alpha^*)$ 'dır yani, "dualite boşluğu" sıfırdır.

Herhangi bir kısıtlı optimizasyon probleminin (konveks veya değil) en iyi çözümünün Karush-Kuhn-Tucker (KKT) koşullarını [42, 43] sağlaması gerekmektedir. KKT koşulları, bir DVM probleminde $\vec{w}$, b ve α 'da çözüme sahip olunması için gerek ve yeter beş koşullardır. Bu koşulların ilk ikisi primal Lagrange fonksiyonunun $\vec{w}$ ve b 'ye göre türevlerinin sıfır olmasıdır. Diğer iki koşulu ise primal ve dual problemlerin kısıtlarını oluşturmaktadır. Son koşul ise en önemlisidir ve "tümler gevşeklik (complementary slackness)" adı verilir. Bu koşullar sırasıyla aşağıda tanımlanmıştır.

$$\frac{\partial}{\partial \vec{w}} L_P(\vec{w}, b, \alpha) = 0 \Rightarrow \vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i, \quad i = 1, 2, ..., n \tag{4.26}$$

$$\frac{\partial}{\partial b} L_P(\vec{w}, b, \alpha) = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0, \quad i = 1, 2, ..., n \tag{4.27}$$

$$y_i[(\vec{w} \cdot \vec{x}_i) + b] - 1 \geq 0, \quad i = 1, ..., n \tag{4.28}$$

$$\alpha_i \geq 0, \quad i = 1, ..., n \tag{4.29}$$

$$\alpha_i \left\{ y_i[(\vec{w} \cdot \vec{x}_i) + b] - 1 \right\} = 0, \quad i = 1, ..., n \tag{4.30}$$

Eşitlik (4.30) ile verilen tümsel gevşeklik koşuluna göre, eğer α_i veya $y_i[(\vec{w} \cdot \vec{x}_i) + b] - 1$'in herhangi biri pozitif ise, diğeri kesinlikle sıfır olmalıdır. Bu sadece destek hiperdüzlemleri üzerinde bulunan noktalar sıfır olmayan α çarpanına sahip anlamına gelmektedir. Destek hiperdüzlemleri üzerinde bulunan ve ayırıcı hiperdüzlemi destekleyen bu öğrenme noktalarına destek vektörleri dendiği daha önceden belirtilmişti. Bu durumda, diğer tüm öğrenme noktaları için $\alpha_i = 0$ ($i = 1, 2, ..., n$) olacaktır. Yani, bu noktalar dual problemin çözümünü etkilemeyecektir. Bu özelliğe *seyreklik (sparsity)* denir. Seyreklik özelliği, DVM'i, lojistik regresyon gibi geleneksel istatistiksel sınıflandırma algoritmalarından ayıran en önemli özelliktir. Çünkü lojistik regresyon analizinde problemin çözümü eğitim kümesindeki tüm noktalara bağlıdır. Seyreklik özelliği ise sayısal nedenler bakımından istenilen bir özelliktir. Böylelikle tüm eğitim örnekleri (n) ile çalışmak yerine sadece hiperdüzlemi etkileyen daha az sayıda örnekle çalışmaktadır.

KKT koşulları ile destek vektörleri bulunduktan sonra, dual problemde gelecek test örneklerinin çıktılarını hesaplamak için kullanılacak tahmin fonksiyonu Eşitlik (4.22), Eşitlik (4.3)'te yazılarak aşağıdaki gibi bulunur.

$$f(\vec{x}) = (\vec{w} \cdot \vec{x}) + b = \vec{w}^T \vec{x} + b = \left(\sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \right)^T \vec{x} + b = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i^T \vec{x} + b \tag{4.31}$$

KKT koşullarının diğer bir uygulaması b eşik/eğilim değerinin hesaplanmasıdır. Herhangi bir $\vec{x}_j$ destek vektörü için ($\alpha_j > 0$) $y_j \left[(\vec{w} \cdot \vec{x}_j) + b \right] = y_j \vec{w}^T \vec{x}_j + y_j b = 1$ yazılabilir. Böylelikle b değeri, $\frac{1}{y_j} = y_j$ olmak üzere,

$$b = \frac{1 - y_j \vec{w}^T \vec{x}_j}{y_j} = \frac{1}{y_j} - \vec{w}^T \vec{x}_j = \frac{1}{y_j} - \left(\sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \right)^T \vec{x}_j = y_j - \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i^T \vec{x}_j \tag{4.32}$$

şeklinde hesaplanır. Sağlam bir çözüm elde etmek için bu katsayı genellikle tüm destek vektörler üzerinden aşağıdaki gibi hesaplanır:

$$b = \frac{1}{m} \sum_{j=1}^{m} \left(y_j - \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i^T \vec{x}_j \right) \tag{4.33}$$

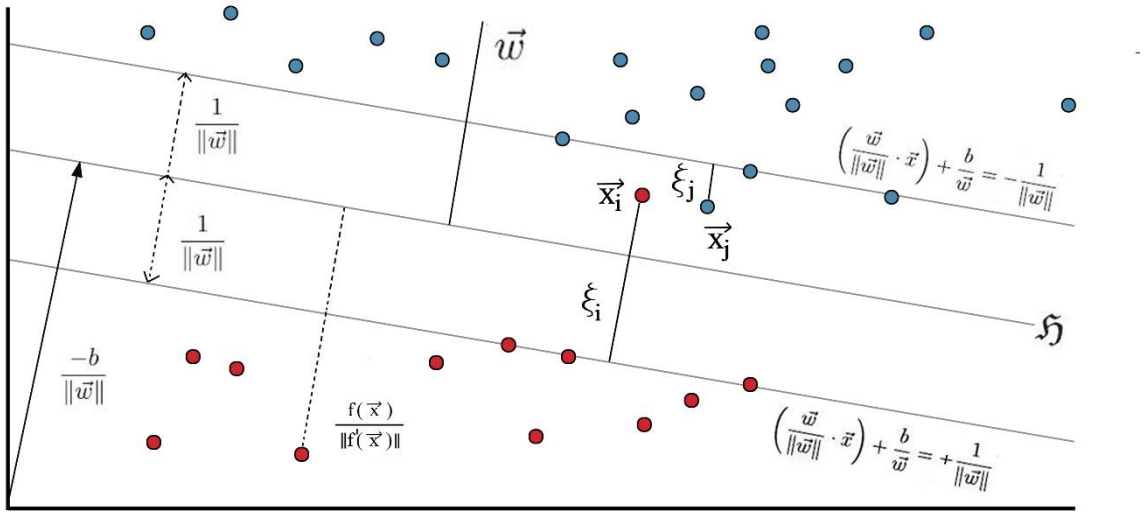
Burada m destek vektörlerinin sayısıdır.

Her konveks optimizasyon problemi için elde edilecek yerel optimal çözüm aynı zamanda globaldir. Bu noktalar yerel minimumlara takılmazlar. Ancak, burada, $\vec{w}$'nin tek çözümü olsa da, α_i katsayıları tek olmak zorunda değildir, yani aynı $\vec{w}$ için farklı sayıda destek vektörü bulunabilir. Bu durumda en az sayıda destek vektörü veren çözüm seçilmelidir. Böylelikle uzay karmaşıklığı ve test aşamasında hesaplama süresi en aza indirgenir.

4.1.4. Yumuşak Sınır Sınıflandırıcıları

Gerçek yaşam problemleri genellikle Şekil 4.4'te görüldüğü gibi kolay ayrılabilir değildir ve doğrusal olarak ayrılabilme varsayımı geçerliliğini yitirir. Verinin doğrusal olarak ayrılabilir olmadığı durumlarda ve/veya sınıflar çakıştığında, dual Lagrange fonksiyonuna ait amaç fonksiyonunun keyfi olarak büyümesinden dolayı optimizasyon problemi için uygun bir çözüm bulunamaz. Hatta kimi zaman verinin *gürültülü (noisy)* olması nedeniyle, veriyi mükemmel şekilde ayıran uygun bir hiperdüzlem bulmak zorlaşır. Bu gibi durumlarda, aşırı uyum denilen kavram gerçekleşir ve bu sınıflandırma algoritması yeni gelecek örnekler üzerinde genelleştirilemeyebilir. Bölüm 4.1.3'te anlatılan sert sınır sınıflandırıcılarının en büyük dezavantajı, sınıflandırma hatasına izin vermemesidir. Bu sınıflandırıcıda ya sınıflandırma hatası olmaksızın bir çözüm bulunur ya da hiç çözüm

bulunamaz. En büyük sınıra sahip hiperdüzlem konseptini genelleştirmek ve doğrusal olarak ayrılamayan verilen için de sınıflandırma yapabilmek için, Eşitlik (4.19)'daki eşitsizlik kısıtı pozitif gevşek değişkenler tanıtılarak rahatlatılır [38]. Bu gevşek değişkenler sayesinde bazı eğitim örneklerinin sınır kısıtlarını ihlal etmesine ve bu örneklerin yanlış sınıflandırılmasına izin verilir. Böylelikle, bazı eğitim verilerinin $\mathfrak{H}_+$ ve $\mathfrak{H}_-$ destek sınırılarını veya $\mathfrak{H}$ sınıflandırma sınırını geçmesi hoş görülür. *Yumuşak sınır sınıflandırıcılarının (soft margin classifiers)* basit bir örneği Şekil 4.5'te gösterildiği gibidir.



Şekil 4.5. Yumuşak sınır sınıflandırıcısı

ξ_i $(i=1,2,...,n)$ pozitif gevşek değişkenler olmak üzere, Eşitlik (4.17) ve Eşitlik (4.18)'de verilen kısıtlar aşağıdaki şekilde yeniden yazılabilir:

$$(\vec{w}\cdot\vec{x}_i)+b\geq+1-\xi_i,\quad y_i=+1\,için \tag{4.34}$$

$$(\vec{w}\cdot\vec{x}_i)+b\leq-1+\xi_i,\quad y_i=-1\,için \tag{4.35}$$

$$\xi_i\geq 0,\quad \forall i$$

Sert sınıf sınıflandırıcılarında olduğu gibi burada da bu iki ayırıcı kısıtlar tek bir eşitsizlik olarak aşağıdaki gibi yazılabilir:

$$y_i\left[(\vec{w}\cdot\vec{x}_i)+b\right]-1+\xi_i \geq 0, \quad i=1,2,...,n \tag{4.36}$$

$$\xi_i \geq 0, \quad \forall i$$

Veri setinde bulunan herhangi bir eğitim örneği için ξ_i gevşek değişkeni 4 farklı değer alabilir.

(i) $\xi_i < 0$: $\vec{x}_i$ doğru sınıflandırılır ve sınırın dışında yer alır.

(ii) $\xi_i = 0$: $\vec{x}_i$ doğru sınıflandırılır ve sınırınüzerinde yer alır.

(iii) $0 < \xi_i \leq 1$: $\vec{x}_i$ doğru sınıflandırılır ve sınıflandırma sınırı ile kendi sınıfına ait sınır arasında yer alır.

(iv) $\xi_i > 1$: $\vec{x}_i$ yanlış sınıflandırılır ve sınıflandırma sınırının yanlış tarafindadır.

ξ_i değişkenleri bu kısıtların bozulmasına izin verir. Ancak, bu bozulmalar için optimizasyon problemindeki amaç fonksiyonu cezalandırılmalıdır. Aksi takdirde bu kısıtlar geçerliliğini yitirecektir Çünkü araştırmacı, ξ_i değişkenlerini çok büyük seçerek sınıflandırmaya gidebilir. Bu durum analiz sonuçlarının güvenilir olmamasına yol açar. Bu nedenle, optimizasyon problemindeki amaç fonksiyonuna $C\left(\sum_{i=1}^{n}\xi_i\right)^k$ formundaki bir ceza fonksiyonu eklenir. Burada C katsayısı, ceza ve sınır arasında denge görevi gören düzenleme katsayısı olarak geçer ve kullanıcı tarafından belirlenir. Bu parametre için olması gerekenden çok küçük veya çok büyük değerler seçilmesi durumunda optimal hiperdüzlem doğru belirlenemeyeceğinden sınıflandırma doğruluğunda ciddi düşüş beklenir. k pozitif tamsayıları için, amaç fonksiyonu hala konveks bir fonksiyondur. $k=1,2$ için bir karesel optimizasyon problemi elde edilir. Burada, avantajlı olması nedeniyle $k=1$ olarak seçilir. L^1-norm problemi de denilen bu problemde ne ξ_i gevşek değişkenleri ne de herhangi bir Lagrange çarpanı, Wolfe dual probleminde görülür [40].

Ayırıcı hiperdüzlemin yanlış tarafında yer alan veri noktaları için $\xi_i > 1$ olacaktır. Böylelikle, $\sum_i \xi_i$ toplamı bozulmalarının ne kadar "kötü" olduğunu gösteren ve bozulmaların sayısının bir üst sınırı olan bir ölçek olarak yorumlanabilir.

Tüm bu bilgiler ışığında doğrusal eşitsizlik kısıtları ile karesel amaç fonksiyonuna sahip yeni primal problem aşağıdaki şekilde yazılabilir.

$$\begin{aligned} \min_{\vec{w},b,\xi} \quad & \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \textit{kısıtlar} \quad & y_i\left[(\vec{w}\cdot\vec{x}_i)+b\right]-1+\xi_i \geq 0, \quad i=1,2,...,n \\ & \xi_i \geq 0, \quad \forall i \end{aligned} \tag{4.37}$$

Bu primal probleme ait primal Lagrange fonksiyonu ise,

$$L_P(\vec{w},b,\xi,\alpha,\mu) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left\{y_i\left[(\vec{w}\cdot\vec{x}_i)+b\right]-1+\xi_i\right\} - \sum_{i=1}^{n}\mu_i\xi_i \tag{4.38}$$

şeklindedir. Eşitlik (4.38)'de, α_i'ler ve μ_i'ler Lagrange çarpanlarıdır ($\alpha_i \geq 0$ ve $\mu_i \geq 0$). Primal değişkenlere ($\vec{w}$, b ve ξ_i) göre bu Lagrange fonksiyonunun türevleri sıfırdır.

$$\frac{\partial}{\partial b}L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{4.39}$$

$$\frac{\partial}{\partial \vec{w}}L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow \vec{w} = \sum_{i=1}^{n}\alpha_i y_i \vec{x}_i \tag{4.40}$$

$$\frac{\partial}{\partial \xi_i}L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \tag{4.41}$$

Elde edilen bu eşitlikler, Eşitlik (4.38)'te verilen primal Lagrange fonksiyonunda yerine yazılırsa,

$$\begin{aligned} & \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left\{y_i\left[(\vec{w}\cdot\vec{x}_i)+b\right]-1+\xi_i\right\} - \sum_{i=1}^{n}\mu_i\xi_i \\ & = \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j + \sum_{i=1}^{n}\xi_i\underbrace{(C-\alpha_i-\mu_i)}_{0} + \sum_{i=1}^{n}\alpha_i - \underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{0} b - \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \\ & = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \end{aligned}$$

elde edilir. Elde edilen bu eşitliğe dual Lagrange fonksiyonu denir ve aşağıdaki gibi yazılabilir:

$$L_D(\vec{w},b,\xi,\alpha,\mu) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j \tag{4.42}$$

O halde, Eşitlik (4.37)'de verilen primal optimizasyon probleminin duali aşağıdaki şekilde yazılır:

$$\begin{aligned} \underset{\alpha}{maks} \quad & L_D(\vec{w},b,\xi,\alpha,\mu) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^T \vec{x}_j \\ & 0 \le \alpha_i \le C, \quad i = 1,...,n \\ kısıtlar \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{4.43}$$

Yumuşak sınır sınıflandırıcıları için çözülmesi gereken bu karesel programlama, sert sınır sınıflandırıcıları için kullanılan amaç fonksiyonuna benzerdir. Ancak burada, $0 \le \alpha_i \le C$ gibi bir ekstra kısıt daha vardır. Bu kısıta çoğu zaman "*kutu kısıtı (box constraint)*" adı verilir. Bunun nedeni, C değerinin, α_i çarpanlarını sınırlı bir bölge içinde tutmasıdır. Bu üst sınır $\alpha_i \ge 0$ ve $\mu_i \ge 0$ iken $C - \alpha_i - \mu_i = 0$ eşitliğinden gelir.

Sert sınıf sınıflandırıcılarında olduğu gibi burada da bu optimizasyon probleminin KKT koşullarını sağlaması gerekmektedir. İlk üç koşul, primal problemin primal değişkenlere göre türevlerinin sıfır olmasıdır. Sonraki dört koşulu kıstılar ve çarpan koşulları oluşturur. Son iki koşul ise tümsel gevşeklik koşullarıdır.

$$\frac{\partial}{\partial b} L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{4.44}$$

$$\frac{\partial}{\partial \vec{w}} L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow \vec{w} = \sum_{i=1}^{n}\alpha_i y_i \vec{x}_i \tag{4.45}$$

$$\frac{\partial}{\partial \xi_i} L_P(\vec{w},b,\xi,\alpha,\mu) = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \tag{4.46}$$

$$y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right] - 1 + \xi_i \ge 0 \tag{4.47}$$

$$\varepsilon_i \ge 0 \tag{4.48}$$

$$\alpha_i \ge 0 \tag{4.49}$$

$$\mu_i \ge 0 \tag{4.50}$$

$$\alpha_i\left\{y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right] - 1 + \xi_i\right\} = 0 \tag{4.51}$$

$$\mu_i \xi_i = 0 \tag{4.52}$$

Yukarıdaki kısıtlardan bazı çıkarsamalar yapılabilir. Eğer $\xi_i > 0$ ise $\mu_i = 0$'dır, böylelikle $\alpha_i = C$'dir. Bu durumda $\xi_i = 1 - y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right]$'dir. Ayrıca, $\xi_i = 0$ ise $\mu_i > 0$'dır ve bu durumda, $\alpha_i < C$'dir. Ek olarak, $\xi_i = 0$ olduğu zaman, $y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right] = 1$'dir, böylelikle, $\alpha_i > 0$'dır. Aksi takdirde, $y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right] - 1 > 0$ ise $\alpha_i = 0$'dır. Özetle, destek hiperdüzlemleri üzerinde olmayan ancak doğru bölgede bulunan veri noktaları için $\xi_i = \alpha_i = 0$'dır. Destek hiperdüzlemi üzerinde ve doğru bölgede bulunan noktalar için ise $\xi_i = 0$ ve $\alpha_i > 0$'dır. Son olarak, ayırıcı hiperdüzlemin yanlış bölgesinde bulunan veri noktaları için α_i en büyük C değerini alır ve ξ_i, $y_i\left[(\vec{w}\cdot\vec{x}_i) + b\right] - 1 + \xi_i = 0$ eşitliği ile kısıt bozulumlarını dengeler.

Yumuşak sınır sınıflandırıcılarına ait karar fonksiyonu aşağıdaki gibidir:

$$f(\vec{x}) = (\vec{w}\cdot\vec{x}) + b = \vec{w}^T\vec{x} + b = \left(\sum_{i=1}^{n}\alpha_i y_i \vec{x}_i\right)^T \vec{x} + b = \sum_{i=1}^{n}\alpha_i y_i \vec{x}_i^T\vec{x} + b \qquad (4.53)$$

Eşitlik (4.53)'teki b değeri, sert sınıf sınıflandırıcılarında olduğu gibi KKT koşullarının yardımıyla destek vektörleri kullanılarak hesaplanır.

4.1.5. Doğrusal Olmayan Sınıflandırıcılar

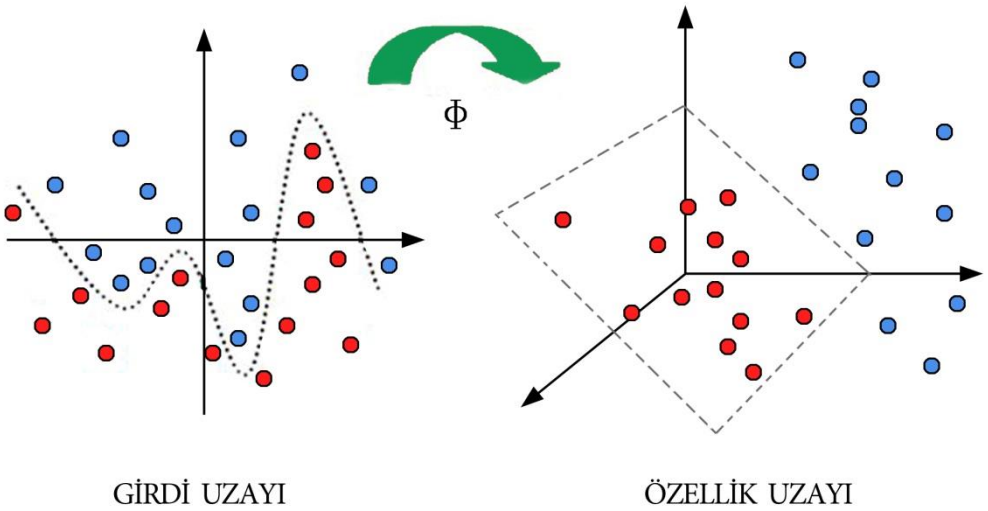
Destek vektör sınıflandırıcıları veriyi ayırmak için doğrusal karar fonksiyonlarını kullanırlar. Ancak, kimi zaman bu karar fonksiyonları eğitim veri kümesinin doğrusal bir fonksiyonu olarak yazılamayabilir. Bu durumda hem yumuşak sınır sınıflandırıcılarının hem de sert sınır sınıflandırıcılarının doğrusal olmayan karar fonksiyonlarına sahip veriler için genelleştirilmesi gerekmektedir. Bu nedenle Boser, Guyon ve Vapnik [44] sundukları çalışmalarında, Aizerman ve arkadaşları [45] tarafından yayınlanan bir makaledeki eski bir hilenin bu genelleştirmeyi kolay ve açık bir şekilde yapabileceğini göstermişlerdir.

Buradaki ana fikir, girdi verilerini yüksek boyutlu bir uzaya haritalamak ve bu uzayda doğrusal bir sınıflandırıcı kullanarak veriyi ayırmaktır. Bu işlem girdi uzayında *doğrusal olmayan bir sınıflandırıcı (nonlinear classifiers)* sonucunu doğurur.

Eğitim verilerinin gösterimi, veriyi, sonlu boyutlu bir Hilbert uzayına (H) haritalayarak değiştirilebilir. Genellikle, bu H uzayı, eğitim verisinin tanımlandığı orijinal girdi uzayı olan $\mathbb{R}^d$ uzayından daha yüksek boyutludur (Burada, $\mathbb{R}^d$ 'düşük boyutlu' uzay, H

‘yüksek boyutlu’ uzay olarak adlandırılır). Eğitimden önce her veri örneğine $\phi:\mathbb{R}^d \to H$ haritalaması uygulanır ve optimal ayırıcı hiperdüzlem H uzayında oluşturulur. Burada H uzayına, özellik uzayı ve $\phi(\vec{x})$ ’in i’ninci bileşenine ise ϕ haritalaması altında i’ninci özellik denir. Veriye en uygun gösterimin seçilmesi ise *özellik seçimi (feature selection)* olarak bilinir.

Hem sert sınıf hem de yumuşak sınır sınıflandırıcılarında, optimizasyon problemlerinin çözümü, eğitim örneklerinin doğrusal bileşeni olarak yazılabilen bir $w \in H$ vektörüdür. Öğrenme görevinin dual formülasyonlarında da görüleceği üzere, algoritma, veri noktalarının nokta çarpımlarına $\left(\vec{x}_i \cdot \vec{x}_j\right) = \vec{x}_i^T \vec{x}_j$ bağlıdır. Doğrusal karar fonksiyonları ile ayrılamayan veriler için ise, veri yüksek boyutlu bir uzaya haritalandığında, öğrenme algoritması H uzayında tanımlanan nokta çarpımlarına bağlıdır, yani $\phi(\vec{x}_i)\cdot\phi(\vec{x}_j)$ formundaki ikili çarpımlar üzerinde yapılacak işlemler sonucu çözüm elde edilir. K gibi bir “çekirdek fonksiyonu” $K(\vec{x}_i,\vec{x}_j) = \left(\phi\left(\vec{x}_i\right)\cdot\phi\left(\vec{x}_j\right)\right) = \phi\left(\vec{x}_i\right)^T \phi\left(\vec{x}_j\right)$ şeklinde tanımlandığında, öğrenme algoritmasında sadece K fonksiyonu kullanılır ve böylelikle ϕ haritalama fonksiyonunun ne olduğunun açıkça bilinmesine gerek yoktur.



Şekil 1.6. ϕ dönüşüm operatörü ile haritalama

Tüm bu bilgiler ışığında çekirdek tabanlı sınıflandırıcılara ait optimizasyon problemleri yazılabilir. Burada yapılan işlem, yüksek boyutlu uzaya çekirdek tabanlı haritalama ile

optimal ayırıcı hiperdüzlem fikrinin kombine edilmesidir. Şekil (4.6)'da da görüleceği üzere çekirdek tabanlı bir DVM algoritması, orijinal girdi uzayında değerlendirilen yüksek boyutlu en büyük sınırlı bir hiperdüzlemdir. Bu bölümde, düşük boyutlu girdi uzayında elde edilen sonuçlar yüksek boyutlu uzayda yeniden özetlenmiştir. Tek fark ise formülasyonlarda, özellik uzayında tanımlı ve bir çekirdek fonksiyonuyla üretilen ikili iç çarpımların kullanılmasıdır.

Sert Sınır Sınıflandırıcıları. Hiçbir öğrenme hatasına izin vermeyen bir DVM algoritması inşa etmek için, aşağıdaki optimizasyon probleminin çözülmesi gerekmektedir.

$$\begin{aligned} \underset{\alpha}{maks} \qquad & L_D(\vec{w},b,\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \phi\left(\vec{x}_i\right)^T \phi\left(\vec{x}_j\right) \\ & = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\vec{x}_i,\vec{x}_j\right) \\ & \alpha_i \geq 0, \quad i=1,\ldots,n \\ kısıtlar \qquad & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{4.54}$$

Yumuşak Sınır Sınıflandırıcıları. Yüksek uzayda doğrusal olarak ayrılamayan veri kümeleri için belirli bir hatayı göze alan bir DVM algoritması inşa etmek için aşağıdaki optimizasyon problemi çözülmesi gerekir.

$$\begin{aligned} \underset{\alpha}{maks} \qquad & L_D(\vec{w},b,\xi,\alpha,\mu) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \phi\left(\vec{x}_i\right)^T \phi\left(\vec{x}_j\right) \\ & = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(\vec{x}_i,\vec{x}_j\right) \\ & 0 \leq \alpha_i \leq C, \quad i=1,\ldots,n \\ kısıtlar \qquad & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{4.55}$$

Çekirdek tabanlı sınıflandırıcılarına ait karar fonksiyonu aşağıdaki gibidir.

$$f(\vec{x}) = \sum_{i=1}^{n}\alpha_i y_i \phi(\vec{x}_i)^T \phi(\vec{x}) + b = \sum_{i=1}^{n}\alpha_i y_i K(\vec{x}_i,\vec{x}) + b \tag{4.56}$$

Çekirdek fonksiyonlarının getirdiği bir takım zorluklar da vardır. Uygun çekirdek türünün ve çekirdek parametresinin seçilmesi analizin güvenirliliği açısından oldukça önemlidir.

Çoğunlukla farklı tür çekirdekler, değişik parametrelerle, ayrı bir *geçerlilik kümesi (validation set)* üzerinde denenerek sınıflandırıcıların performansları karşılaştırılır ve en iyi sonucu veren parametreler eğitim kümesi üstünde çalıştırılarak hataların ölçümü yapılır.

4.2. Destek Vektör Regresyonu

DVM, alternatif bir kayıp fonksiyonu yardımıyla bağlanım yani bağlanım problemleri için kullanılmaktadır [46, 47]. En büyük sınır algoritmasını karakterize eden tüm temel özellikler burada da korunur. Doğrusal olmayan bir fonksiyon, çekirdek tabanlı özellik uzayında bir doğrusal öğrenme makinesi tarafından tahmin edilmeye çalışılır. Aynı zamanda, sistemin kapasitesi, uzayın boyutuna bağlı olmayan bir parametre tarafından kontrol edilir. Sınıflandırma probleminde olduğu gibi, bağlanım analizinde de öğrenme algoritması konveks bir fonksiyoneli minimize etmeye çalışır. Bu işlem sonucunda da elde edilen çözüm seyrektir.

Sınıflandırma probleminde olduğu gibi, $P(\vec{x}, y)$ gibi bilinmeyen bir olasılık dağılımından türetilmiş, $\mathbb{R}^d$ uzayında tanımlı n tane girdi vektörü ve bu girdi vektörlerine karşılık gelen reel değerler alabilen çıktılar aşağıdaki gibi gösterilebilir.

$$S = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_n, y_n)\}, \quad \vec{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

Bu veri kümesini tanımlayan fonksiyonların bir kümesi,

$$F = \{ f \mid f : \mathbb{R}^d \to \mathbb{R} \}$$

olduğunda, en temel problem aşağıdaki risk fonksiyonelini minimize eden $f \in F$ gibi bir fonksiyon bulmaktır.

$$R_X[f] = \int_{X \times Y} \ell(y - f(\vec{x}), \vec{x}) dP(\vec{x}, y) = \int_X \int_Y \ell(y - f(\vec{x}), \vec{x}) dP(y \mid \vec{x}) dP(\vec{x}) \qquad (4.57)$$

Burada, ℓ bir kayıp fonksiyonudur ve y ile $f(\vec{x})$ arasındaki farkın ne kadar cezalandırılması gerektiğini gösterir. $P(\vec{x}, y)$ bilinmediğinden $R_X[f]$ beklenen riski hesaplanamaz, bunun yerine aşağıdaki deneysel risk hesaplanır:

$$R_{deney}[f] = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i - f(\vec{x}_i), \{\vec{x}_i, y_i\}\right) \qquad (4.58)$$

Deney riskin hesaplanmasıyla, risk R, $R_{deney}[f]+R_{gen}[f]$ olacak biçimde sınırlandırılır. Burada, $R_{gen}[f]$, kullanılan veriyi en iyi şekilde tahmin edebilen fonksiyonlara bağlı genelleştirme hatasının üst sınırıdır [48].

Sınıflandırma problemlerinde olduğu gibi, burada da, elde edilecek tahmin fonksiyonunun eğitim veri kümesine ne kadar iyi uyum sağladığı veya yeni gözlemleri ne kadar iyi tahmin ettiği bir kayıp fonksiyonu yardımıyla bulunabilir. Bağlanım analizinde kullanılan 4 temel kayıp fonksiyonu vardır ve aşağıdaki gibi tanımlanır [6]:

(i) *Karesel Kayıp Fonksiyonu (Quadratic Loss Function):* En çok kullanılan kayıp fonksiyonudur. Kimi zaman L_2 kayıp fonksiyonu olarak da adlandırılır.

$$\ell(f,\{\vec{x},y\}) = \ell(y-f(\vec{x})) = (y-f(\vec{x}))^2$$

(ii) *Mutlak Kayıp Fonksiyonu (Absolute Loss Function):* L_1 kayıp fonksiyonu olarak da bilinir ve aşağıdaki gibi tanımlanır:

$$\ell(f,\{\vec{x},y\}) = \ell(y-f(\vec{x})) = |y-f(\vec{x})|$$

(iii) *Huber Kayıp Fonksiyonu (Huber Loss Function):* L_1 kayıp fonksiyonunun aksine her yerde türevlenebilirdir ve

$$\ell(f,\{\vec{x},y\}) = \ell(y-f(\vec{x}))_\varepsilon = \begin{cases} \frac{1}{4\varepsilon}(y-f(\vec{x}))^2, & |y-f(\vec{x})| \le 2\varepsilon \\ |y-f(\vec{x})|-\varepsilon, & \textit{öteki değerler} \end{cases}$$

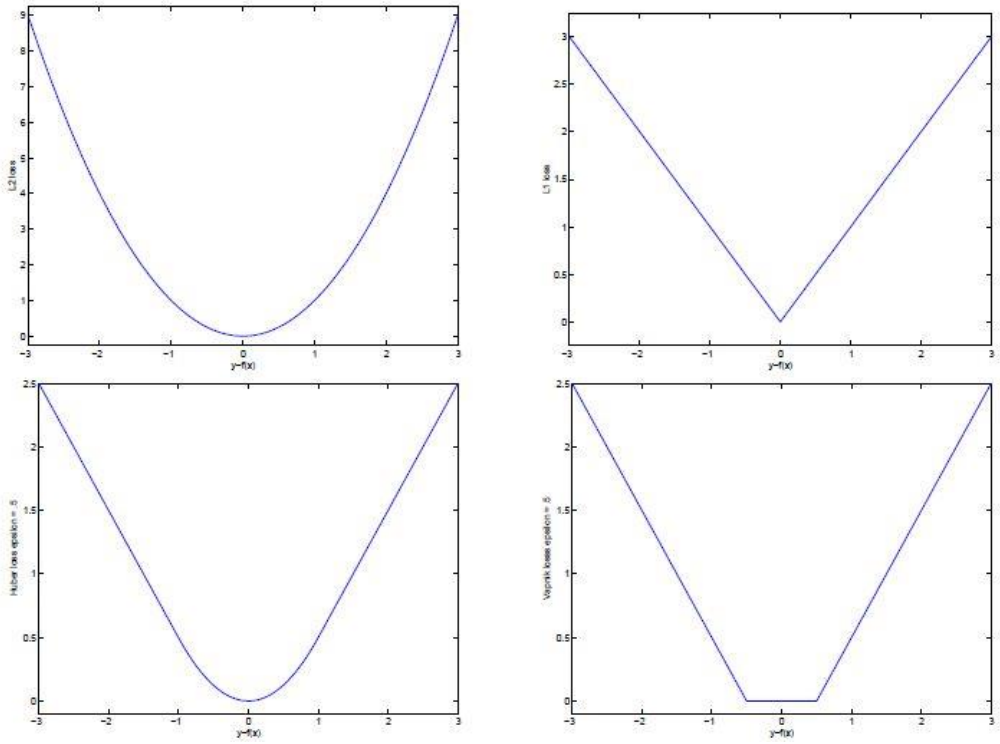
şelindedir.

(iv) *Vapnik Kayıp Fonksiyonu:* Vladamir Naumovich Vapnik tarafından önerilen bir kayıp fonksiyonudur [36]. ε-duyarsız kayıp fonksiyonu olarak da bilinir ve

$$\ell(f,\{\vec{x},y\}) = \ell(y-f(\vec{x}))_\varepsilon = \begin{cases} 0, & |y-f(\vec{x})| \le \varepsilon \\ |y-f(\vec{x})|-\varepsilon, & \textit{öteki değerler} \end{cases}$$

şeklinde tanımlanır.

Şekil 4.7'de bağlanım analizinde kullanılan kayıp fonksiyonları gösterilmektedir.



Şekil 4.7. Bağlanım problemleri için kullanılan bazı kayıp fonksiyonları

Şekil 4.7(a)'da gösterilen karesel kayıp fonksiyonu, geleneksel en küçük kareler hata kriterine karşılık gelmektedir. Şekil 4.7(b)'deki mutlak kayıp fonksiyonu, karesel kayıp fonksiyonuna göre aykırı değerlere daha az duyarlıdır. Şekil4.7(c)'de gösterilen Huber kayıp fonksiyonu, Peter Huber (1964) [49] tarafından verinin türetildiği dağılım bilinmediğinde, kayıp fonksiyonunun *sağlam (robust)* ve optimal özelliklere sahip olması için önerilmiştir. Ancak bu üç kayıp fonksiyonunun hiç biri destek vektörleri açısından seyrekliğe neden olmamaktadır. Bu problemin üstesinden gelmek için, Vapnik, Huber kayıp fonksiyonuna bir yaklaşım olarak, elde edilen destek vektörlerinin seyrek bir kümesinin oluşmasına izin veren ε -duyarsız kayıp fonksiyonunu önermiştir [50]. ε -duyarsız kayıp fonksiyonu kullanıldığından bağlanım çözümlemesi için inşa edilen DVM algoritmasına, $\varepsilon-$ Destek Vektör Regresyonu veya kısaca *Destek Vektör Regresyonu (Support Vector Regression)* adı verilir [36].

DVR'de4 amaç, tüm eğitim örneklerinin çıktıları olan y_i'leri en çok ε kadar bir sapmayla tahmin edecek ve aynı zamanda mümkün olabilecek en düz yani karmaşıklığı en az bir $f(\vec{x})$ fonksiyonunu bulmaktır. Diğer bir ifadeyle, ε gibi bir değerden küçük olduğu sürece hatalarla ilgilenilmez ancak bu değerden yüksek sapmalar da kabul edilemez. Örneğin, herhangi bir para birimi için art arda günlerde ölçülen döviz kuru değerleri ve bu değerleri tanımlayan bir kaç ekonometrik göstergelere sahip olunsun. Döviz kuru ile ilgilenildiğinde, yukarıda anlatılan kavram, ε gibi bir miktardan fazla para kaybetmek istemeyenler için önemlidir.

Sınıflandırma probleminde olduğu gibi $\mathbb{R}^d$ uzayında tanımlı bir doğrusal fonksiyon tanımlansın. $\vec{w} \in \mathbb{R}^d$ ağırlık vektörü ve $b \in \mathbb{R}$ eşik/eğilim değeri olmak üzere, toplamda $d+1$ tane serbest değişkene sahip bu doğrusal fonksiyonu aşağıdaki gibi yazılabilir:

$$f(\vec{x}) = (\vec{w} \cdot \vec{x}) + b \tag{4.59}$$

Düz bir fonksiyon elde etmek için $\vec{w}$ değeri küçük olmalıdır. Bunu sağlamanın bir yolu $\|\vec{w}\|^2 = \langle w, w \rangle = (w \cdot w)$ 'i minimize etmektir. O halde, bu problem, bir konveks optimizasyon problemi olarak yazılabilir:

$$\begin{array}{ll} \min\limits_{\vec{w},b} & \frac{1}{2}\|\vec{w}\|^2 \\ \textit{kısıtlar} & \begin{array}{l} y_i - (\vec{w} \cdot \vec{x}_i) - b \le \varepsilon \\ (\vec{w} \cdot \vec{x}_i) + b - y_i \le \varepsilon \end{array}, \quad i = 1, \ldots, n \end{array} \tag{4.60}$$

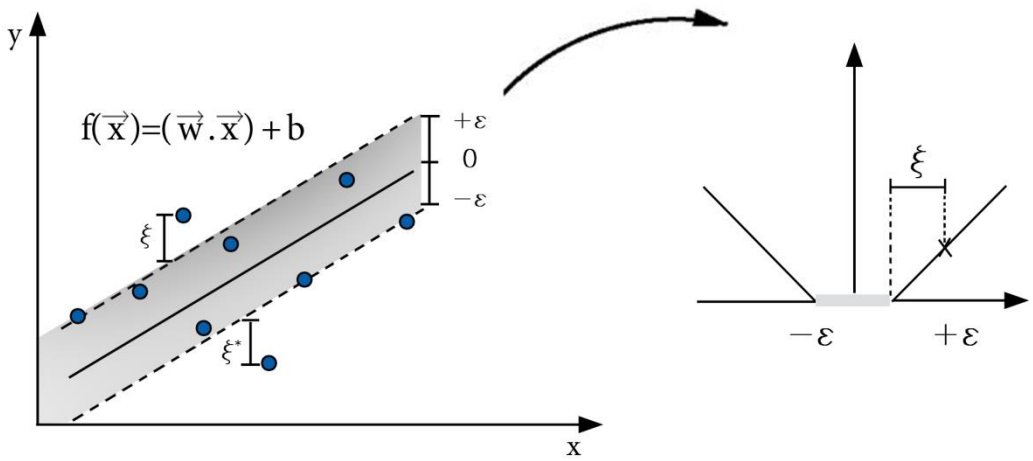
Burada açıkça belirtilmemiş olan bir varsayım, tüm $(\vec{x}_i, y_i)$ çiftlerini ε gibi bir hassasiyetle tahmin eden f gibi bir fonksiyonun mevcut olmasıdır. Yani, bu konveks optimizasyon problemi uygulanabilirdir. DVR, doğrusal bir fonksiyon ve onu ε gibi bir uzaklıkla çevreleyen bir tüp olarak düşünüldüğünde, veri noktaları her zaman bu tüp içine düşmeyebilir. Bu ε tüpüne, tolerans sınırı da denebilir. Bu durumda bazı hataların gerçekleşmesine göz yumulur. Bu nedenle, yumuşak sınır sınıflandırıcılarına benzer olarak, ξ_i ve ξ_i^* gibi negatif olmayan gevşek değişkenler tanımlanır ve bu optimizasyon probleminin uygun bir çözüme sahip olması sağlanır. Böylelikle aşağıdaki optimizasyon problemi çözülmeye çalışılır:

$$
\begin{aligned}
\min_{\vec{w},b,\xi} \quad & \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \\
& y_i - (\vec{w}\cdot\vec{x}_i) - b \le \varepsilon + \xi_i, \quad i = 1,\ldots,n \\
\textit{kısıtlar} \quad & (\vec{w}\cdot\vec{x}_i) + b - y_i \le \varepsilon + \xi_i^*, \quad i = 1,\ldots,n \\
& \xi_i, \xi_i^* \ge 0, \quad \forall i
\end{aligned}
\tag{4.61}
$$

$C > 0$ parametresi, f tahmin fonksiyonunun düzlüğü ve ε 'dan ne kadar büyük sapmaların kabul edileceği bilgisi arasında takas görevi gören bir parametredir. Bu parametre kullanıcı tarafından belirlenir. Bu bilgiler ışığında ε-duyarsız kayıp fonksiyonu aşağıdaki gibi tanımlanabilir.

$$
|\xi|_\varepsilon = \begin{cases} 0 \quad , & |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & \textit{öteki değerler} \end{cases}
\tag{4.62}
$$

Şekil 4.8'de doğrusal bir DVR grafiksel olarak görülebilir.



Şekil 4.8. Doğrusal DVR için yumuşak sınır kayıp düzeni

Burada sadece taralı bölge dışındaki noktalar kayıp fonksiyonunda maliyet yaratır ve tüm bu sapmalar doğrusal bir şekilde cezalandırılır. Sınıflandırma problemlerinde olduğu gibi, çoğu durumda, optimizasyon problemi dual olarak yazıldığında kolay bir şekilde çözülebilir. Dual formülasyonun kullanılmasının diğer bir sebebi ise, bu doğrusal bir DVM algoritmasını doğrusal olmayan fonksiyonlarla da kullanılmasını sağlamaktır. O halde, Eşitlik (4.61)'de verilen primal optimizasyon probleminin Lagrange fonksiyonu,

$$L_P(\vec{w},b,\xi,\alpha,\eta)=\frac{1}{2}\|\vec{w}\|^2+C\sum_{i=1}^{n}\left(\xi_i+\xi_i^*\right)-\sum_{i=1}^{n}(\eta_i\xi_i+\eta_i^*\xi_i^*)$$
$$-\sum_{i=1}^{n}\alpha_i\left(\varepsilon+\xi_i-y_i+(\vec{w}\cdot\vec{x}_i)+b\right) \tag{4.63}$$
$$-\sum_{i=1}^{n}\alpha_i^*\left(\varepsilon+\xi_i^*+y_i-(\vec{w}\cdot\vec{x}_i)-b\right)$$

şeklindedir. Burada η_i, η_i^*, α_i ve α_i^*'ler Lagrange çarpanlarıdır. Bu dual değişkenlerin pozitiflik kısıtını sağlaması gerekmektedir, yani $\eta_i>0$, $\eta_i^*>0$, $\alpha_i>0$ ve $\alpha_i^*>0$. Optimalite için primal değişkenlere ($\vec{w}$, b, ξ_i ve ξ_i^*) göre bu Lagrange fonksiyonunun kısmi türevleri yok edilmelidir.

$$\frac{\partial}{\partial b}L_P(\vec{w},b,\xi,\alpha,\eta)=0\Rightarrow\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0 \tag{4.64}$$

$$\frac{\partial}{\partial \vec{w}}L_P(\vec{w},b,\xi,\alpha,\eta)=0\Rightarrow\vec{w}=\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)\vec{x}_i \tag{4.65}$$

$$\frac{\partial}{\partial \xi_i}L_P(\vec{w},b,\xi,\alpha,\eta)=0\Rightarrow C-\alpha_i-\eta_i=0 \tag{4.66}$$

$$\frac{\partial}{\partial \xi_i^*}L_P(\vec{w},b,\xi,\alpha,\eta)=0\Rightarrow C-\alpha_i^*-\eta_i^*=0 \tag{4.67}$$

Eşitlik (4.64)-(4.67), primal Lagrange fonksiyonunda yerine yazılırsa, dual optimizasyon problemi aşağıdaki gibi elde edilir:

$$\begin{aligned}
&\underset{\alpha}{maks} \qquad L_D(\vec{w},b,\xi,\alpha,\eta)=\begin{cases}-\frac{1}{2}\sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)\vec{x}_i^T\vec{x}_j\\ -\varepsilon\sum_{i=1}^{n}\left(\alpha_i+\alpha_i^*\right)+\sum_{i=1}^{n}y_i\left(\alpha_i-\alpha_i^*\right)\end{cases}\\
&kısıtlar \qquad \sum_{i=1}^{n}\left(\alpha_i-\alpha_i^*\right)=0\\
&\qquad\qquad\qquad 0\le\alpha_i,\alpha_i^*\le C,\quad\forall i
\end{aligned} \tag{4.68}$$

Dual optimizasyon probleminin elde edilmesiyle, dual değişkenler η_i ve η_i^*, Eşitlik (4.66) ve Eşitlik (4.67) yardımıyla $\eta_i^{(*)}=C-\alpha_i^{(*)}$ şeklinde yok edilmiş oldu. Böylelikle Eşitlik (4.65) yardımıyla DVR'ye ait karar fonksiyonu aşağıdaki gibi yazılır:

$$f(\vec{x}) = (\vec{w} \cdot \vec{x}) + b = \vec{w}^T \vec{x} + b = \left(\sum_{i=1}^{n} \left(\alpha_i - \alpha_i^* \right) \vec{x}_i \right)^T \vec{x} + b = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^* \right) \vec{x}_i^T \vec{x} + b$$

Destek vektör genişlemesi denilen bu durumda, $\vec{w}$ vektörü, $\vec{x}_i$ eğitim noktalarının doğrusal bir birleşimi şeklinde yazılabilir. Dual problemin yardımıyla, fonksiyonun karmaşıklığının destek vektörleri ile gösteriminin, girdi uzayı X 'in boyutundan bağımsız ve sadece destek vektörlerin sayısına bağlı olduğu görülebilir.

Karar fonksiyonundaki b katsayı ise, KKT koşulları sayesinde bulunur. Çözüm noktasında dual değişkenler ile kısıtların çarpımı yok edilmelidir.

$$\alpha_i \left(\varepsilon + \xi_i - y_i + (\vec{w} \cdot \vec{x}_i) + b \right) = 0 \tag{4.69}$$

$$\alpha_i^* \left(\varepsilon + \xi_i^* + y_i - (\vec{w} \cdot \vec{x}_i) - b \right) = 0 \tag{4.70}$$

$$(C - \alpha_i)\xi_i = 0 \tag{4.71}$$

$$(C - \alpha_i^*)\xi_i^* = 0 \tag{4.72}$$

Eşitlikler (4.69) – (4.72)'den şu çıkarımlar yapılabilir. (i) Sadece $\alpha_i^{(*)} = C$ olan $(\vec{x}_i, y_i)$ noktaları f bağlanım fonksiyonunu çevreleyen ε-duyarsız tüpünün dışında yer alır. (ii) $\alpha_i \alpha_i^* = 0$ 'dır yani α_i ve α_i^* dual değişkenlerinin her ikisi de aynı anda sıfırdan farklı değer alamaz. α_i sıfırdan farklı ise, α_i^* sıfırdır ve tam tersi. (iii) $0 < \alpha_i < C$ iken $\xi_i = 0$ 'dır ve benzer şekilde, $0 < \alpha_i^* < C$ iken $\xi_i^* = 0$ 'dır.

Son olarak eklenmesi gerekenler ise destek vektör açılımının seyrekliğine ilişkindir. $\left| y_i - f\left(\vec{x}_i\right) \right| \geq \varepsilon$ olan tüm noktalar için Lagrange çarpanları sıfırdan farklıdır, diğer bir ifadeyle ε-tüpü içinde (Şekil 4.8'deki taralı bölge) bulunan tüm eğitim verileri için α_i ve α_i^* yok olur, yani $\left| y_i - f\left(\vec{x}_i\right) \right| < \varepsilon$ ise α_i ve α_i^* sıfır olmalıdır ki KKT koşulları sağlansın. O halde, $\vec{w}$ 'nin $\vec{x}_i$ 'ler türünden seyrek açılımı elde edilebilir, yani $\vec{w}$ 'i tanımlamak için tüm $\vec{x}_i$ 'lerin kullanılmasına gerek yoktur. Yok edilemeyen katsayılara sahip eğitim örnekleri, bağlanım çözümlemesi için, destek vektörleri oluşturur [51].

5. UYGULAMA

Bu çalışmada iki sınıflı sınıflandırma ve bağlanım analizi için iki ayrı uygulama orijinal veri kümeleri üzerinden gösterilecektir. Çalışma için LibSVM [52] ve Python 2.7. programları kullanılmıştır.

5.1. Wisconsin Göğüs Kanseri Veri Seti

Göğüs kanseri olarak bilinen habis tümör, göğüs hücrelerin anormal biçimde ve kontrolsüz olarak büyüyüp çoğalmasıyla oluşur. Bu kanser türünün oluşmasında yaş, cinsiyet, ırk, aile geçmişi, genetik ve kişisel davranışlar (sigara içme, alkol tüketimi vb.) gibi birçok faktör etkilidir. Ancak, bu risk faktörleri hastanın kanser olup olmadığı ile ilgili kesin bilgiler sağlamaz. Bu nedenle, hasta bireylere ait dokulardan alınan örnekler kullanılarak belirli bir doğruluk oranıyla gerekli analizlerin yapılması gerekmektedir.

Bu çalışmada kullanılan Wisconsin Göğüs Kanseri Veri Seti orijinal olup, Dr. William H. Wolberg tarafından Wisconsin Üniversitesi Hastanelerinde yatan 699 hastadan ince iğne aspiratları ile toplanmıştır [53]. Veri setindeki her örnek, Çizelge 5.1'de verilen 9 nitelik ile açıklanmaktadır. Bu niteliklere 1 ile 10 arasında değerler atanmıştır. 10'a yakın değerlere sahip hastanın habis tümöre sahip olma olasılığı daha fazladır. Veri setinde ayrıca 16 kayıp gözlem vardır. Birçok araştırma makalesi, bu kayıp gözlemleri sınıflandırma algoritmasının değerlendirilmesi aşamasında veri kümesinden çıkarmıştır. Bu çalışmada da bu kayıp gözlemlere sahip hastalar veri kümesinden silinmiştir. 16 örnek eksik özellik içerdiğinden çalışmalar 683 örnek üzerinde yürütülmüştür. 444 tane örnek iyi huylu kanser sınıfına, 239 örnek ise kötü huylu kanser sınıfına aittir.

Çizelge 5.1. Wisconsin göğüs kanseri veri seti

#	*Nitelik*	*Aralık*	*Ortalama*	*Standart Sapma*
1	Kütle Kalınlığı	1-10	4.4422	2.82076
2	Hücre büyüklüğünün üniformluğu	1-10	3.1508	3.06514
3	Hücre şeklinin üniformluğu	1-10	3.2152	2.98858
4	Marjinal adezyon	1-10	2.8302	2.86456
5	Tek epitelyal hücre büyüklüğü	1-10	3.2343	2.22309
6	Yalın çekirdek	1-10	3.5447	3.64386
7	Hafif kromatin	1-10	3.4451	2.44970
8	Normal nükleoli	1-10	2.8697	3.05267
9	Mitozlar	1-10	1.6032	1.73267

DVM ile sınıflandırma yapmadan önce ilk yapılması gereken ve "Veri Ön İşleme (Data Pre-processing)" adı verilen yöntemlerden biri olan *ölçeklendirme (scaling)* işlemi yapılmalıdır. Bu işlemin yapılması hem hesaplamaya dayalı zamanı azaltmak hem de yüksek aralığa sahip niteliklerin, düşük aralığa sahip niteliklere baskın gelerek analiz sonuçlarını etkilemesini engellemektir. Bu nedenle veri kümesindeki her nitelik analizden önce doğrusal olarak $[-1,\ 1]$ aralığına ölçeklendirilmiştir.

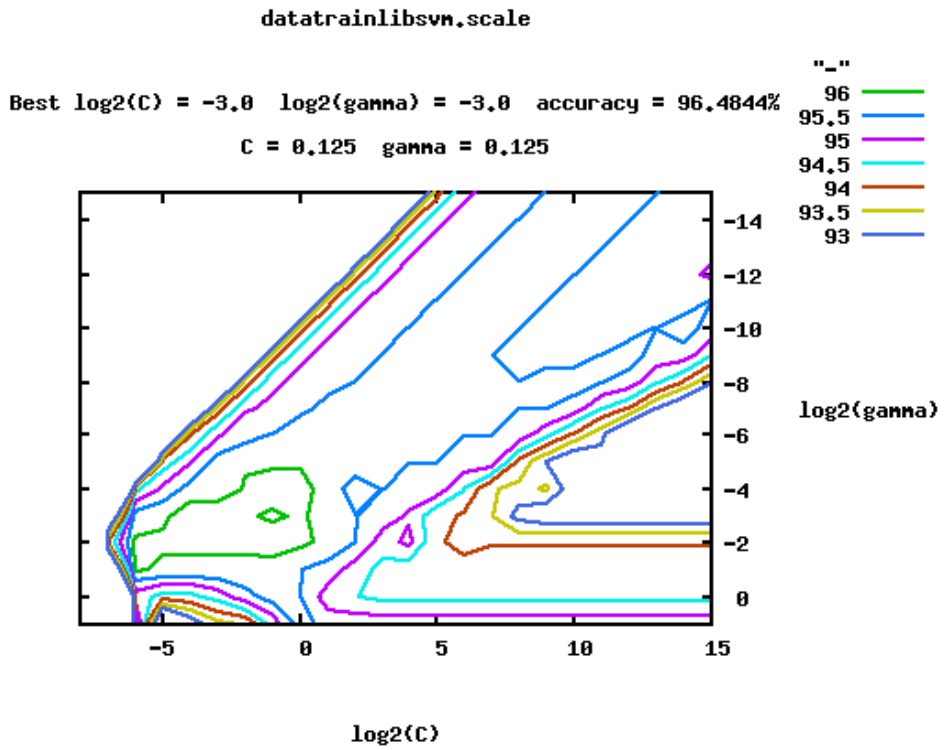
İkinci dikkatli olunması gereken nokta ise kullanılacak çekirdek fonksiyonu seçimidir. Literatürde en çok kullanılan çekirdek *fonksiyonu radyal tabanlı çekirdek fonksiyonudur (radial basis kernel function)* ve çoğunlukla analize bu çekirdek fonksiyonu kullanılarak başlanır. Bunun nedeni, hiperparametrelerin sayısının modelin karmaşıklığını etkilemesidir. Örneğin, polinomiyal çekirdek fonksiyonu, radyal tabanlı çekirdek fonksiyonundan daha fazla parametreye sahiptir, bu nedenle kullanılması tercih edilmez. Radyal tabanlı sınıflandırma işlemi, tüm diğer çekirdek fonksiyonları gibi, doğrusal olmayan veriyi daha yüksek boyutlu bir uzaya taşıyarak işlemi gerçekleştirir. Bu işlem için $\exp\left(-\gamma\left|\vec{x}_i-\vec{x}_j\right|^2\right)$ fonksiyonu kullanılır.

Çekirdek fonksiyonu seçiminden sonra yapılması gereken diğer işlem hiperparametrelerin değerlerinin elde edilmesidir. Bu nedenle bu parametreler için bir model seçimi işlemi yapılması gerekmektedir. Radyal tabanlı çekirdek fonksiyonundan gelen γ parametresi ve optimizasyon probleminden gelen ve düzenleme terimi olan C parametresi burada bulunması gereken parametrelerdir. Bu problem için, parametre araştırma uzayı iki boyutludur. Burada yapılması gereken işlem (C,γ) parametrelerini bulmak ve bilinmeyen bir test verisi üzerindeki performansını ölçmektir. Yani, eğitim kümesinde yüksek doğruluk oranı elde etmek kendi başına yeterli değildir.

Bu nedenle veri kümesi eğitim kümesi ve test kümesi olarak ikiye parçalanmalıdır. Eğitim kümesi üzerinde hesaplanan parametrelerin test kümesi üzerindeki doğruluk oranlarına bakılmalıdır. Bu veri seti için, verinin %75'i (yaklaşık 512 hastaya ait veriler) eğitim kümesi olarak, geri kalan %25'i (171 hasta) ise test kümesi olarak kullanılmıştır. Bu parçalama oranı genellikle veri setinin büyüklüğüne bağlı olarak sezgisel olarak yapılır. Tüm sınıflandırma algoritmaları için en uygun parçalama oranının ne olacağı sorusu halen araştırma konusudur.

Eğitim kümesinden parametre seçimi ise, aşırı uyumu engellemek için 10-katlı çapraz geçerlilik (10-fold cross validation) kullanılarak yapılmıştır. 10-katlı çapraz geçerlilik testinde eğitim verileri 10 gruba ayrılır ilk aşamada birinci grup test, diğer gruplar eğitim için kullanılır. Bu süreç her defasında bir grubun test diğer grupların eğitim amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen on hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır.

10-katlı çapraz geçerlilik ile parametre seçimi *iki adımlı ağ-arama algoritması (two-step grid-search algorithm)* kullanılarak yapılmıştır [54]. Parametre seçimi için literatürde kullanılan ve iterasyona dayalı birçok yöntem vardır ancak bu yöntemler hesaplamaya dayalı yarattıkları maliyetlerin çok olması sebebiyle ağ-arama algoritması tercih edilmiştir. İlk olarak C ve γ parametrelerinin sırasıyla $(2^{-8}, 2^{-7}, ..., 2^{15})$ ve $(2^{1}, 2^{0}, ..., 2^{-15})$ üstel büyüyen dizileri işleme alınmıştır.



Şekil 5.1. C ve γ parametreleri için ağ-arama algoritması

Şekil 5.1'de Pyhton 2.7 ile C ve γ parametreleri seçimi ve gnuplot ile elde edilen çapraz geçerlilik doğruluk konturu verilmiştir. İşlem sonunda %96.4844 çapraz geçerlilik oranıyla

en iyi (C,γ) ikilisi $(2^{-3},2^{-3})$ bulunmuştur. En iyi parametrelerin arandığı alan biraz daha daraltılarak işlem bir kez daha tekrarlanırsa, $C = 0.105112051907$ ve $\gamma = 0.125$ elde edilir. En iyi parametreler bulunduktan sonra eğitim kümesi, son modeli elde etmek için bu parametreler ile bir daha eğitilir. Elde edilen model 171 hastaya ait bilgilerin bulunduğu test kümesine uygulandığında, %99,4152 (170/171) doğru sınıflandırma oranı elde edilir. Destek vektörlerinin sayısının 116 çıkması seyrek ve doğru bir çözüm elde edildiğinin göstergesidir.

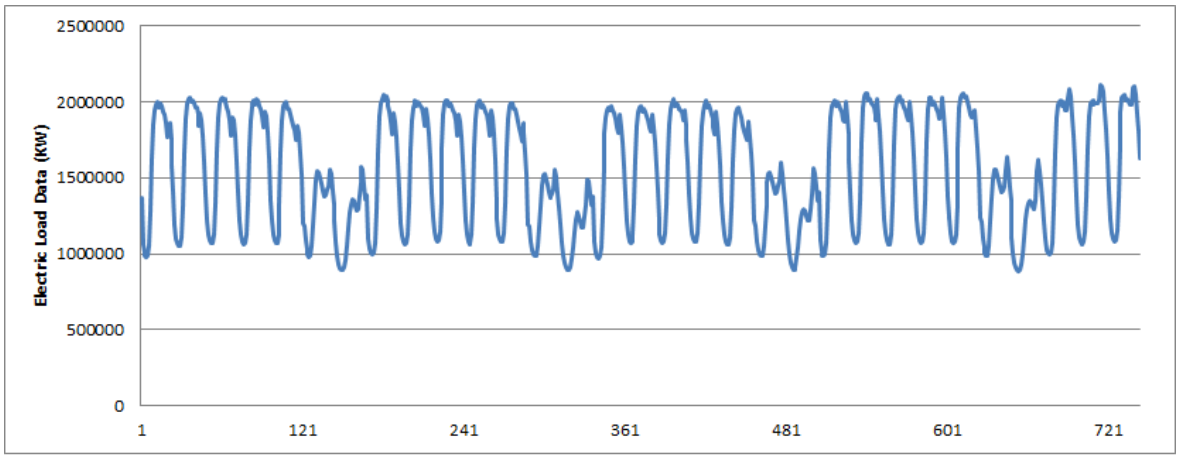
5.2. Kısa Dönemli Elektrik Yük Tahmini

Elektrik enerjisi sektörünün gelişiminde arz, talep, iletim, dağıtım ve fiyatlandırma konularında planlama çok önemlidir. Planlama çalışmalarının daha sağlıklı yapılabilmesi için kullanılacak yük talep verileri mümkün olduğunca doğru tahmin edilmelidir. Elektrik enerjisi tüketimine yönelik tahminler yük tahmini olarak da isimlendirilebilir. Yük tahminlerini kısa, orta ve uzun dönemli olmak üzere üç değişik boyutta incelemek mümkündür. Kısa dönemli tahminler birkaç dakikadan bir güne kadar güç santralleri arasında yük paylaşımı, en iyi grup belirleme ve ekonomik işletmenin yapılabilmesi açısından büyük bir öneme sahiptir. Orta dönemli tahminler bir günden bir yıla kadar dönem için yakıt kaynaklarının dağılımını ve bakım işlemlerinin zamanlamasının belirlenmesi için kullanılır. Uzun dönemli analizler (bir yıldan daha uzun) iletim ağlarının ve yeni üretim kapasitelerinin ekonomik planlamasında, sistemin modifiye edilmesinde önemli olmaktadır.

Kısa dönemli yük tahmini, elektrik güç sistemlerinin günlük operasyonlarının planlanması için zorunlu yapılması gereken işlemlerden biridir [55]. Doğru yapılan tahminler, gün içinde maliyetli birçok operasyon gerçekleştiren enerji şirketleri için hayati önem taşımaktadır. Yapılan deregülasyonlar ve piyasadaki mevcut rekabet nedeniyle değişen enerji sektörünün yapısı, kısa dönemli yük tahminlerinin iyileştirilmesini zorunlu kılmaktır. Bu nedenle, enerji şirketleri, daha doğru tahminler yapmak için bu konuda olası en yüksek verim ile çalışmaktadır. [56].

Bu çalışma kapsamında Hollanda'nın güneyinde bulunan Brabant şehrine ait 12 hafta boyunca kaydedilmiş saatlik elektrik yük verisi kullanılmıştır. Deneysel saatlik veri kümesi, 29.09.2008 23:00 ve 21.12.2008 23:00 tarihleri arasında toplanmıştır. Saatlik arayla kaydedilen elektrik yük verisine ait bir zaman serini birden fazla mevsimsel bileşen

içerir. Şekil 5.2'de Brabant şehrinin Ekim ayı için kaydedilmiş saatlik talep grafiği görülebilir. Bu veri kümesinde gün-içi 24 saatlik ve hafta-içi 168 saatlik mevsimsel periyodlar vardır.



Şekil 5.2. 29 Eylül 2008 Perşembe ve 29 Ekim 2008 Çarşamba tarihleri arası Brabant şehri saatlik elektrik yük talebi

Yük tahminlerinin doğruluk oranlarının iyileştirilmesi problemi araştırmacılar için uzun süredir üzerinde çalışılan önemli bir konu olmuştur. Bu amaçla çeşitli yöntemler ile tahmin modelleri oluşturulmuştur. Bu yöntemleri, bağlanım tabanlı modeller, zaman serileri yaklaşımları ve *bilgi-işlemsel zekâ (computational intelligence)* teknikleri olarak sınıflandırmak mümkündür [57]. İlk kategoriye basit doğrusal bağlanım modelleri, yerel polinomiyal bağlanım modelleri [58], parametrik olmayan bağlanım modelleri [59], sağlam bağlanım yöntemleri [60], çoklu bağlanım modelleri [61] ve doğrusal olmayan bağlanım modelleri [62] konulabilir. Zaman serileri yaklaşımları yük tahminlerine uygulanan en eski yöntemlerdir. Otoregresif hareketli ortalamalar (ARMA), Dışsal açıklayıcı değişken içeren ARMA (ARMAX), otoregresif bütünleşik hareketli ortalamalar (ARIMA) modelleri, iyi bilinen ve en çok kullanılan temel zaman serileri yöntemleridir [63]. Ek olarak, mevsimsellik bileşenlerini de hesaplamaya katan mevsimsel ARMA (SARMA) ve mevsimsel ARIMA (SARIMA) tabanlı modeller de önerilmiştir [64]. James W. Taylor [65] ise, standart *Holt-Winters üstel düzleştirme tekniğini (Holt-Winters exponential smoothing)* iki mevsimsel bileşen için genişleterek *çift-mevsimsel üstel düzleştirme (double seasonal exponential smoothing)* yöntemini önermiş ve önerdiği bu yeni yöntemi çift-mevsimsel ARIMA modeli ile karşılaştırmıştır.

Bilgi-işlemsel zekâ yöntemleri literatüre göreceli yeni dâhil edilmiştir. Bu kategoride bulunan ve elektrik yük tahmini için en çok kullanılan yöntem ise YSA'dır. Yük tahmini için literatüre giren sinir ağlarının bazıları, ileri-beslemeli sinir ağları [66, 67], radyal tabanlı fonksiyon ağları [68], *öz düzenleyici haritalar (self-organizing maps)* [69] ve yinelenen sinir ağlarıdır [70]. Bu kategorideki diğer yöntem ise DVM algoritmasıdır. DVM, yük tahmini modellemesi için üzerinde halen çalışılan yöntemlerden biridir [71, 72, 73, 74]. Parametrik varsayımlara dayanmaması ve doğrusal olmayan veriler üzerinde iyi bir şekilde çalışabilmesi, DVM'i elektrik yük tahmin modelleri için üzerinde çalışılması gereken konulardan biri yapmaktadır.

Bir yük tahmin modeli inşa edebilmek için, uygun girdi değişkenlerinin seçilmesi en önemli ve en temel görevlerden biridir. Bu süreçte takip edilmesi gereken genel bir kural yoktur. Mühendis görüşü, deneme yanılma yöntemi ile elde edilen deneyim genellikle işe yarayan sezgisel tekniklerdir. Ayrıca birkaç istatistiksel ve grafiksel analizler de girdi seçiminde kullanılan bazı yöntemlerdir.

Elektrik talebi ile demografik, meteorolojik, ekonomik faktörler ve takvim faktörleri arasında ilişki olup olmadığını araştıran birçok çalışma yayınlanmıştır [75]. Literatürde bulunan ve elektrik talebini etkilediği düşünülen bu dışsal değişkenler Çizelge 5.2'de özetlenmiştir.

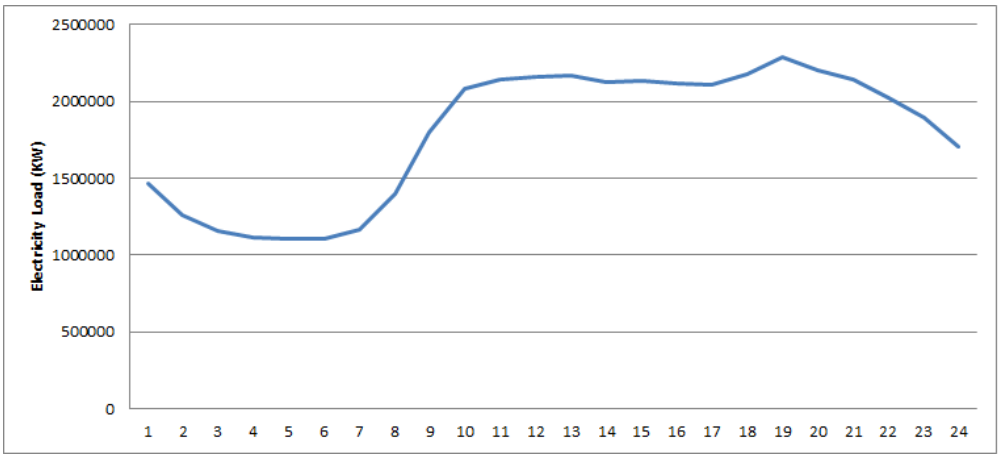
Çizelge 5.2. Elektrik talebini etkileyen faktörler

Takvim Verisi	Günün saati, haftanın günleri, tatiller, okul tatilleri, 'bridge günleri', yaz saati uygulaması vb.
Meteorolojik Veri	Sıcaklık, nem, bulut örtüsü, aydınlık, rüzgâr hızı, iklim değişikliği, güneşin yüksekliği, güneş ısınımı, tutulma esnasında dünyanın pozisyonu vb.
Ekonomik Veri	Gayri Safi Yurtiçi Hasıla (GSYİH), Kişi başı GSYİH, tüketici fiyatları endeksi, ortalama gelir, elektrik fiyatı, endüstriyel genişleme, şirketlerin üretim planları vb.
Demografik Veri	Hane halkı sayısı, nüfus artışı, yerel bölge gelişimi vb.

Çizelge 5.2'de görüleceği üzere, elektrik yükü bir çok değişken tarafından etki altındadır. Ancak burada, yapılacak tahminin süresi önemlidir. Bu çalışmada kısa dönemli elektrik

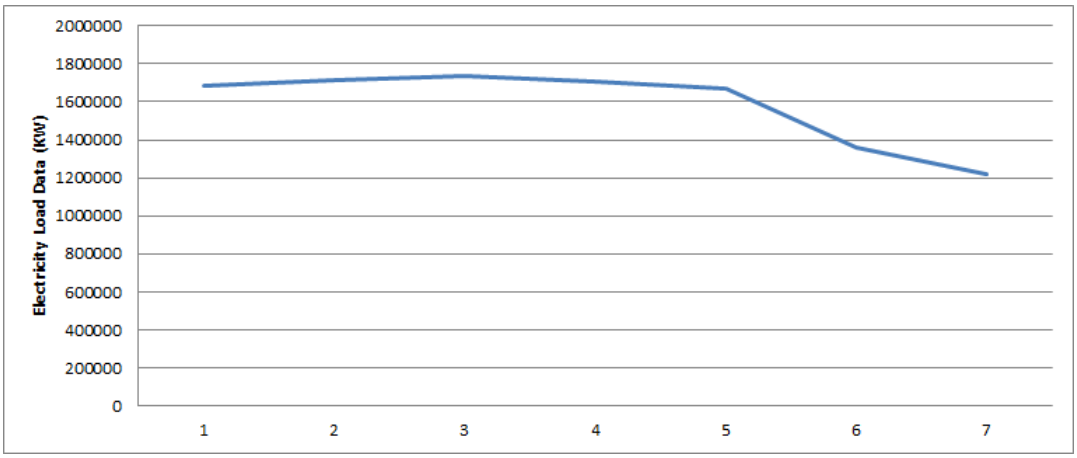
yük modellemesi yapılacağı için, sadece geçmiş verilere dayanan girdi değişkenlerinin kullanılması yeterlidir.

Talep edilen elektrik yükü bir saatten diğerine gün içinde değişebilir. Şekil 5.3, Brabant şehrinin 11.12.2008 günü içinde değişen elektrik yükünü göstermektedir. Şekil 5.3'ten anlaşılacağı üzere gün içerisinde yük şiddetinde anlamlı bir farklılık vardır. Özellikle, çalışma saatleri içerisinde maksimum elektrik tüketimi yaşanmıştır. Bu nedenle, kısa dönemli elektrik yük tahmini modellemesi için zamanın 1'den 24'e değiştiği bir saat göstergesi girdi olarak alınabilir.



Şekil 5.3. 11.12.2008 günü içerisindeki saatlik yük değişimi

Elektrik yükü bir hafta için de günden güne değişim de gösterebilir. Şekil 5.4, Branbant şehrine ait 03.11.2008-09.11.2008 tarihleri arasındaki kırk beşinci haftaya ait veriyi göstermektedir. Hafta sonunda çeşitli nedenler yüzünden elektrik tüketimin azaldığı aşikârdır. Ayrıca, hafta içinde yük tüketimi sabit bir oranda gitmektedir. Bu nedenle hafta sonlarını modelleyebilmek amacıyla gün değişkenine ait 1'den 7'ye değişen bir girdi kullanmak yararlı olacaktır.



Şekil 5.4. 03.11.2008-09.11.2008 haftasındaki ortalama yük tüketimi

Gün ve saat göstergelerinin yanında geçmiş verilere dayanan girdi değişkenlerinin kullanılması tahmin modelinin altındaki önemli yapıları açığa çıkaracaktır. d günü, i saati göstermek üzere, tahmin modelini etkileyen aşağıdaki faktörler de modele dahil edilmiştir.

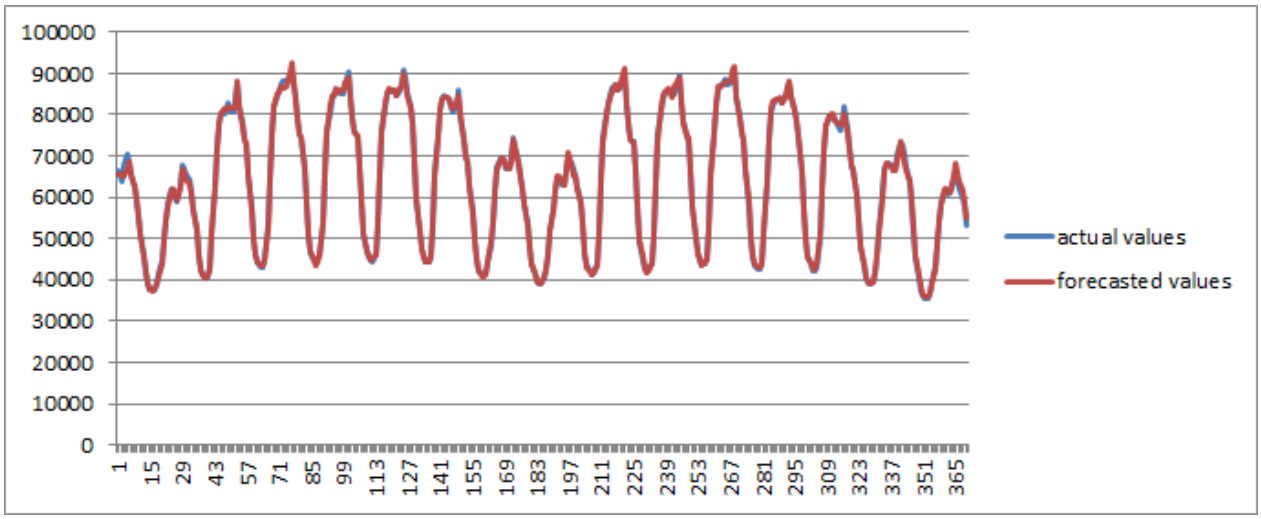
$$\{L_{d-7}^{i-2}, L_{d-7}^{i-1}, L_{d-7}^{i}, L_{d-1}^{i-2}, L_{d-1}^{i-1}, L_{d-1}^{i}, L_{d}^{i-2}, L_{d}^{i-1}\}$$

Burada, L_{d-7}^{i-2}, tahmini yapılacak zaman noktasına karşılık gelen yedi gün ve iki saat önceki elektrik yükünü; L_{d-7}^{i-1}, tahmini yapılacak zaman noktasına karşılık gelen yedi gün ve bir saat önceki elektrik yükünü; L_{d-7}^{i} tahmini yapılacak zaman noktasına karşılık gelen yedi gün önceki elektrik yükünü; L_{d-1}^{i-2}, tahmini yapılacak zaman noktasına karşılık gelen bir gün ve iki saat önceki elektrik yükünü; L_{d-1}^{i-1} tahmini yapılacak zaman noktasına karşılık gelen bir gün ve bir saat önceki elektrik yükünü; L_{d-1}^{i}, tahmini yapılacak zaman noktasına karşılık gelen bir gün önceki elektrik yükünü; L_{d}^{i-2}, tahmini yapılacak zaman noktasına karşılık gelen iki saat önceki elektrik yükünü ve L_{d}^{i-1}, tahmini yapılacak zaman noktasına karşılık gelen bir saat önceki elektrik yükünü gösterir.

Girdi değişkenleri belirlendikten sonra analize başlanabilir. Bu çalışma kapsamında verilerin %80'i, yani, 06.10.2008 01:00 - 06.12.2008 13:00 tarihleri arasındaki 1475 veri noktası eğitim kümesi olarak, verilerin geri kalan %20'si yani 06.12.2008 14:00 - 21.12.2008 23:00 tarihleri arasındaki 370 veri noktası test kümesi olarak kullanılmıştır

Daha sonra, büyük aralığa sahip girdilerin küçük aralığa sahip girdilere baskın olmasını engellemek amacıyla, hem eğitim kümesi hem de test kümesi aynı ölçekleme faktörleri kullanılarak ölçeklendirilmiştir. Analizin kalitesini arttırmak ve tahmin sonuçlarını iyileştirmek için yapılan bu ölçeklendirmede, her özellik bileşeni bağımsız bir şekilde $[0,1]$ aralığına normalleştirilmiştir.

DVR modellemesinde öncelik, kullanılacak çekirdek fonksiyonunun seçilmesindedir. Elektrik yük tahmininde en çok kullanılan haritalama fonksiyonu radyal tabanlı çekirdek fonksiyonudur [76]. Bu nedenle, bu çalışma kapsamında da bu çekirdek fonksiyonun kullanılmasına karar verilmiştir. Radyal tabanlı çekirdek fonksiyonundan gelen γ parametresi ile optimizasyon probleminden gelen, düzenleme terimi olan C parametresi ve göz ardı edilmesi istenilen hata anlamına gelen, duyarsız tüpün büyüklüğünü belirleyen ε burada bulunması gereken meta-parametrelerdir [77]. Bu parametrelerin hepsi aynı anda elde edilmelidir. Sınıflandırma probleminde olduğu gibi burada da bu parameterelerin seçiminde genel geçer bir yöntem bulunmamaktadır. Bu nedenle, Hsu ve ark. [54] çalışması takip edilerek, 10-katlı çapraz geçerlilik kullanılarak, en iyi parametreler iki adımlı ağ-arama algoritması ile bulunmuştur. Öncelikle, tüm bu parametrelere üstel olarak büyüyen rastgele aralıklar verilmiştir ($C = 2^{-6}, 2^{-5}, ..., 2^{9}$ ve $\varepsilon = 2^{3}, 2^{2}, ..., 2^{-11}$ ve $\gamma = 2^{-8}, 2^{-7}, ..., 2^{3}$). Daha sonra bulunan parametrelerin daha dar aralıklarında algoritma bir daha çalıştırılarak, en küçük ortalama hata karesini veren $C = 512$, $\varepsilon = 0.0078125$ ve $\gamma = 0.03125$ değerler optimal parametreler olarak kabul edilmiştir. Bu parametreler kullanılarak test kümesi üzerinde yapılan analiz sonucu, tahmini değerler ile gerçek değerler arasındaki ortalama mutlak yüzde hata 0.870311 olarak bulunmuştur. Test kümesindeki asıl ve tahmini değerler arasındaki ilişki miktarı ise %99.87 olarak elde edilmiştir. Bu değerin yüksek olması DVR'nin bu veri kümesinde iyi sonuç verdiğinin göstergesidir. Şekil 5.5'te gerçel değerler ile tahmini değerlerinin grafikleri birlikte görülebilir.



Şekil 5.5. DVR'nin performansı

Şekil 5.5 incelendiğinde doğrusal olmayan bir veri, yapısına DVR algoritmasının iyi bir şekilde uyum sağladığı söylenebilir. Ancak şekil dikkatli bir şekilde incelendiğinde elektrik yükünün dip değerleri ve zirve noktalarında modellemenin o kadar başarılı olmadığı görülebilir. Gelecek çalışma konularından biri olarak, bu noktaların daha iyi nasıl modelleneceği olabilir.

6. SONUÇLAR

DVM, genellikle sınıflandırma ve bağlanım işlemlerinde kullanılan başarılı bir makine öğrenmesi yöntemidir. Güçlü bir teorik alt yapıya ve iyi bir genelleme yeteneğine sahiptir. Bu amaçla bu tez çalışmassında DVM algoritması ayrıntılı olarak ele alınmış, algoritma hakkında bilgi verilmiş, analizlerde kullanılan optimizasyon problemleri incelenmiş ve her iki tür analiz için de orijinal veri setleri üzerinde DVM'nin nasıl uygulandığı, nelere dikkat edilmesi gerektiği vurgulanmıştır.

İlk olarak doğrusal olmayan sınıflandırma ve bağlanım analizlerin de yapılabilmesi için kullanılması gereken çekirdek fonksiyonları tartışılarak, doğrusal bir methodoloji detaylı bir şekilde incelenmiştir. Daha sonra, özellik uzayındaki iç çarpımların, girdi uzayında çekirdek değerlendirmesi ile yer değiştirmesine izin veren çekirdek hilesi tanıtılmıştır. Böylelikle, DVM'nin bir DÇHU'da nasıl çalıştığı gösterilmiştir. Sonraki tartışmalar ise destek vektör sınıflandırma ve bağlanım analizlerinin temel işleyişlerini gösteren ve öğrenmenin hangi durumlar altında mümkün olduğunu tanımlayan istatistik öğrenme teoremi üzerinde yoğunlaşmıştır. Daha sonra bu iki analiz türü için karesel optimizasyon problemleri tanıtılıp global çözümün nasıl elde edildiği ve seyreklik özelliğinin nasıl sağlandığı detaylı bir şekilde ele alınmıştır.

DVM'nin teorik alt yapısı geçtiğimiz birkaç yıl boyunca yoğun bir şekilde üzerinde çalışılmıştır ve halen çalışılmaktadır. Yeni görev odaklı çekirdek fonksiyonlarının tanıtılması, bu fonksiyonlara ait hiperparametrelerin elde edilmesi için yeni yöntemlerin önerilmesi, hipotez/tahmin fonksiyonunun daha hızlı elde edilmesi ve DVM çözümünün sonsal olasılıkla kalibre edilmesi şu zamana kadar bulunmuş gelişmelerdir. Optimizasyon teoresindeki gelişmeler ise "*Ardışık Minimal Optimizasyon (Sequential Minimal Optimization)*" gibi daha hızlı eğitim yöntemlerinin ortaya çıkmasına neden olmuştur. Bununla birlikte DVM'nin kullanıldığı uygulama alanlarında da artış meydana gelmiştir. Hava durumu öngörüsü yüz tanıma, ses doğrulama, kaotik zaman serilerinin tahmini, yardımcı teminatların fiyatlarının kestirimi bu alanlardan bir kaçıdır.

DVM uygulanması basit bir yöntem gibi düşünülse de, doğru ve güvenilir sonuçlar elde edebilmek için, analiz edilecek verilerin dikkatli bir şekilde ön işleme alınması, kullanılacak çekirdek fonksiyonun verinin türüne uygun olarak seçilmesi ve bu çekirdek

fonksiyonunun parametresiyle, çözülmesi gereken karesel optimizasyon problemlerinden gelen hiperparametrelerin uygun bir şekilde optimize edilmesi gerekmektedir.

KAYNAKLAR

[1] Bellman, R.E., *Adaptive Control Processes: a guided tour*, Princeton University Press, **1961**.

[2] Rosenblatt, F., The Perceptron-a perceiving and recognizing automaton, Report 85-460-1, Cornell Aeronautical Laboratory, **1957**.

[3] Poggio, T., Torre, V., Koch, C., Computational vision and regularization theory, *Nature*, 317, 314-319, **1985**.

[4] Kecman,V., *Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models*, MIT Press, Cambridge, Massachusetts, **2001**.

[5] Lu, W., Wang, W., Leung, A., Lo, S., Yuen, R., Xu, Z., Fan, H., Air Pollutant Parameter Forecasting Using Support Vector Machines*: In Proceedings of the 2002 International Joint Conference on Neural Networks*, 1, 630-635, **2002**.

[6] Mukherje, S., Statistical Learning: Algorithms and Theory, Erişim Tarihi: **03.02.2014**, www.stat.duke.edu\~sayan\statlearn.pdf

[7] Nilsson, N.J., **Learning Machines**, McGraw-Hill, New York, **1965**.

[8] Jäkel, F., Schölkopf, B., Wichmann, F.A., A tutorial on kernel methods for categorization, *Journal of Mathematical Psychology*, 51, 343-358, **2007**.

[9] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines and Other Kernel-based Learning Algorithms*, Cambridge University Press, **2000**.

[10] Schölkopf, B., Smola, A.J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, **2001**.

[11] Mercer, J., Functions of positive and negative type and their connection with theory of integral equations, *Philosophical Transactions of the Royal Society*, 209, 415-446, **1909**.

[12] Genton, M.G., Classes of Kernels for Machine Learning: A Statistics Perspective, *Journal of Machine Learning Research*, 2, 299-312, **2001.**

[13] Hunter, J.K., Nachtergaele, B., *Applied Analysis*, World Scientific Publishing, **2001**.

[14] Ciesielski, K., *Set theory for the working mathematician*, Cambridge University Press, **1997**.

[15] Heil, C., Banach and Hilbert Spaces Review, Erişim Tarihi: **14.03.2014**, www.people.math.gatech.edu/~heil/handouts/banach.pdf

[16] Meyer, C.D., *Matrix Analysis and Applied Linear Algebra*, SIAM: Society for Industrial and Applied Mathematics, **2001**.

[17] Hartig, D.G., The Riesz representation theorem revisited, *American Mathematical Monthly*, 90(4), 277-280, **1983**.

[18] Rynne, B., Youngson, M.A., *Linear Functional Analysis (Springer Undergraduate Mathematic Series)*, 2. Baskı, Springer, **2007**.

[19] Horváth, J., On the Riesz-Fischer Theorem, Erişim Tarihi: **21.03.2014**, www.univie.ac.at/NuHAG/FEICOURS/ws0506/Riesz-Fisher.pdf

[20] Saze, K., *Beginning Functional Analysis (Undergraduate Texts in Mathematics)*, Springer, **2010**.

[21] Herbrich, R., *Learning Kernel Classifiers: Theory and Algorithms (Adaptive Computation and Machine Learning)*, MIT Press, **2001**.

[22] Cucker, F., Smale, S., On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, 39(1), 1-49, **2001**.

[23] Aronszajn, N., Theory of reproducing kernels, *Transactions of the American Mathematical Society*, 68, 337-404, **1950**.

[24] Berlinet, A., Thomas-Agnan, C., *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, 2004 Edition, Springer, **2003**.

[25] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction (Springer Series in Statistics)*, 7. Baskı, Springer, **2013**.

[26] Ivanov, V.K., Vasin, V.V., Tanana, V.P., *Theory of Ill-Posed Problems and its Applications*, 2. Baskı, VSP, **2002**.

[27] Ivanov, V.V., *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*, Noordhoff International Publishing, **1976**.

[28] Tikhonov, A.N., Arsenin, V.Y., *Solutions of Ill-Posed Problems*, Wiley, New York, **1977**.

[29] Rakhlin, A., Applications of Empirical Processes in Learning Theory: Algorithmic Stability and Generalization Bounds, Doktora Tezi, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, **2006**.

[30] Schölkopf, B., Herbrich R., Smola, A.J., A Generalized Representer Theorem, *Computational Learning Theory:Lecture Notes in Computer Science*, 2111, 416-426, **2001**.

[31] Trafalis, T.B., Ince, H., Suppport vector machine for regression and applications to financial forecasting: *: In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks,* 6, 348-353, **2006**.

[32] Alpaydın, E., *Introduction to Machine Learning*, 2. Baskı, MIT Press, **2004**.

[33] Suykens, J., Gestel, T., Brabanter, J., Moor, B., Vandewalle, J., *Least Squares Support Vector Machines*, World Scientific Pub. Co., **2002**.

[34] Vapnik, N.V., Chervonenkis, A.Y., On the Uniform Convergence of Relative Frenquescies of Events to Their Probabilities, *Theory of Probability and its Applications*, 16(2), 264-280, **1971**.

[35] Cherkassky, V., *Learning from Data*, John Wiley & Sons, **1998**.

[36] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, NY, **1995**.

[37] Vapnik, V.N., *Statistical Learning Theory*, John Wiley & Sons, Toronto, CA, **1998**.

[38] Cortes, C., Vapnik, V., Support-Vector Networks, *Machine Learning*, 20, 273-297, **1995**.

[39] Frogner, C., Lecture Notes: Support Vector Machines, Erişim Tarihi: **20.11.2013**, www.mit.edu/~9.520/spsring11/slides/class06-svm.pdf

[40] Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2, 121-167, **1998**.

[41] Fletcher, R., *Practical Methods of Optimization*, 2. Baskı, Wiley, **2013**.

[42] Karush, W., *Minima of Functions of Several Variables with Inequalities as Side Conditions*, Yüksek Lisans tezi, Department of Mathematics, University of Chicago, **1939**.

[43] Kuhn, H.W., Tucker, A.W., Nonlinear programming: *In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 481–492, **1951**.

[44] Boser, B.E., Guyon, I.M., Vapniki V.N., A training algorithm for optimal margin classifiers: *In Proceedings of the 5th annual workshop on Computational Learning Theory*, New York, NY, **1992**.

[45] Aizerman, M.A., Braverman, E.A., Rozonoer, L., Theoretical foundations of the potential function method in pattern recognition learning, *Automation and Remote Control*, 25, 821-837, **1964**.

[46] Smola, A.J., Regression estimation with support vector learning machines, Yüksek Lisans tezi, Technische Universität München, **1996**.

[47] Vapnik, V., Golowich, S., Smola, A., Support Vector Method for Function Approximation, Regression Estimation and Signal Processing, *Neural Information Processing Systems*, 9, 281-287, **1996**.

[48] Vapnik, V.N., *Estimation of Dependences Based on Empirical Data (Information Science and Statistics Series),* Springer, **2006**.

[49] Huber, P.J., Robust Estimation of a Location Parameter, *Annals of Statistics*, 53, 73–101, **1964**.

[50] Gunn, S.R., Support Vector Machines for Classification and Regression, Technical Report, Erişim Tarihi: **19.06.2013**, www.ce.sharif.ir/courses/85-86/ce725/resources/root/LECTURES/SVM.pdf

[51] Smola, A., Schölkopf, B., A Tutorial on Support Vector Regression, *Statistics and Computing*, 14, 199-222, **2004**.

[52] Chih-Chung, C., Chih-Jen, L., LIBSVM : a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27, **2011**.

[53] Mangasarian, O. L., Wolberg, W. H., Cancer diagnosis via linear programming, *SIAM News*, 23(5), 1-18, **1990**.

[54] Hsu, C.-W., Chang, C.-C, Lin, C.-J., A practical guide to support vector classiffication, Reliability Engineering & System Safety, **2003**.

[55] Senjyu, T., Takara, H., Uezato, K., Funabashi T., One-hour-ahead load forecasting using neural network, *IEEE Transactions on Power Systems*, 17(1), 113-118, **2002**.

[56] Khotanzad, A., Afkhami-Rohani, R., Maratukulam, D., ANNSTF- Artificial neural network short-term load forecaster generation three, *IEEE Transactions on Power Systems*, 13(4), 1413-1422, **1998**.

[57] Hahn, H., Meyer-Nieberg, S., Pickl S., Electric load forecasting methods: Tools for decision making, *European Journal of Operational Research*, 199(3), 902-907, **2009**.

[58] Zivanovic R., Local regression-based short-term load forecasting, *Journal of Intelligent and Robotic Systems*, 31(1-3), 115-127, **2001**.

[59] Charytoniuk, W., Chen, M.S., van Olinda, P., Nonparametric regression based short-term load forecasting, *IEEE Transactions on Power Systems*, 13(3), 725-730, **1998**.

[60] Jin, L., Lai, Y.J., Long, T.X., Peak load forecasting based on Robust regression: *In Proceedings of 8th International Conference on Probabilistic Methods Applied to Power Systems*, 123-128, **2004**.

[61] Hor, C.-L., Watson, S.J., Majithia, S., Analyzing the impact of weather variables on monthly electricity demands, *IEEE Transactions on Power Systems*, 20(4), 2078-2085, **2005**.

[62] Bruhns, A., Deurveilher, G., Roy, J.-S., A non-linear regression model for midterm load forecasting and improvements in seasonality*: In Proceedings of 15th Power Systems Computation Conference*, **2005**.

[63] Mbamalu, G.A.N., El-Hawary, M.E., Load forecasting via suboptimal seasonal auto regressive models and iteratively reweighted least squares estimation, *IEEE Transactions on Power Systems,* 8(1), 343-348, **1993**.

[64] Espinoza, M., Joye, C., Bemans, R., de Moor, B., Short-term load forecasting, prole identication and customer segmentation: A methodology based on periodic time series, *IEEE Transactions on Power Systems*, 20(3), 1622-1630, **2005**.

[65] Taylor, J.W., Short-term electricity demand forecasting using doubleseasonal exponential smoothing, *The Journal of the Operational Research Society*, 54(8), 799-805, **2003**.

[66] Park, D.C., El-Sharkawi, M.A., Marks, R.J.(II), Atlas, L.E., Damborg, M.J., Electric load forecasting using an artificial neural network, *IEEE Transactions on Power Systems*, 6(2), 442-449, **1991**.

[67] Lu, C.N., Wu, H.T., Vemuri, S., Neural network based short term load forecasting, *IEEE Transactions on Power Systems*, 8(1), 336-342, **1993**.

[68] Ranaweera, D.K., Hubele, N.F., Papalexopoulos, A.D., Application of radial basis function neural network model for short-term load forecasting: *In IEEE Proceedings Generation, Transmission and Distribution*, 142(1), 45-50, **1995**.

[69] Beccali, M., Cellura, M., Lo Brano, V., Marvuglia, A., Forecasting daily urban electric load profiles using artificial neural networks, *Energy Conversion and Management*, 45(18), 2879-2900, **2004**.

[70] Senjyu, T., Mandal, P., Uezato, K., Funabashi, T., Next day load curve forecasting using recurrent neural network structure, *IEEE Transactions on Power Systems*, 20(1), 102-109, **2005**.

[71] Chen, B.J., Chang, M.W., Lin, C.J., Load forecasting using support vector machines: a study on eunite competition 2001, *IEEE Transactions on Power Systems*, 19(4), 1821-1830, **2004**.

[72] Mohandes. M., Support vector machines for short-term electrical load forecasting, *International Journal of Energy Research*, 26(4), 335-345, **2002**.

[73] Li, Y., Fang, T., In Wavelet and Support Vector Machines for Short-Term Electrical Load Forecasting: *In Proceedings of the International Conference on Wavelet Analysis and its Applications*, 1, 399-404, **2003**.

[74] Niu, D., Wang, Y., Wu, D.D., Power load forecasting using support vector machine and ant colony optimization, *Expert Systems with Applications*, 37(3), 2531-2539, **2010**.

[75] Heeren, M.E.M., Improving the Electricity Distribution Loss Forecasting Accuracy for Enexis B.V., Yüksek Lisans Tezi, Eindhoven University of Technology, Eindhoven, Hollanda, **2011**.

[76] Niu, D., Wang, Y., WU, D.D., Power load forecasting using support vector machine and ant colony optimization, *Expert Systems with Applications*, 37, 2531-2539, **2010**.

[77] Cherkassky, V., Ma, Y., Selection of Meta-parameters for Support Vector Regression: *In Proceedings of the International Conference on Artificial Neural Networks*, 687-693, **2002**.

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı: Mustafa Murat ARAT

Doğum Yeri: İskenderun

Medeni Hali: Bekar

E-posta: muratarat@hacettepe.edu.tr

Adresi: Hacettepe Üniversitesi İstatistik Bölümü 06800 Beytepe, Ankara

Eğitim

Lise: Sokullu Mehmet Paşa Lisesi

Lisans: Hacettepe Üniversitesi, İstatistik Bölümü

Yüksek Lisans: Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik A.B.D.

Yabancı Dil ve Düzeyi

İngilizce, İleri Seviye

Fransızca, İleri Seviye

İş Deneyimi

2011-… Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Araştırma Görevlisi

Deneyim Alanları

-

Tezden Türetilmiş Projeler ve Bütçesi

-

Tezden Yazılmış Yayınlar

-

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

-