MAKINE ÖĞRENMESI

Dr.Öğr.Üy. Filiz Gürkan –
Elektrik Elektronik Mühendisliği
İstanbul Medeniyet Üniversitesi

Dr.Öğr.Üyesi Filiz Gürkan
filiz.gurkan@medeniyet.edu.tr
Kuzey Kampüs- F008

- Slaytlar
- Pattern Recognition and Machine Learning, Christopher M. Bishop
- Introduction to Machine Learning, The MIT Press, Ethem ALPAYDIN
- İnternet kaynakları – Makale ve bildiriler

Notlandirma:

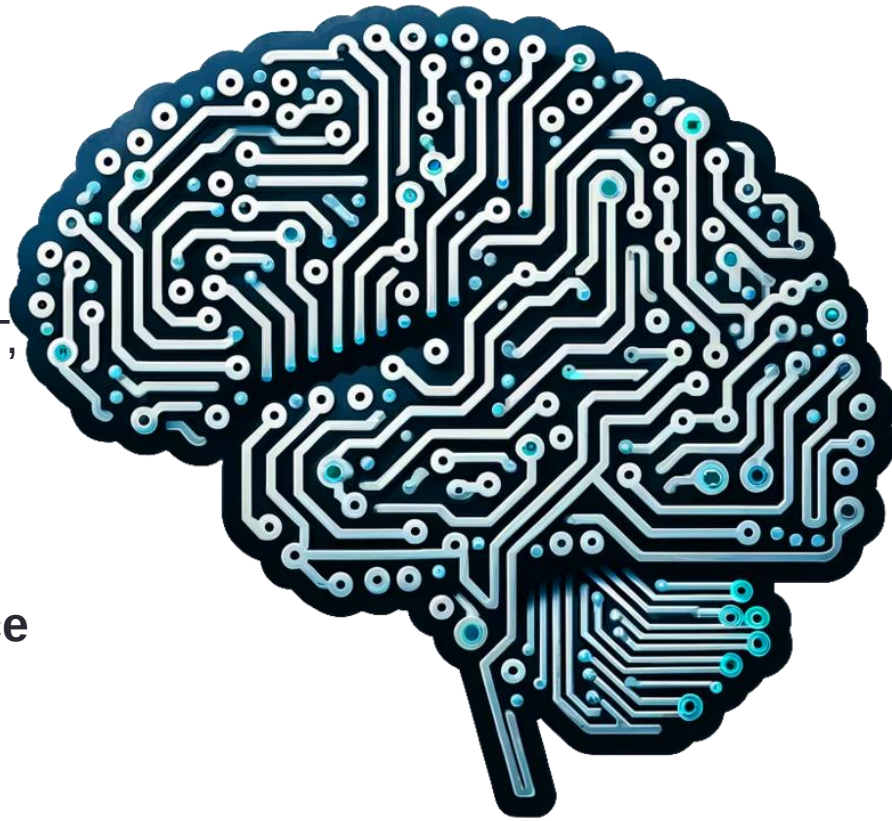
- YIL İÇİ ÖDEVLER+PROJE (%60)
- FİNAL (%40)

23.02.2024	GİRİŞ	
01.03.2024	TEMEL KAVRAMLAR	
08.03.2024	PCA-LDA	
15.03.2024	Öğreticili Öğrenme	ÖDEV1
22.03.2024	Öğreticili öğrenme	
29.03.2024	Öğreticili öğrenme	
05.04.2024	Öğreticili öğrenme	ÖDEV2
12.04.2024	DERS YOK	
19.04.2024	ARA SINAV HAFTASI	
26.04.2024	Sunumlar-Öğreticisiz öğrenme	
03.05.2024	Öğreticisiz öğrenme	
10.05.2024	WEKA	ÖDEV3
17.05.2024	YAPAY SİNİR AĞLARI	
24.05.2024	TEKRAR	
31.05.2024	SUNUMLAR	

Öğrenme ?

- *Makine öğrenmesi*, bilgisayarların bir performans kriterini enbüyük leyecek şekilde programlanması olarak tanımlanabilir

• [Tom Mitchell]
learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E



algorithms that improve their **performance** at some **task** with **experience**


Data
Program
Computer
Output

X→ Y

X: emails

Y: {spam, değil}

AMAÇ : f(X) -> Y yi elde edeceğimiz f(.) yi belirlemek


Data
Output
Computer
Data
Program
Computer
Output

Nasıl ?

- Örnek veri seti / geçmiş deneyim
- Örneklerden genel modelin öğrenilmesi
  - Satış : Müşteri işlemlerinden müşteri profilinin öğrenilmesi
- Belli amaçla toplanmış veriye uygun iyi bir modelleme yapılması
  - Covid verilerinden hastalık teşhisine uygun öznitelikl erin çıkarılması

Arthur Samuel (1952)
Dama oyunu

Frank Rosenblatt (1957)
Perceptron – ikili sınıflandırma

1960-1970s – AI Boom (MLP-SGD- geri yayılım)

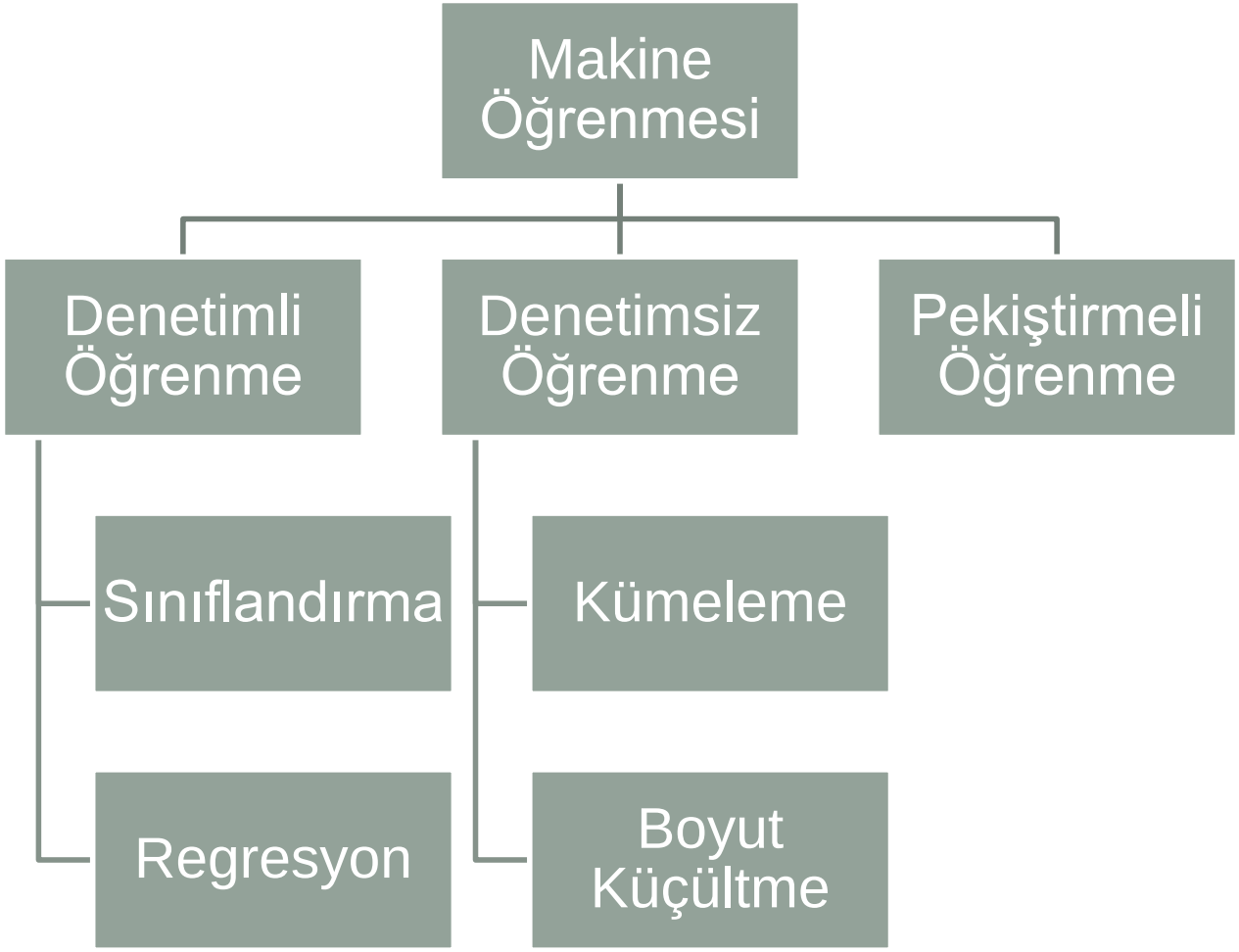
1970s-1980 – AI WINTER

1980-1987 – AI BOOM

1987-1993 – AI WINTER

Gerry Tesauro (1994)
1997

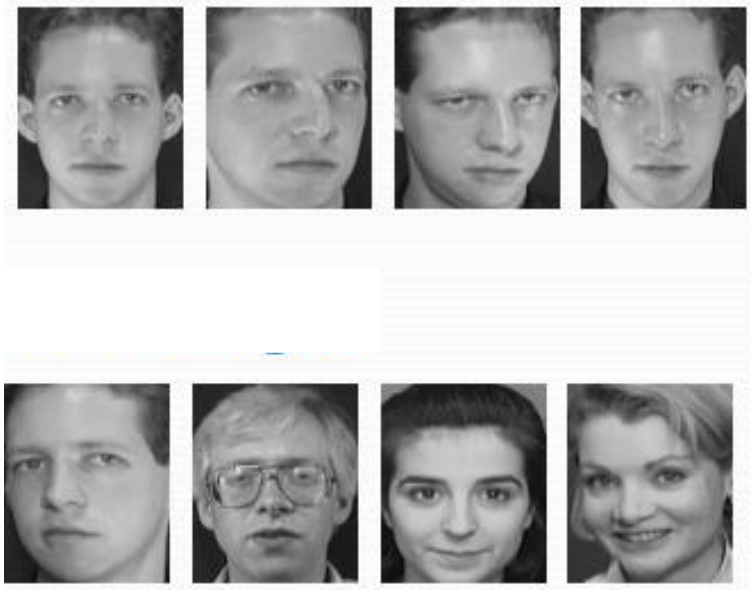
Öğrenme yaklaşımları



Yarı denetimli öğrenme

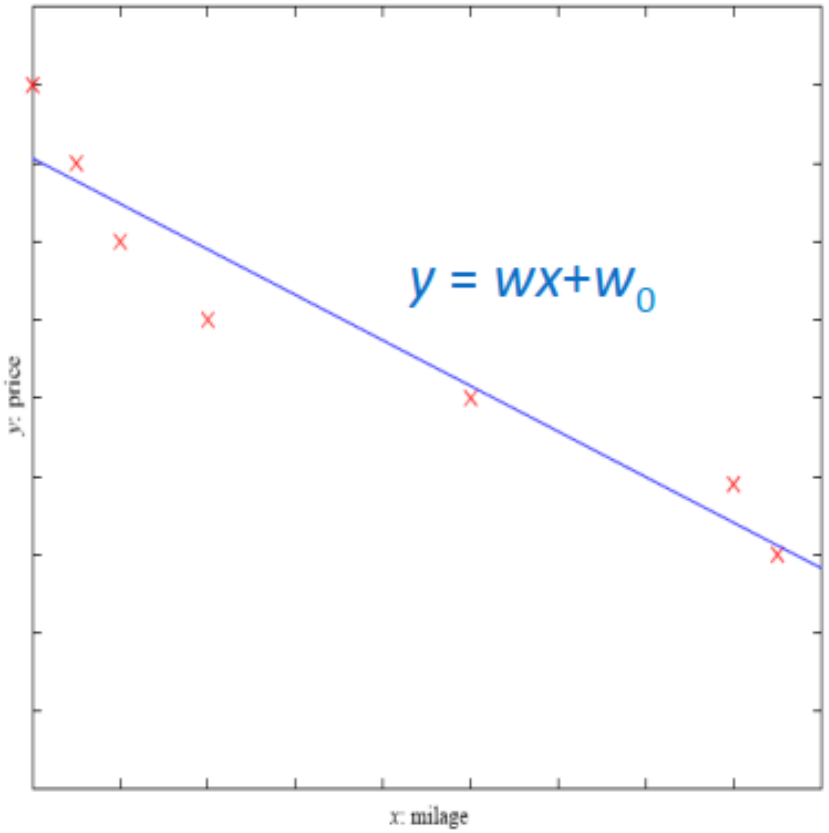
Öğreticili öğrenme

- Sınıflandırma
- Kesikli çıktı (discrete)
  - Yüz tanıma
  - Karakter tanıma
  - Konuşma tanıma
  - Medikal data
  - Biometrik data



Öğreticili öğrenme

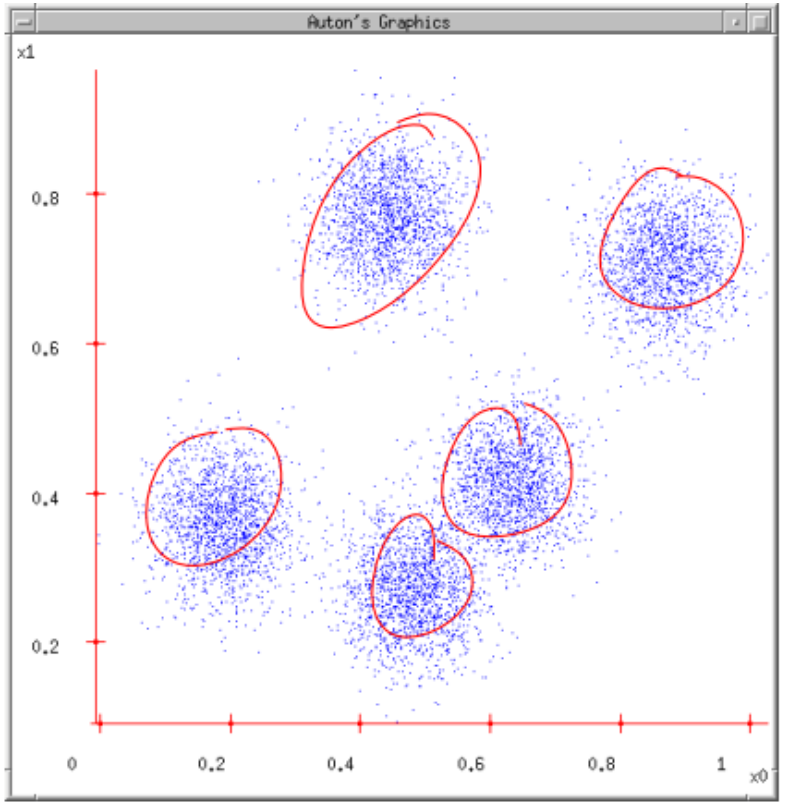
- Regresyon
  - Sürekli çıktı (Continuous)



Öğreticisiz öğrenme

- Çıkışta sınıf etiketi yoktur
  - Öbekleme (Clustering) : Benzerlik kriterlerine göre giriş örneklerini gruplar

- Örnek uygulamalar
  - Müşteri segmentini belirleme
  - Görüntü bölütleme
  - Anomali tespiti



Pekiştirmeli öğrenme

- Geri beslemeden öğrenme
- Öğrenme kuralı : Bir dizi çıktı
  - Oyun algoritmaları

There is only one “supervised” signal at the end of the game.

But you need to make a move at every step

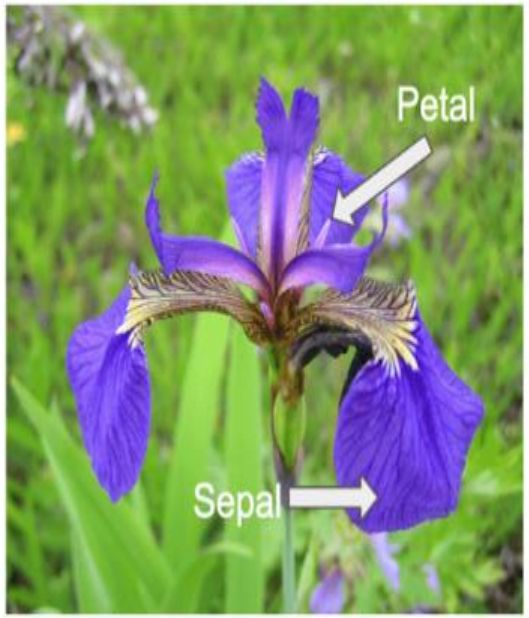
VERİ TÜRLERİ

- Sayisal
  - Sürekli
  - Ayrık
- Kategorik
  - Ordinal
  - Nominal

MAKİNE ÖĞRENMESİ

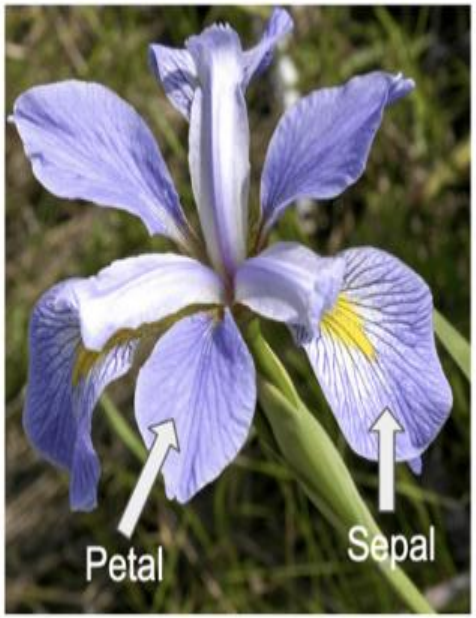
IRIS VERİSETİ

Iris setosa



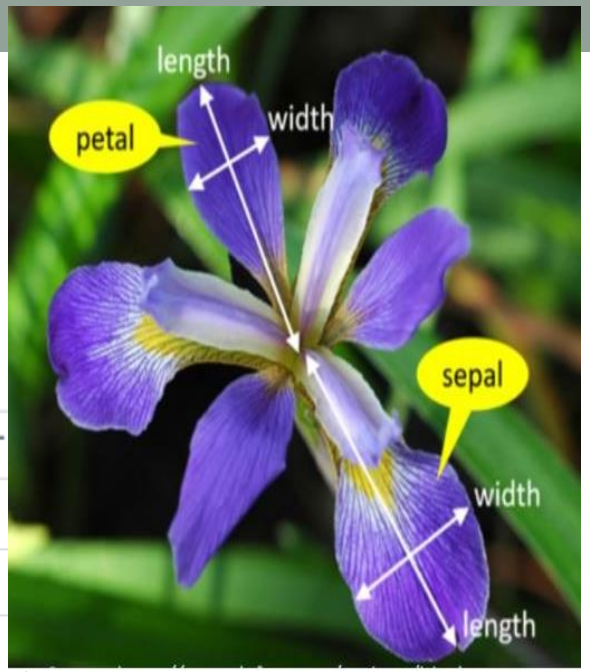
Iris versicolor

Iris virginica



50X3 = 150 adet görsel

# sepal_length	# sepal_width	# petal_length	# petal_width	∆ species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5.0	2.0	3.5	1.0	versicolor
5.9	3.0	4.2	1.5	versicolor
6.0	2.2	4.0	1.0	versicolor
7.7	2.6	6.9	2.3	virginica
6.0	2.2	5.0	1.5	virginica
6.9	3.2	5.7	2.3	virginica
5.6	2.8	4.9	2.0	virginica
5.4	3.7	1.5	0.2	setosa



Öznitelikler(feature-attributes)

Sınıf (class)

Örnekler-Gözlemler (instance-object)

# sepal_length	# sepal_width	# petal_length	# petal_width	∆ species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5.0	2.0	3.5	1.0	versicolor
5.9	3.0	4.2	1.5	versicolor
6.0	2.2	4.0	1.0	versicolor
7.7	2.6	6.9	2.3	virginica
6.0	2.2	5.0	1.5	virginica
6.9	3.2	5.7	2.3	virginica
5.6	2.8	4.9	2.0	virginica
5.4	3.7	1.5	0.2	setosa

Yapısal veriler

Yapısal olmayan veriler

Resim : her bir pikseli, renkli resimlerde R,G,B değerleri, siyah-beyaz resimlerde 1–255 arası gri seviyesi kullanılarak sayılara çevrilir. Renkli resimler 3 adet, siyah beyazlar 1 adet en*boy büyüklüğünde matrisle ifade edilir.

Metin :harfler, heceler ve kelimeler genelde frekanslarına göre kodlanarak sayılara çevrilir.

Hareketli görüntü: Resim bilgisine ek olarak resmin hangi resimden sonra geldiğini gösteren zaman bilgisini de içerir. Bu ek bilgi haricinde yapılan işlem resim ile aynıdır.

Ses : genlik ve frekansın zaman içinde değişimiyle kodlanır.

Veri türleri

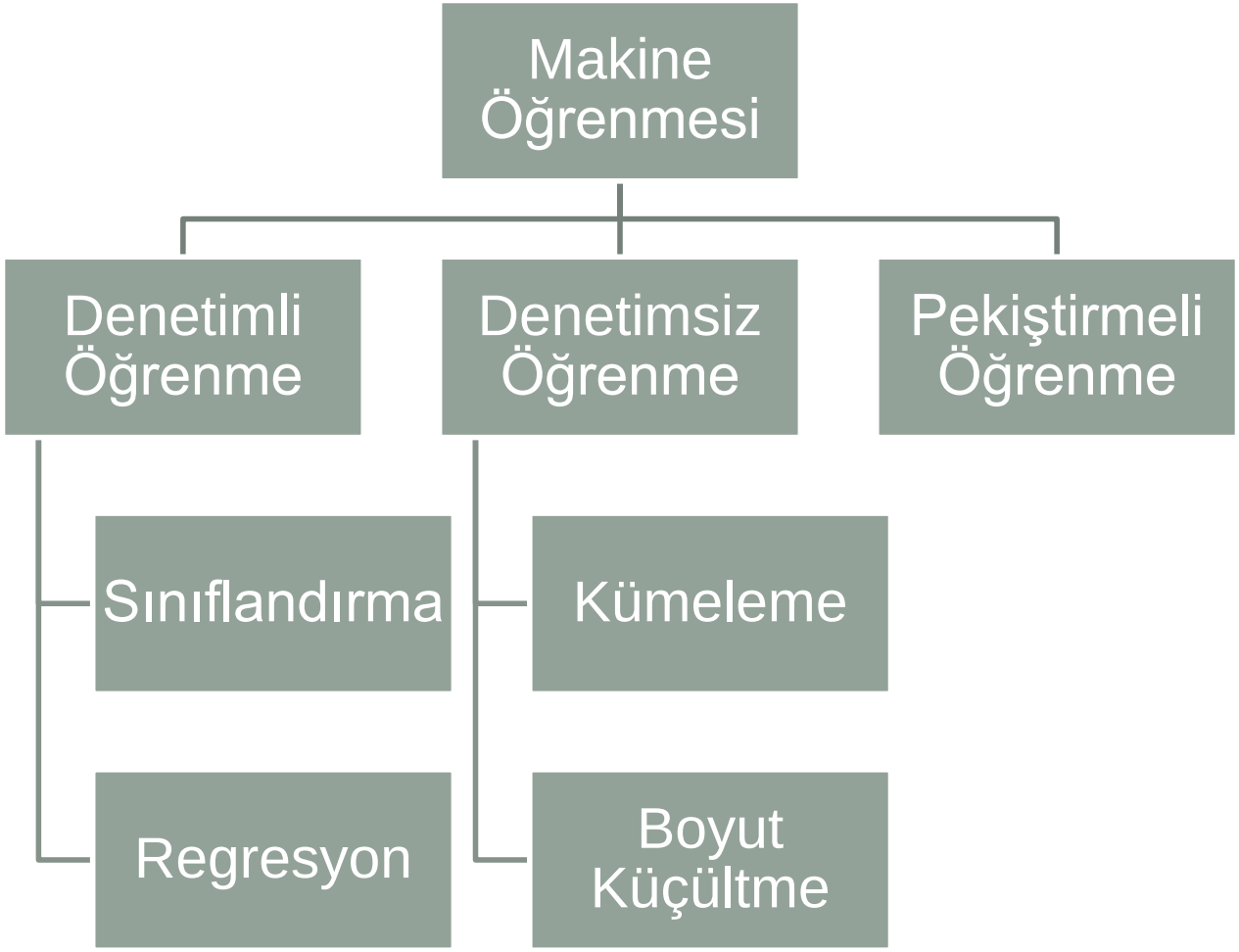
- Sayısal (NÜMERİK) veriler
  - Sürekli → 1.63 cm boy, 12265 km …
  - Ayrık →3 kişi, 5 elma (sayılabilir nicelikler)..

- Kategorik
  - Ordinal (sıralı) → iyi-orta-kötü , yüksek-orta-düşük
  - Nominal → cinsiyet, medeni durum, renk

- Tüm ML algoritmaları için
  - Girişi çıkışa dönüştürecek fonksiyonu bulmak f: X → Y
  - X: emails, Y: {spam, notspam}

- Every machine learning algorithm has three components:
  - Representation – model (derin öğrenme-SVM..)
  - Evaluation – performans değerlendirme aşaması (doğruluk-hassaslık …)
  - Optimizasyon-doğru model ağırlıklarını bulma süreci

Öğrenme yaklaşımları

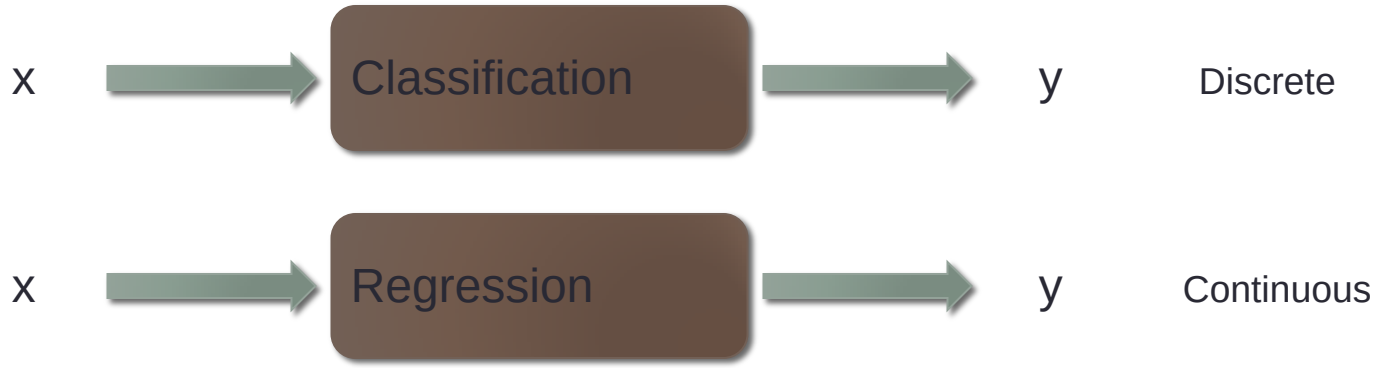


SPAM MAİL SINIFLANDIRMA

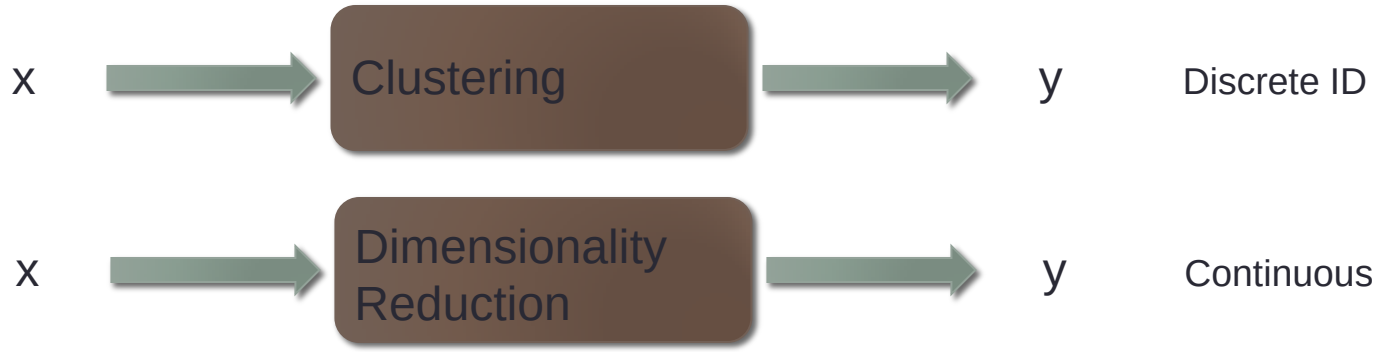
- Spam Emails
  - Bazı belirgin kelimeler
    - “money”
    - “free”
    - “bank account”
    - …..
- Normal Emails
  - Ilgili kelimelerin kullanımı az

Slide Credit: Fei Sha

Supervised Learning



Unsupervised Learning



Öğreticili öğrenme

- Sınıflandırma
- Kesikli çıktı (discrete)
  - Yüz tanıma
  - Karakter tanıma
  - Konuşma tanıma
  - Medikal data
  - Biometrik data

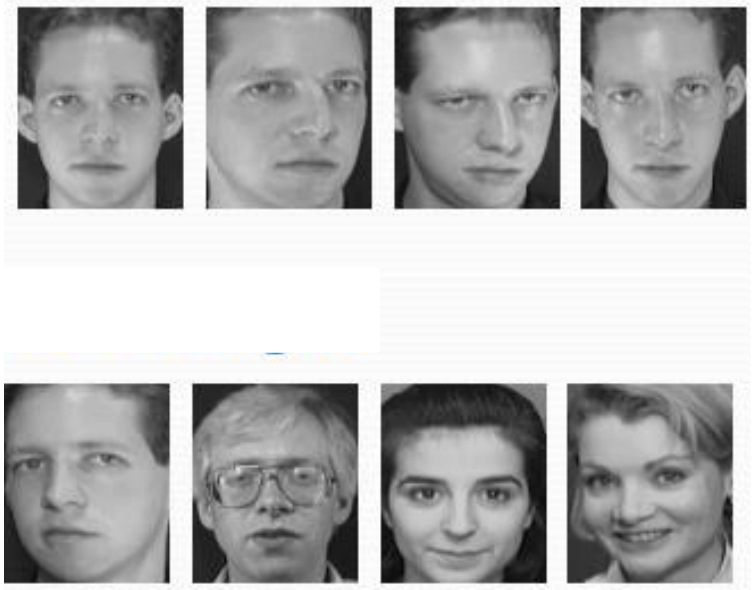


Image Classification

Pizza
Wine
Stove

Yüz Tanıma

Konuşma Tanıma

Slide Credit: Carlos Guestrin

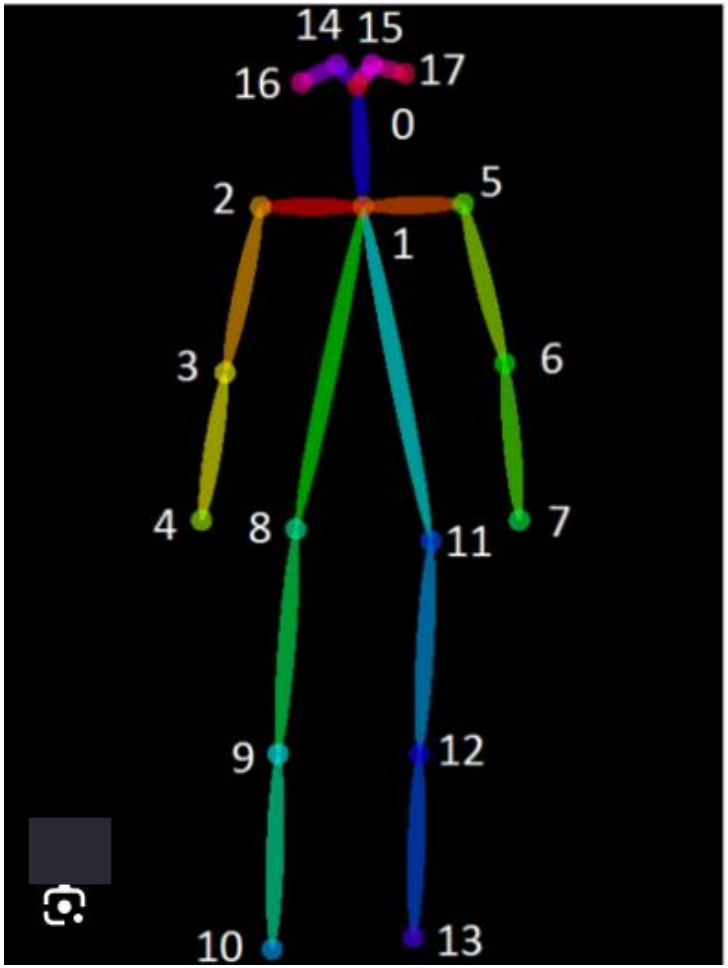
Stock market

Hava tahmini

Slide Credit: Carlos Guestrin

Pose Estimation

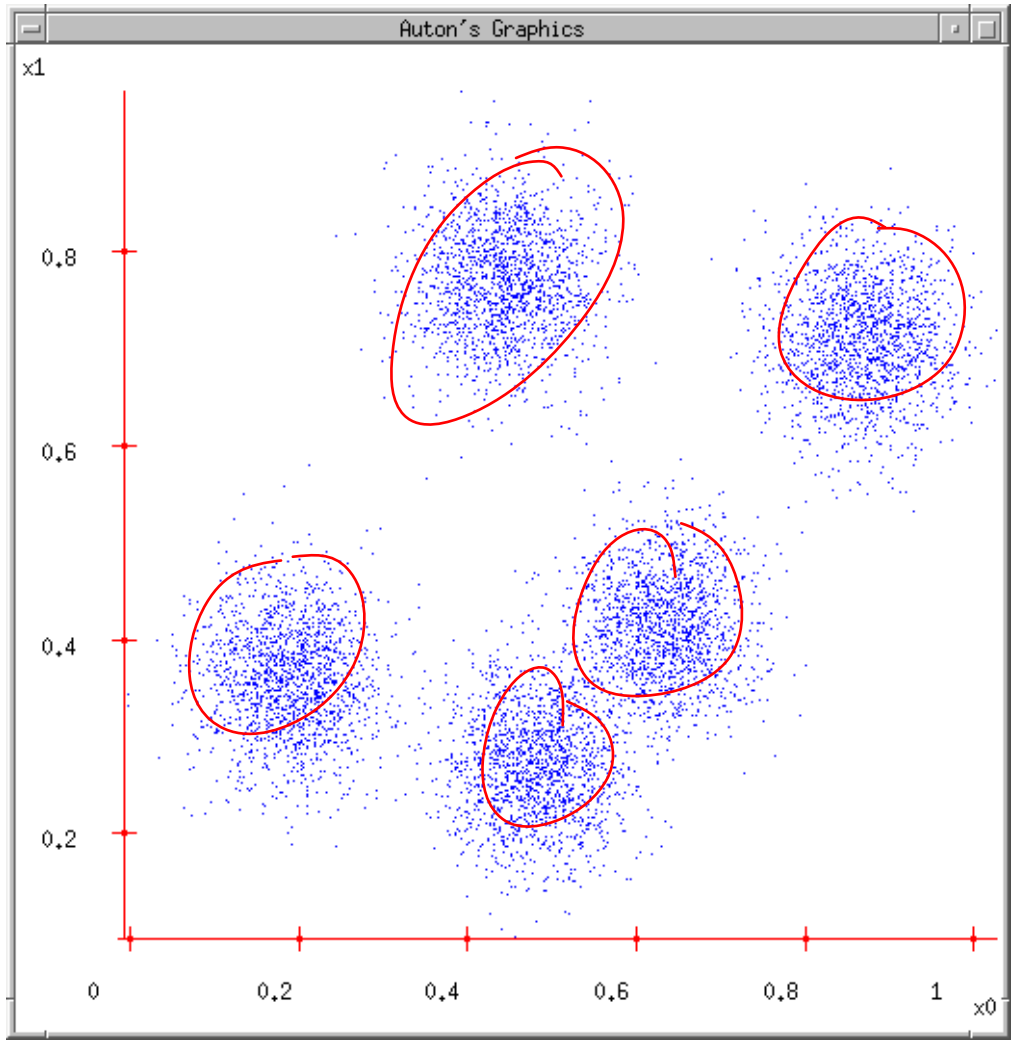
- https://www.youtube.com/watch?v=RMgrAxds3DU



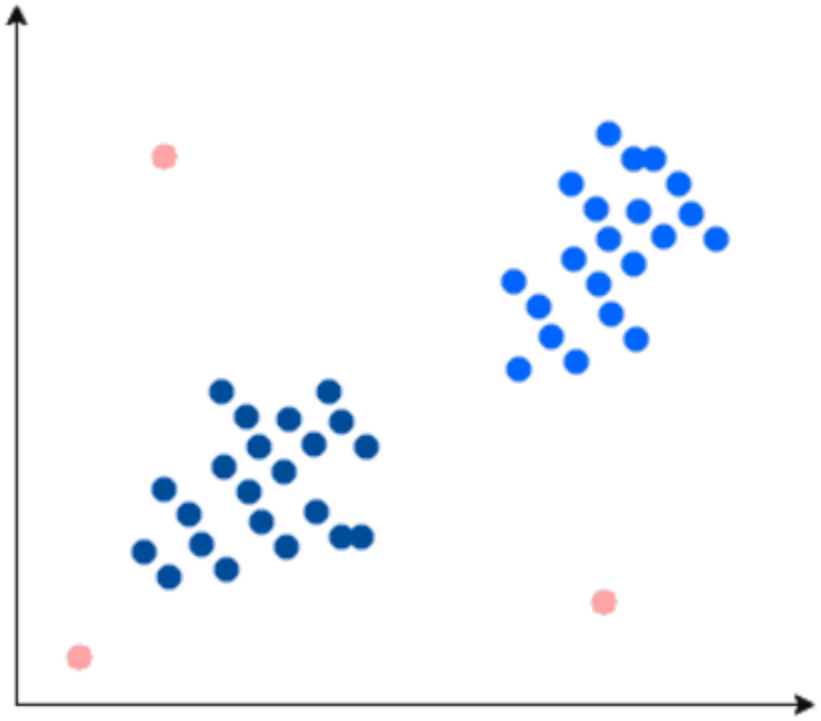
Unsupervised Learning

Y BİLİNMEZ

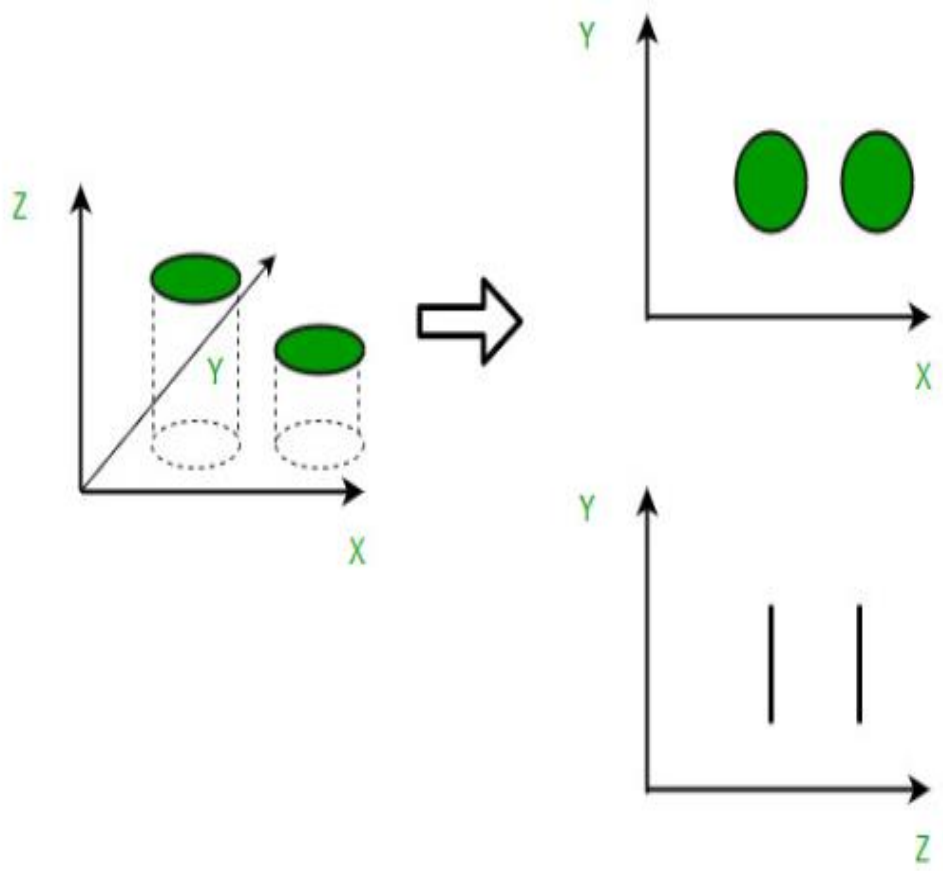
Clustering Data: Benzer ŞEYLERi gruplama

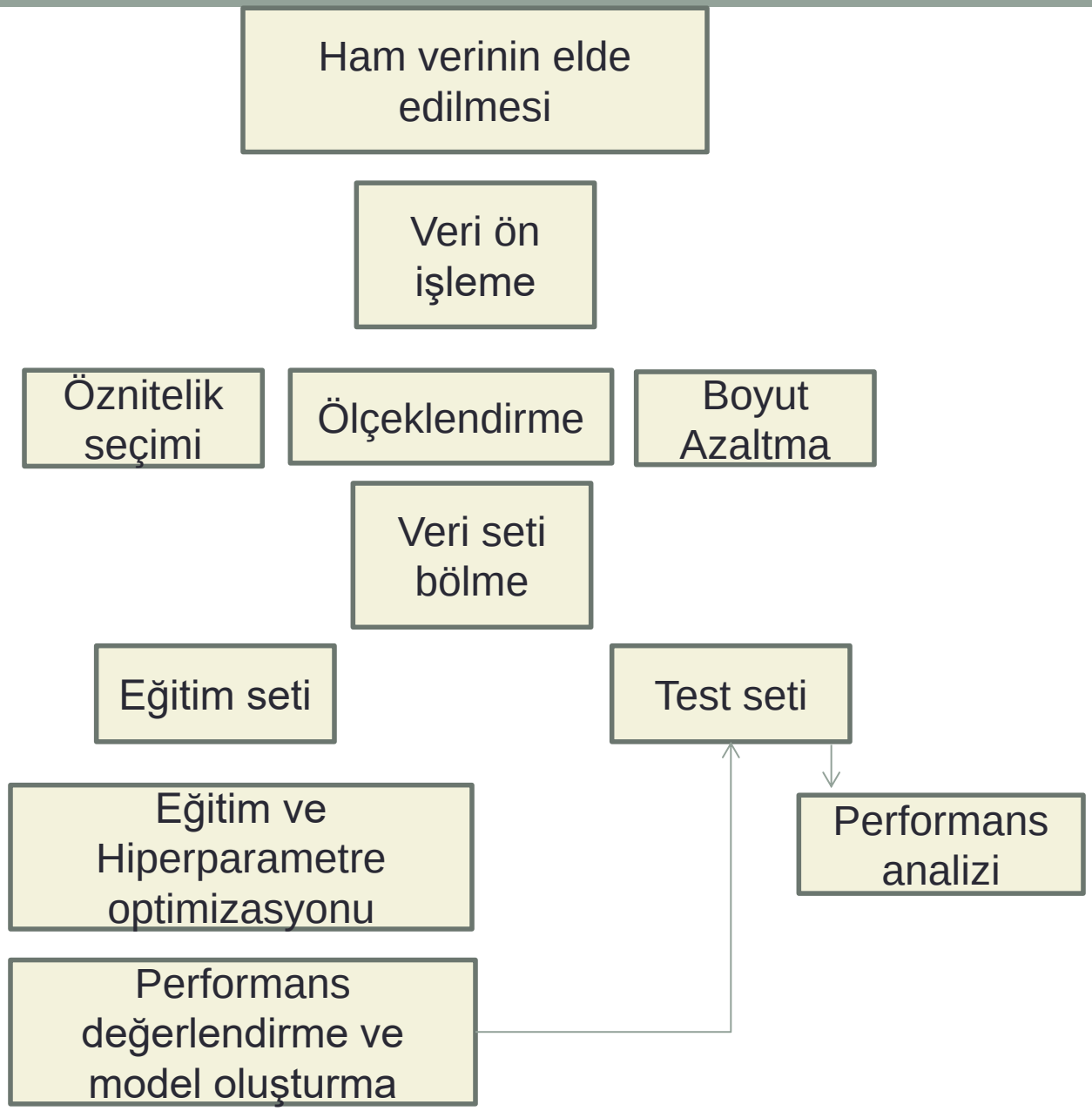


Outlier bulma (AYKIRI-UÇ DEĞER)



Boyut azaltma




Ham verinin elde edilmesi
Veri ön işleme
Öznitelik seçimi
Ölçeklendirme
Boyut Azaltma
Veri seti bölme
Eğitim seti
Test seti
Eğitim ve Hiperparametre optimizasyonu
Performans analizi
Performans değerlendirme ve model oluşturma

Ön işleme

Kayıp veri analizi

- Silme
- Ortalama yazma

Veri temizleme

- Tekrarları yok etme
- Uç noktaların belirlenmesi
- Anlamsız özniteliklerin silinmesi

Kategorik verilerin dönüşümü

- Label encoding (sıralı)
- One hot encoding

Ölçeklendirme

Student	CGPA	Salary '000
1	3.0	60
2	3.0	40
3	4.0	40

Student	CGPA	Salary '000
1	-1.184341	1.520013
2	-1.184341	-1.100699
3	0.416120	-1.100699

- 1 TL ------ 1 Dolar == 1 birim
- -50 – 50 C derece --- 10000 – 100000 km ölçekler çok farklı

Bazı yöntemler için ölçek önemli

Standartizasyon

X_new = (X - mean)/Std

Normalizasyon (min-max scaling) -- outlier varsa iyi sonuç vermez

X_new = (X - X_min)/(X_max - X_min)

Öznitelik seçimi

- N öznitelik varsa → ??? kombinasyon

Öznitelik seçimi

- ***Öznitelik-sınıf arasındaki ilişkiye göre (Filter Methods)***
  i) Pearson Korelasyonu (Pearson Correlation) - sayısal veriler için
  - [-1 1] arası değerler

ii) Ki-Kare Testi (Chi2) – kategorik veriler için
iii) Anova Testi (Anova) - kategorik ile sayısal arasındaki ilişkiyi ölçmek için

- ***Ardışık olarak değişkenleri ekleyerek ve çıkartarak seçme (Sarmal Yöntemler- Wrapped Methods)***
  i) Ardışık İleri Yönde Seçim (Sequential Forward Selection (SFS))
  ii) Ardışık Geri Yönde Seçim (Sequential Backward Selection (SBS))

Öznitelik seçimi

- ***Boyut İndirgeme (Dimensionality Reduction)***
  i) Temel Bileşenler Analizi (Principal Component Analysis (PCA))
  ii) Lineer Diskriminant Analizi (Linear Discriminant Analysis (LDA))

Eğitim – Test seti

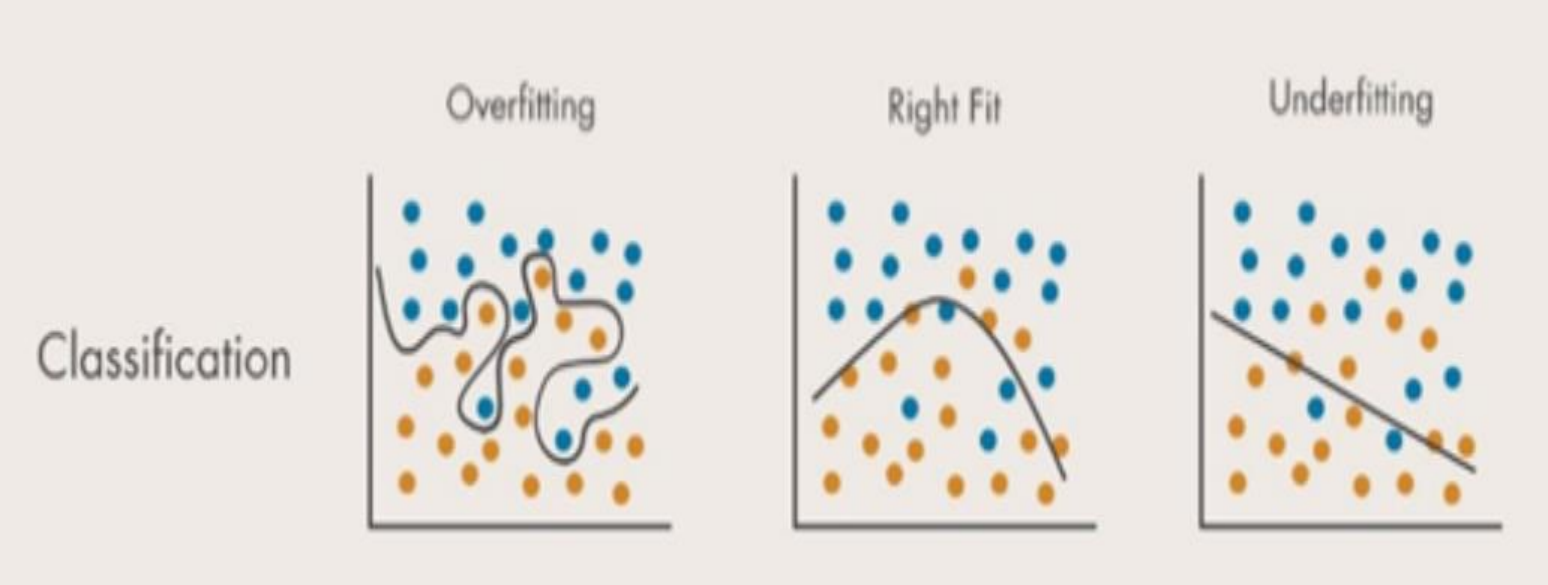
Genelleme (generalization): Test seti ile modelin performansının değerlendirilmesi—oluşan hata genelleme hatası

%80 - %20

Nelere dikkat edilmeli

Eğitim – Test seti

- Overfitting (aşırı öğrenme)



Underfitting: Modelin verilerdeki temel örüntüleri yakalamak için çok basit olması ve bu nedenle kötü performans göstermesi
Overfitting: Modelin çok karmaşık olması nedeniyle verilerdeki gürültüyü veya rastgele dalgalanmaları yakalamaya başlaması ve bu nedenle modelin daha önce karşılamadığı yeni verilere genelleme yaparken kötü performans göstermesi

Validation-doğrulama

Model parametrelerini incelemek için eğitim →test yapılmalı

TEST VERİSİNİ KULLANDIK ?

Eğitim → validation

-eğitim sırasında, eğitim nasıl gidiyor çıkarım yapabiliriz

K-katlı çapraz doğrulama (k-fold cross validation

Temel amacı, modelin farklı veri alt kümesi üzerindeki performansını değerlendirerek genelleştirme kabiliyeti hakkında daha güvenilir ve kararlı bir tahmin yapmaktır

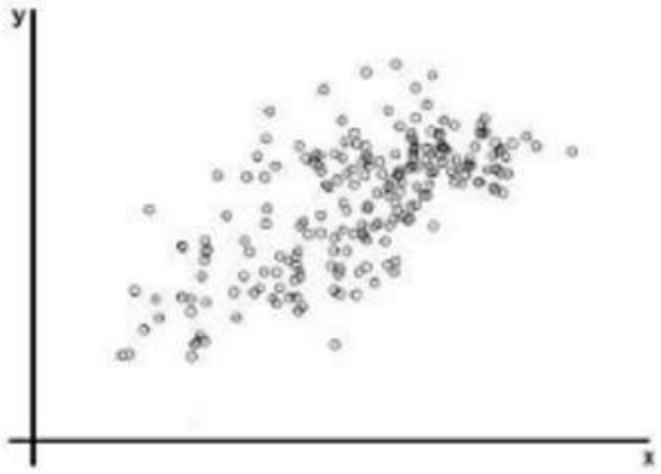
K kere eğitim-test yapılır ortalaması alınır.

Genelde 10-fold

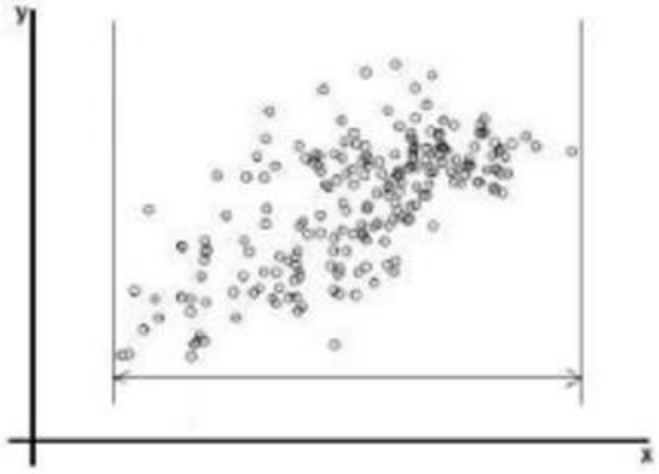
.

PCA- Temel bileşen analizi

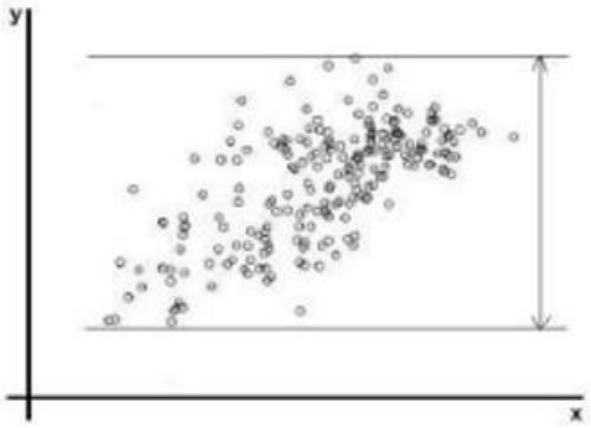
- PCA, veri setindeki toplam varyansın mümkün olan en büyük kısmını ilk birkaç temel bileşenle yakalamaya çalışır. Yani, ilk temel bileşen veri setindeki en yüksek varyansı, ikinci temel bileşen ilk bileşene dik olarak kalan varyansın en yüksek kısmını, ve bu şekilde devam ederek yakalar.



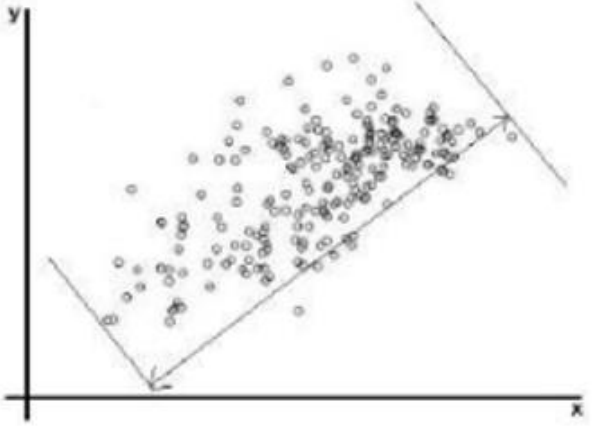
(a) Scatter diagram



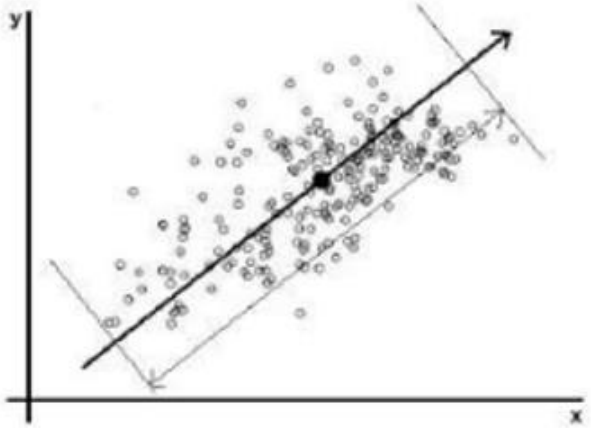
(b) Spread along x-direction



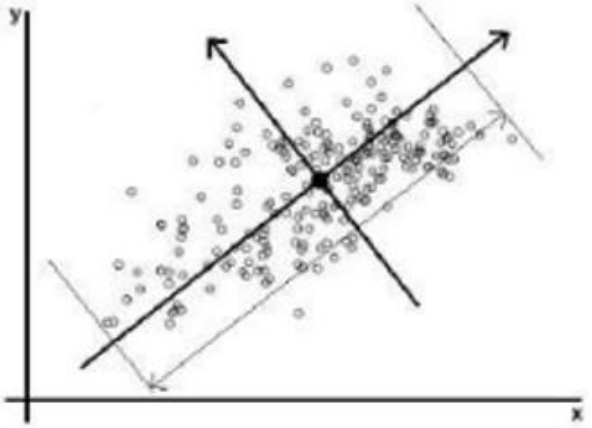
(c) Spread along y-direction



(d) Largest spread



(e) Direction of largest spread : Direction of the first principal component (solid dot is the point whose coordinates are the means of x and y)



(f) Directions of principal components

Features	Example 1	Example 2	$\cdots$	Example N
X_1	X_{11}	X_{12}	$\cdots$	X_{1N}
X_2	X_{21}	X_{22}	$\cdots$	X_{2N}
$\vdots$				
X_i	X_{i1}	X_{i2}	$\cdots$	X_{iN}
$\vdots$				
X_n	X_{n1}	X_{n2}	$\cdots$	X_{nN}

$$\bar{X}_i = \frac{1}{N}(X_{i1} + X_{i2} + \cdots + X_{iN}).$$

$$\text{Cov}(X_i, X_j) = \frac{1}{N-1}\sum_{k=1}^{N}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j).$$

$$S = \begin{bmatrix} \operatorname{Cov}(X_1, X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Cov}(X_2, X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & & & \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Cov}(X_n, X_n) \end{bmatrix}$$

Öz değer bul

$$\det(S - \lambda I) = 0.$$

öz vektör bul

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$(S - \lambda' I)U = 0.$$

$$e_i = \frac{1}{\|U_i\|} U_i, \quad i = 1, 2, \ldots, n.$$

$$X = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_1 & \cdots & X_{1N} - \bar{X}_1 \\ X_{21} - \bar{X}_2 & X_{22} - \bar{X}_2 & \cdots & X_{2N} - \bar{X}_2 \\ \vdots & & & \\ X_{n1} - \bar{X}_n & X_{n2} - \bar{X}_n & \cdots & X_{nN} - \bar{X}_n \end{bmatrix}$$

F seçilen özvektörler olmak üzere

$$X_{\text{new}} = FX.$$

ÖRNEK:

Önceki slyatlardaki adımları kullanılarak, her biri 2 adet öznitelik içeren, 4 örnek için
PCA uygulayın. Tek öznitelik kullanılması durumunda bu öznitelik ne olur, hesaplayın

Feature	Example 1	Example 2	Example 3	Example 4
X_1	4	8	13	7
X_2	11	4	5	14