

Machine Learning Giriş

BTK Akademi



Büyük Veri Dönemi

- İnsanlığın ürettiği veri miktarı eksponansiyel olarak artmakta.
- Üretilen verinin karmaşıklığı da veri miktarıyla birlikte artmakta.
- Mevcut veri miktarını insan eliyle incelemek maliyetli, zaman alıcı ve istenen performansı vermemekte.
- Mevcut veri miktarını insan eliyle incelemek maliyetli, zaman alıcı ve istenen performansı vermemekte.
- Yeni teknolojiyle birlikte büyük miktarda verileri işleyebilecek işlemciler ucuza üretilmekte.
- **Makine öğrenimi**, bu yeni dönemin olmazsa olmaz bir parçası olmuş durumda.

Pek çok sektörde düzenli olmayan ve insan eliyle işlenemeyen büyük miktarda veri bulunmakta.

- **Ticari Faaliyetler**
 - Speech recognition
 - Reklamcılık, kişiye özel pazarlama, kişiye özel sağlık hizmetleri
- **Güvenlik Faaliyetleri**
 - Spam mail tespiti, dolandırıcılık faaliyetleri tespiti
 - Görüntü işleme, gözetleme, takip faaliyetleri
- **Bilimsel Faaliyetler**
 - Ekonomik faaliyetler, piyasa tahmini
 - Biyoloji, astronomi, jeoloji, neuroscience, fizik, vb.

Büyük Veri Kullanım Alanları

Speech recognition: http://y2u.be/JvbHu_bVa_g

Görüntü işleme: <http://y2u.be/50NPqEla0CQ>

İniş yapan roketler: <http://y2u.be/gsliniJMr3E>



Temperature

72° F

STM TOGAN



Chess



AlphaZero vs. Stockfish

W:29.0% D:70.6% L:0.4%



W:2.0% D:97.2% L:0.8%

Shogi



AlphaZero vs. Elmo

W:84.2% D:2.2% L:13.6%



W:98.2% D:0.0% L:1.8%

Go



AlphaZero vs. AGO

W:68.9% L:31.1%

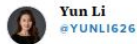


W:53.7% L:46.3%

AZ wins AZ draws AZ loses AZ white AZ black

80% of the stock market is now on autopilot

PUBLISHED SAT, JUN 29 2019-8:30 AM EDT | UPDATED SAT, JUN 29 2019-8:31 AM EDT



Yun Li
@YUNLI626

SHARE f t in

Büyük Veri Kullanım Alanları

Spam Mail Tespiti

☆ **Osman Khan** to Carlos [show details](#) Jan 7 (6 days ago) [Reply](#) ▼

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

☆ **Carlos Guestrin** to 10615-announce, Osman, Miche [show details](#) 3:15 PM (8 hours ago) [Reply](#) ▼

Hi everyone,

Welcome to New Media Installation: Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only \$5.95 for shipping mfw rik [Spam](#) | [X](#)

☆ **Jaquelyn Halley** to nherriein, bcc: thehorney, bcc: ang [show details](#) 9:52 PM (1 hour ago) [Reply](#) ▼

=== Natural WeightLOSS Solution ===

Vital Acai is a natural WeightLOSS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

- * Rapid WeightLOSS
- * Increased metabolism - BurnFat & calories easily!
- * Better Mood and Attitude
- * More Self Confidence
- * Cleanse and Detoxify Your Body
- * Much More Energy
- * BetterSexLife

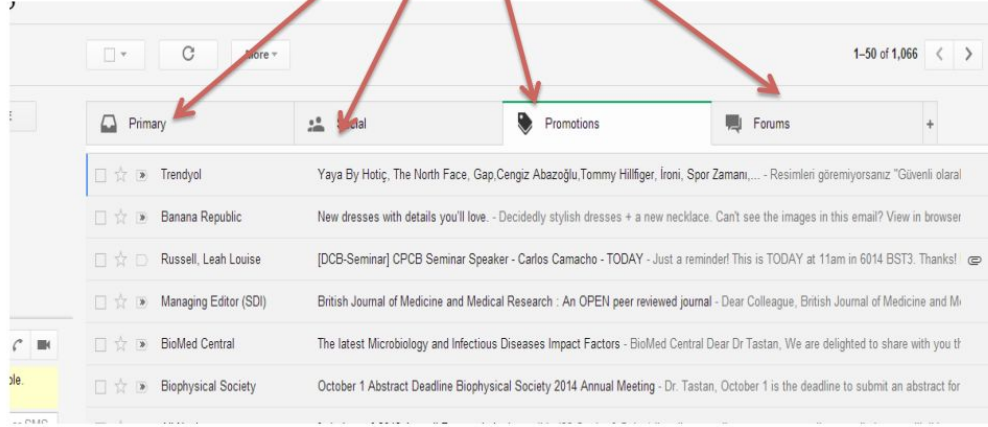
Ham

Spam

Büyük Veri Kullanım Alanları

Metin Sınıflandırma

Incoming e-mail

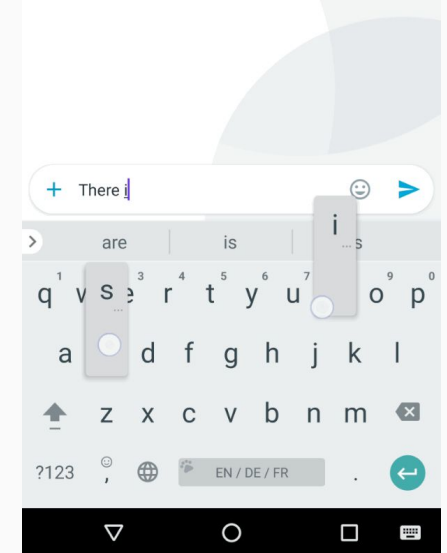


Varlık İsmi Tanıma

LOC 1 ORG o PERSON p OTHER C-o

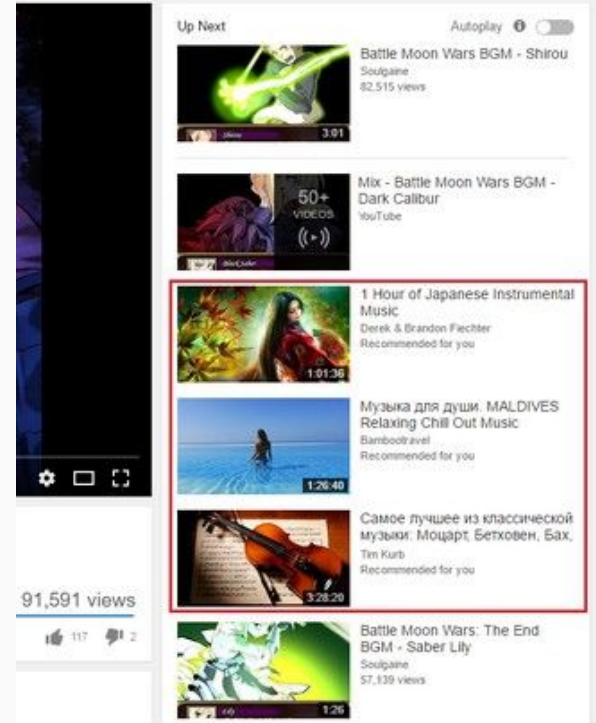
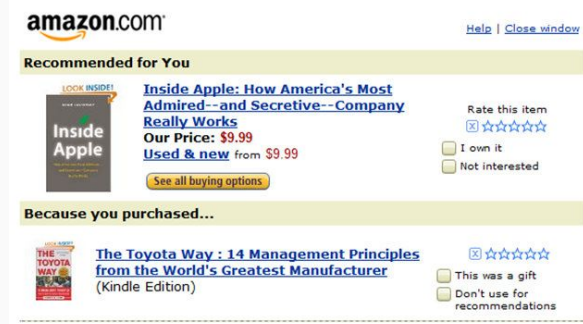
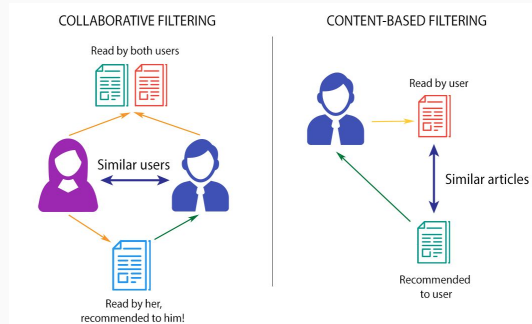
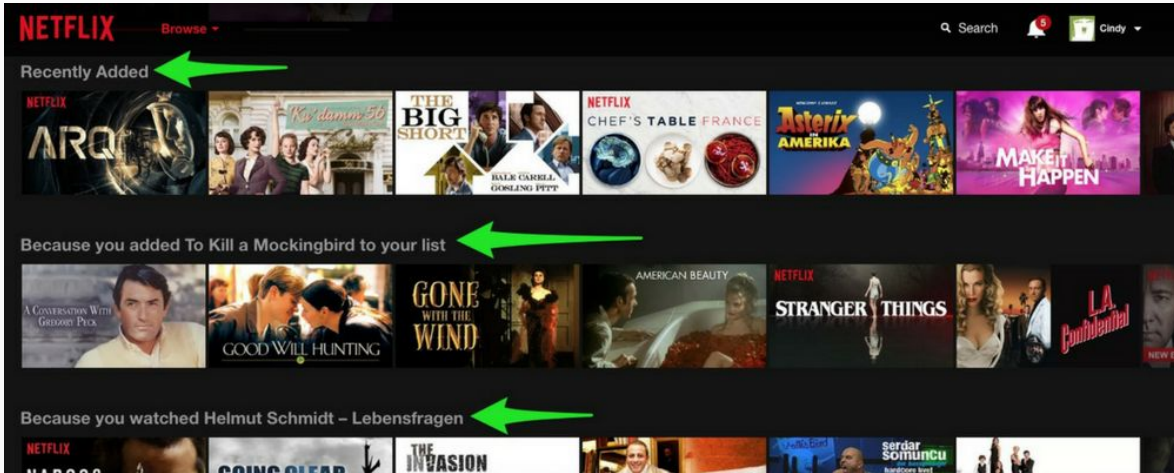
Hatayspor x Galatasaray x maci oncesinde ev sahibi ekibin
tarafarlari Galatasaray Teknik Direktoru Fatih Terim'i x tribune
cagirip Kebapci Selo tezahuratinda bulundu.

Autocomplete



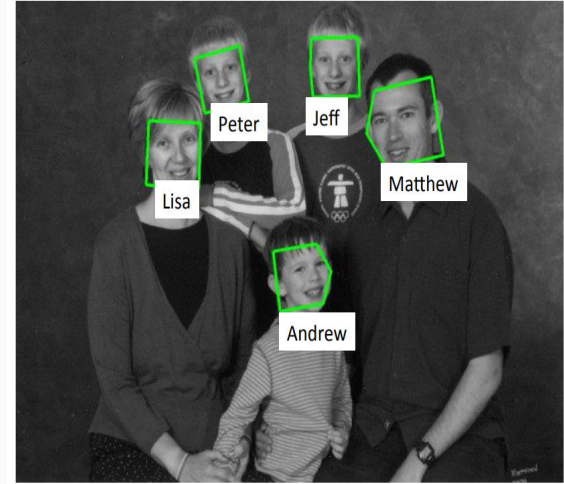
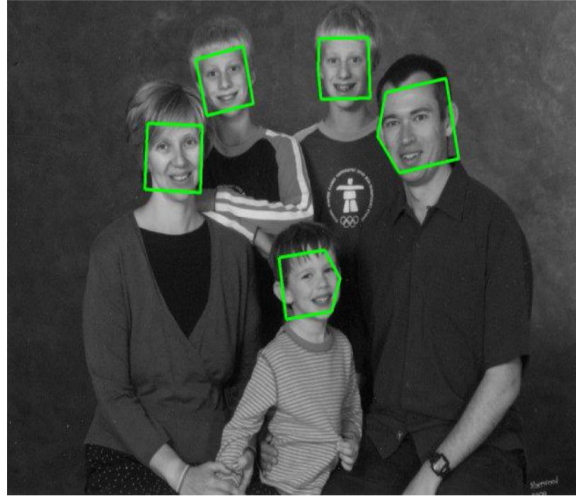
Büyük Veri Kullanım Alanları

Öneri Sistemleri



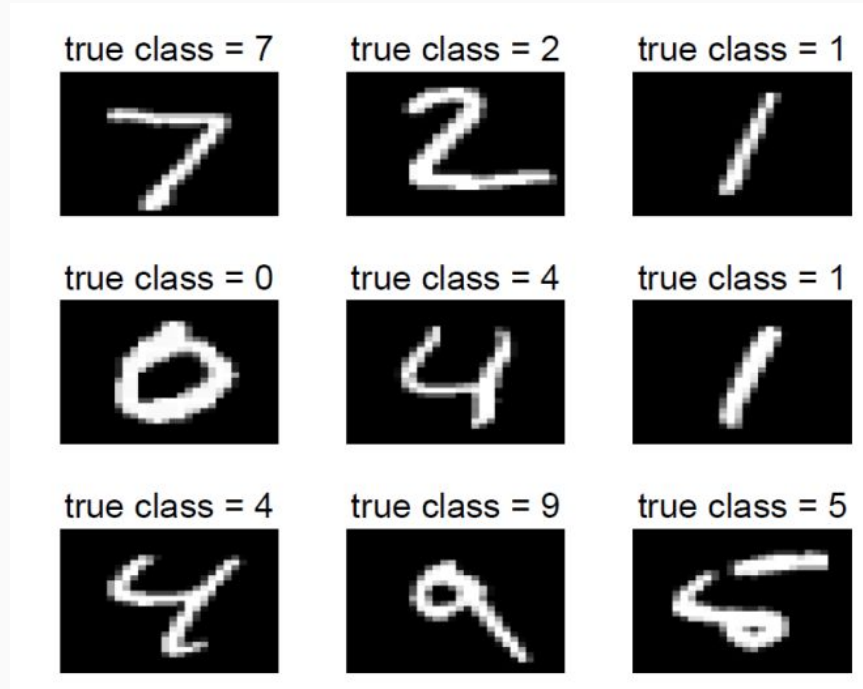
Büyük Veri Kullanım Alanları

Yüz Tespiti



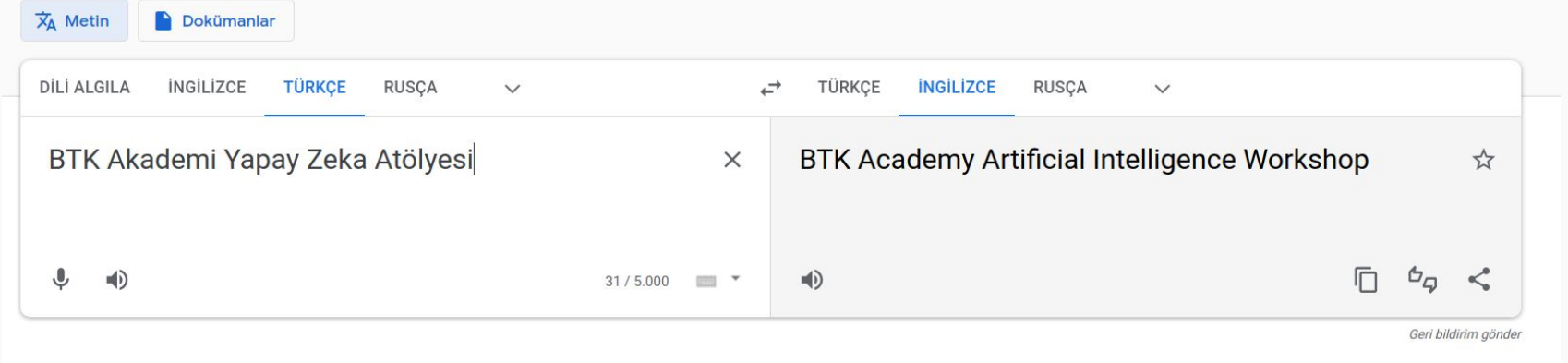
Büyük Veri Kullanım Alanları

Resim Sınıflandırma

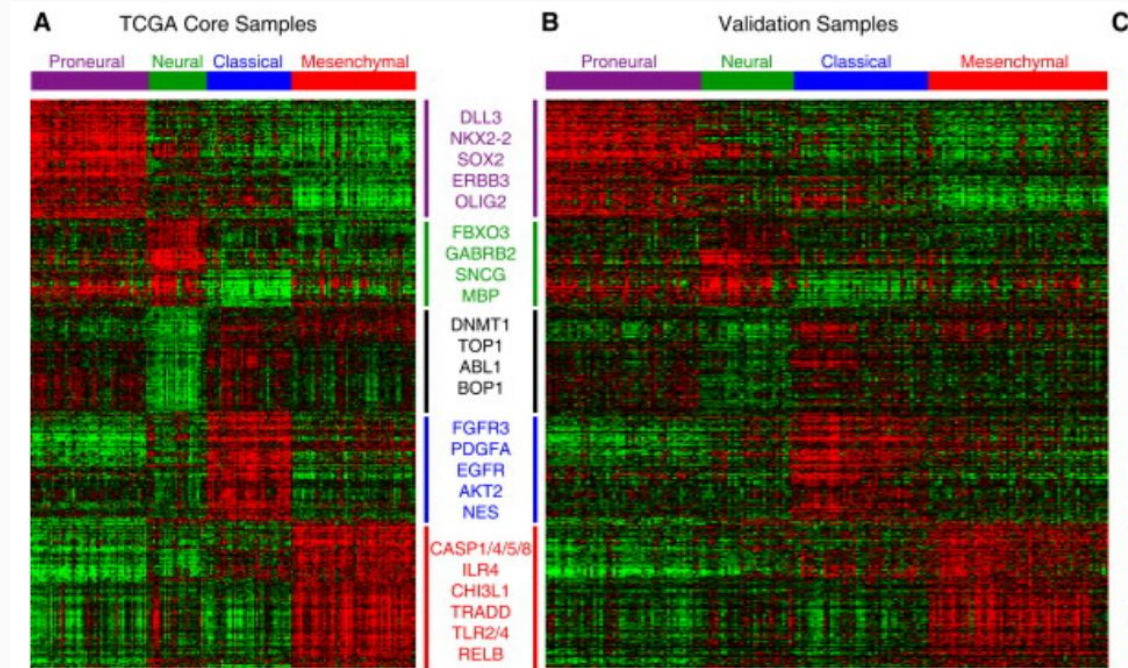


Büyük Veri Kullanım Alanları

Makine Çevirisi



Kanserli Hücre Tespiti



Verhaak et al., Cancer Cell, 2010.

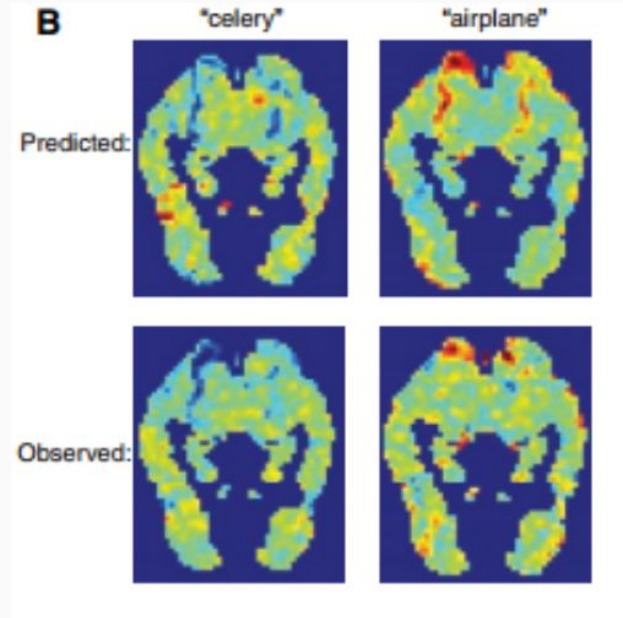
Büyük Veri Kullanım Alanları

Görselden Metne Çevirme



Automatically captioned: "Two pizzas sitting on top of a stove top oven"

FMRI Görsellerinin Analizi / Hastalık Tespiti



Büyük Veri Kullanım Alanları

Sağlık



Büyük Veri Kullanım Alanları

Görüntü Yaratma



Büyük Veri Kullanım Alanları

Deepfake

https://www.youtube.com/watch?v=F4G6GNFz0O8&ab_channel=DiepNep

Star Wars: Rogue One, Princess Leia (Carrie Fisher)



The Mandalorian: Book of Boba Fett, Luke Skywalker (Mark Hamill)



Makine öğrenimi ne zaman kullanışlıdır?

- İnsan bilgisi ve uzmanlığı yeterli değil.
 - Marsa yolculuk
 - İnsanüstü satranç becerisi
- İnsan bilgisi ve uzmanlığı var, fakat modellenemez değil.
 - Speech-to-text
 - speech recognition
 - image recognition
- Belirli bir alanda insan bilgisi ve uzmanlığı var, fakat farklı bir alana aktarılmak isteniyor.
 - Kişiselleştirilmiş araçlar

Makineler nasıl “öğrenir”?

- İnsanlar yeteneklerini geliştirmeyi zaman içerisinde pratikle öğrenirler.
- Makine öğrenmesinin amacı **tecrübe** ile gelişebilen algoritmalar kurmaktır.
 - Makine öğrenmesi algoritmaları için tecrübe **verilerden** gelir.
- **Veriler** nereden gelir?
 - İnsanlar tarafından yapılan veri etiketlemeleri
 - Otonom veri toplama/üretme

Makine öğrenimi amaçları nelerdir?

- Algoritmalar:
 - Büyük çapta problemlere hızlı çalışan, yüksek başarımlı, genelleştirilmiş çözümler sunmak
 - Daha önce görmediği örneklerde de başarılı tahminler yapmak
 - Farklı alanlarda ve problemlerde kullanılabilir olmak

Örnek:

- Veri içerisindeki tek bir obje veya sayı

Feature (Öznitelik) (X ile gösterilir):

- Bir örneğe ait olan, genellikle numerik formda bulunan özellik. Örneğe ait bilgi. Örn. bir kişiye ait boy, yaş, kilo, cinsiyet

Label (Etiket) (Y):

- Öznitelikler aracılığıyla tahmin edilmek veya bulunmak istenen bilgi.
- Sınıflandırma problemlerinde örneğin ait olduğu kategori bilgisi
- Regresyon problemlerinde numerik bir değer

Eğitim veri seti:

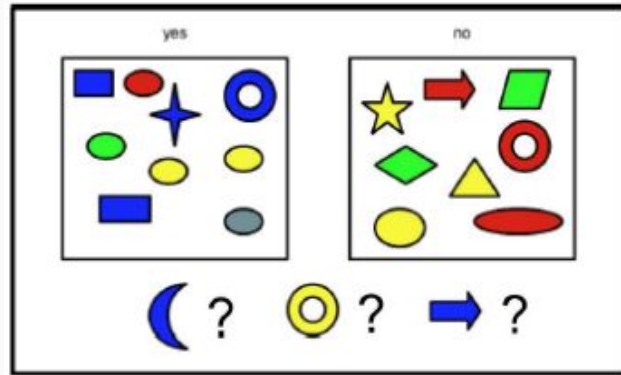
- Model eğitimi amacıyla kullanılan veri seti

Test veri seti:

- Model başarımını ölçmek amacıyla kullanılan veri seti

Feature Extraction (Veri Çıkarma):

- Ham haldeki veri içerisinde çıkarılan anlamlı bilgiler, veriyi işlenebilir hale getirme, veri içerisinde özellikler yaratma işlemi



(a)

D features (attributes)			Label
Color	Shape	Size (cm)	
Blue	Square	10	
Red	Ellipse	2.4	
Red	Ellipse	20.7	0

(b)

Feature Extraction (Veri Çıkarma):

Örnek: Çiçek sınıflandırma



Setosa

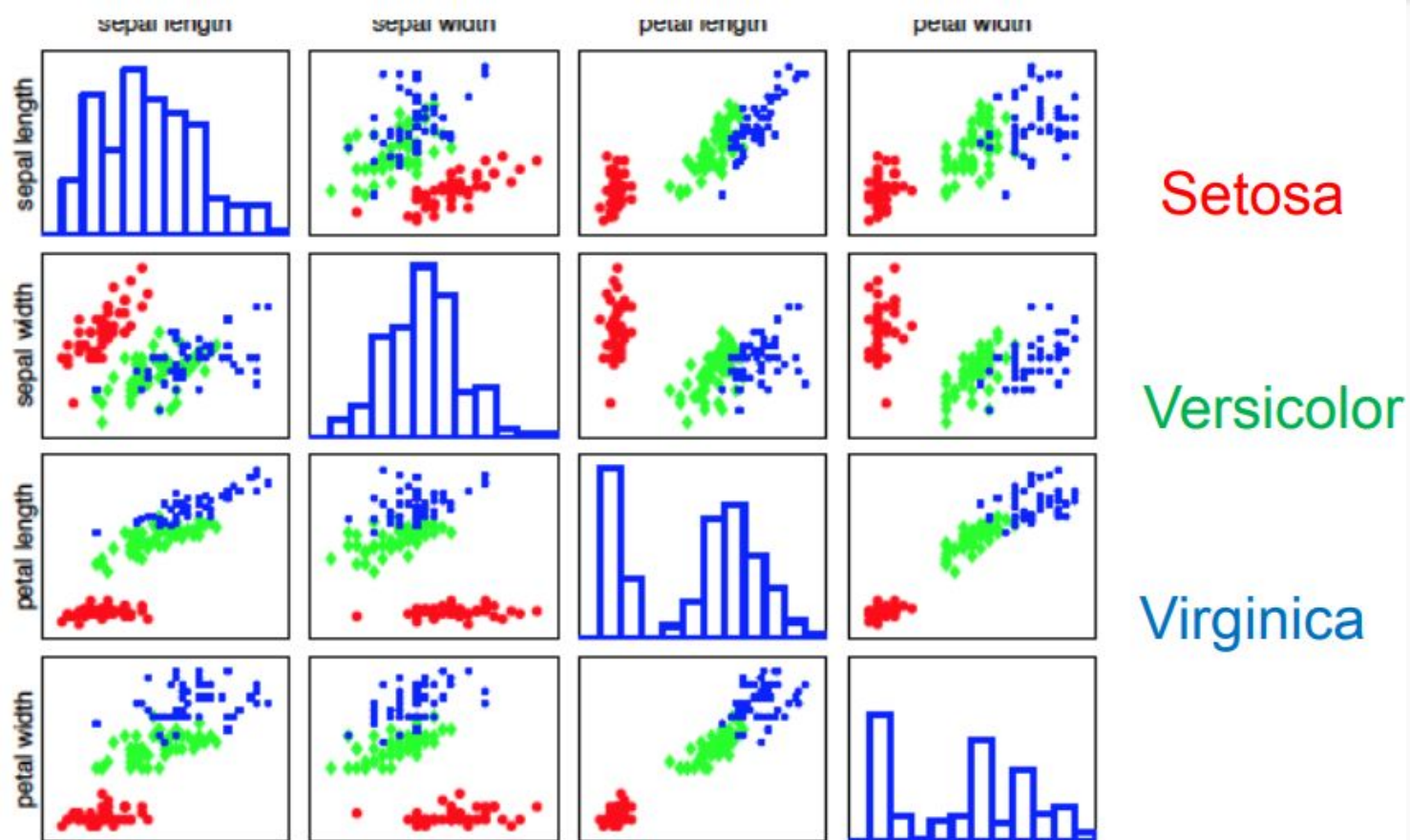


Versicolor



Virginica

Tanımlar



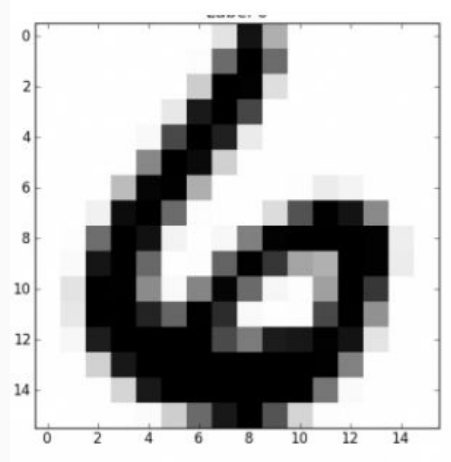
Feature Extraction (Veri Çıkarma):

Örnek: El Yazısı Tanıma



Feature Extraction (Veri Çıkarma):

Örnek: El Yazısı Tanıma

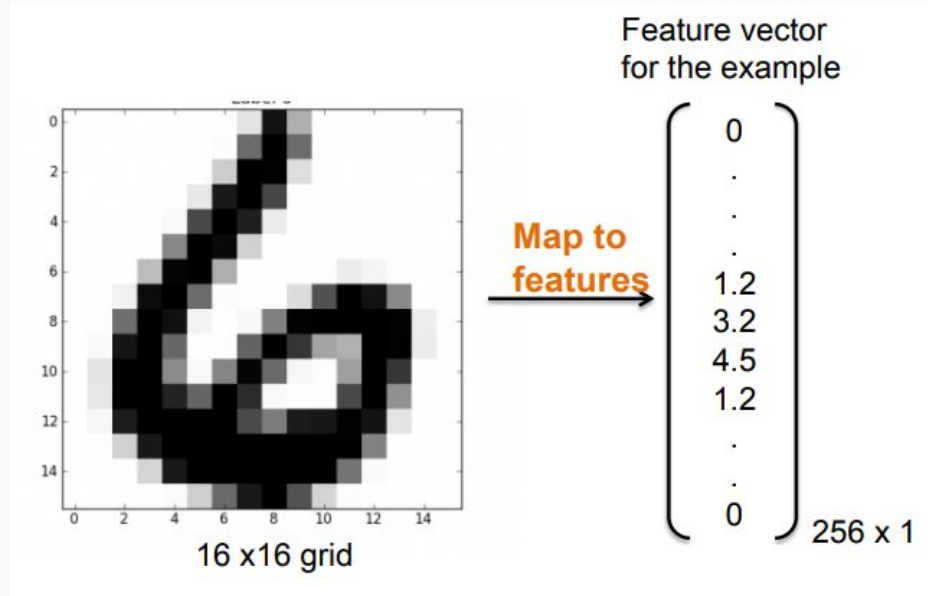


$$Y \in \{0, 1, \dots, 9\}$$

Bu örneğin etiketi: 6

Feature Extraction (Veri Çıkarma):

Örnek: El Yazısı Tanıma



Training set

$$D_{train} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

000000000
111111111
222222222
333333333
444444444
555555555
666666666
777777777
888888888
999999999

+ their labels

Test set

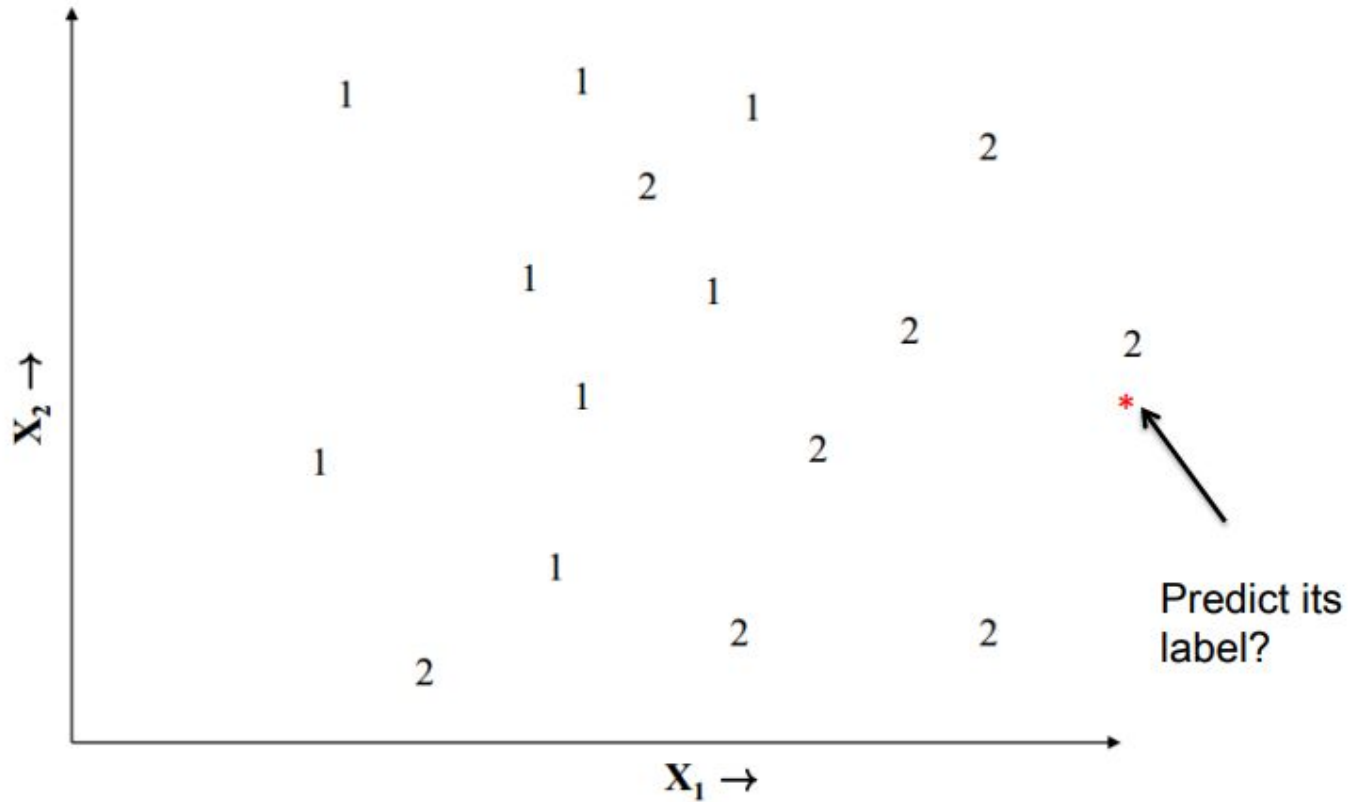
$$D_{test} = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$$

00000
11111
22222
33333
44444
55555
66666
77777
88888
99999

+ their labels

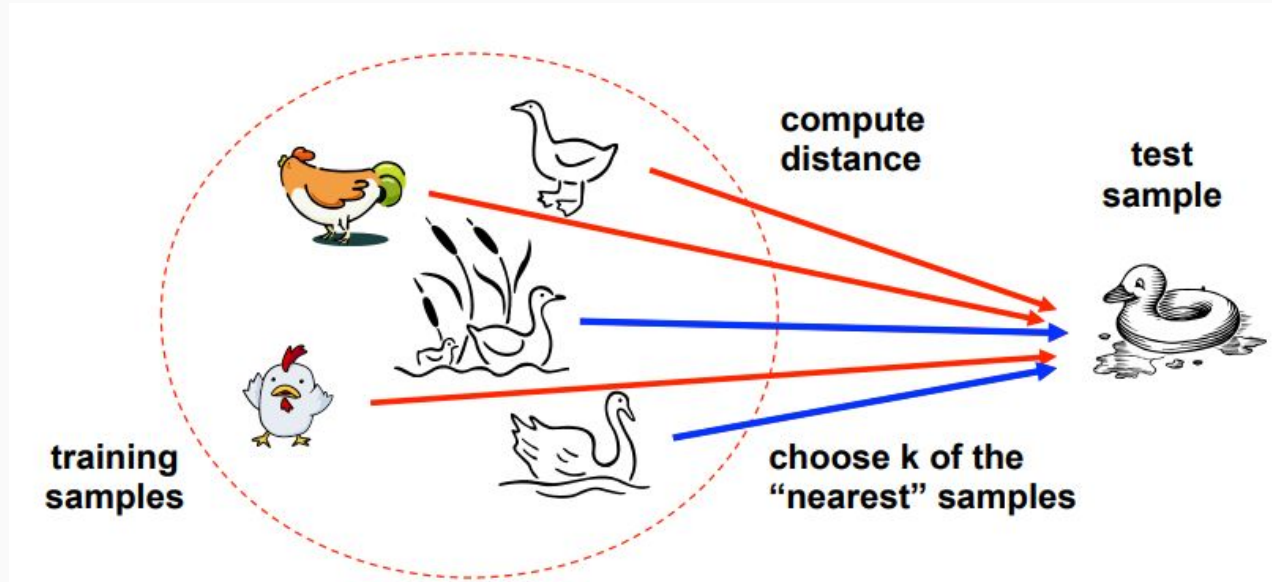
Should not be overlapping

Bu Örneği Nasıl Sınıflandırırsınız?

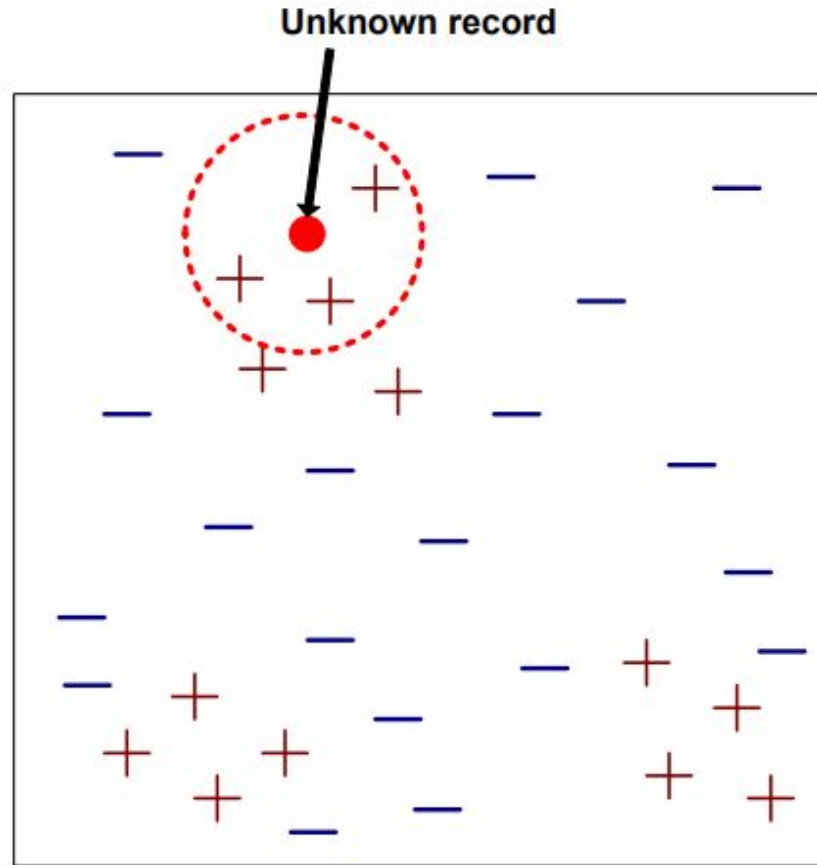


Nearest Neighbor Classifiers

Temel fikir: Ördek gibi yürüyorsa, ördek gibi ses çıkarıyorsa, ördek gibi yüzüyorsa, o zaman bir ördektir.

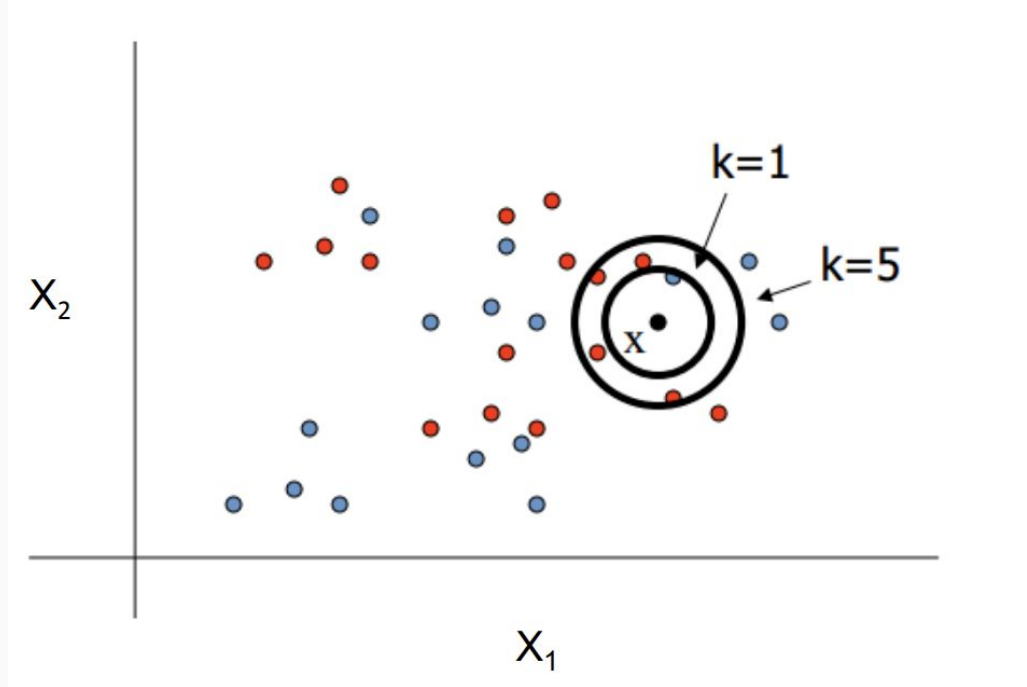


Nearest Neighbor Classifiers

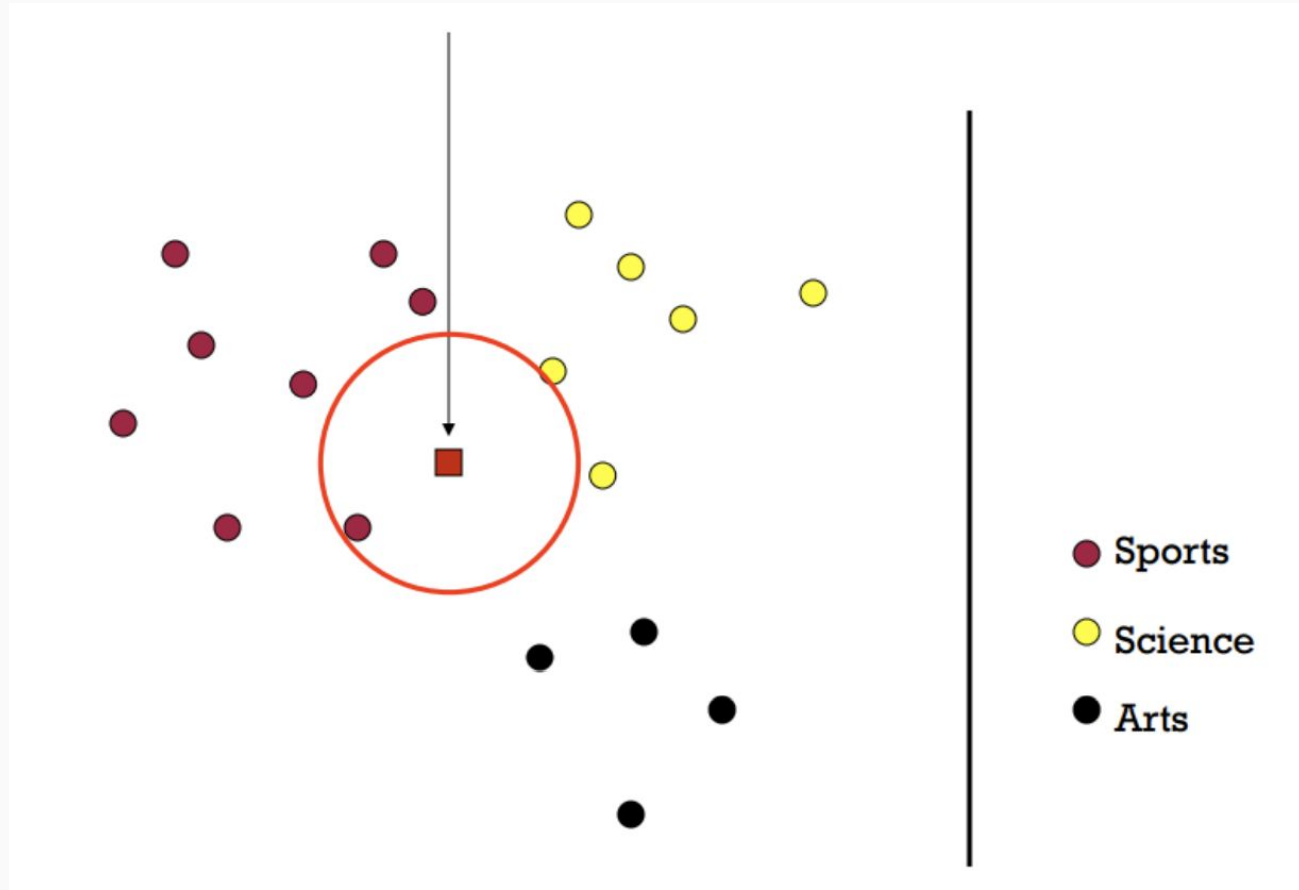


KNN Sınıflandırıcı

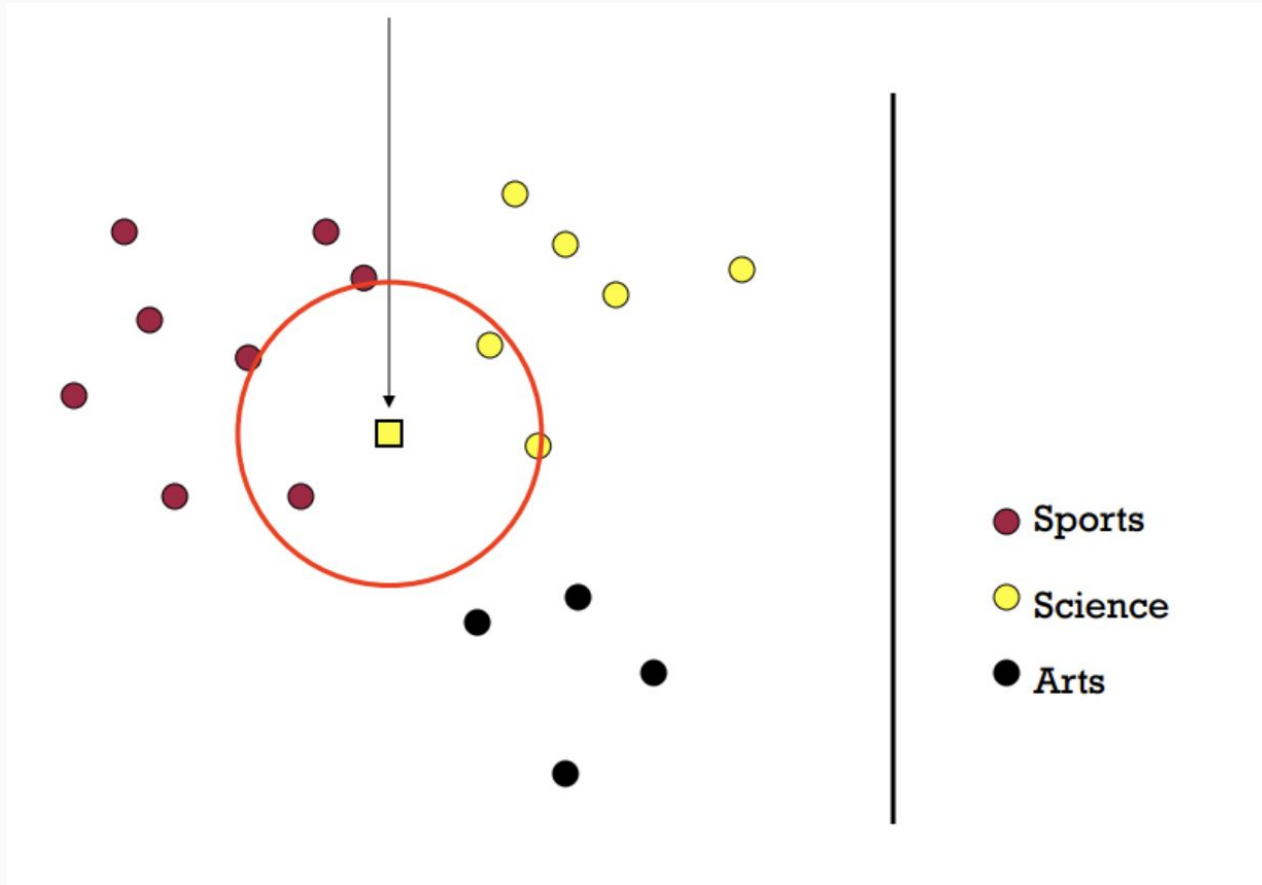
Gelen bir örneği sınıflandırmak için örneğin vektör uzayındaki konumunu analiz et ve en yakınındaki örneklerle göre bir sınıf ata.



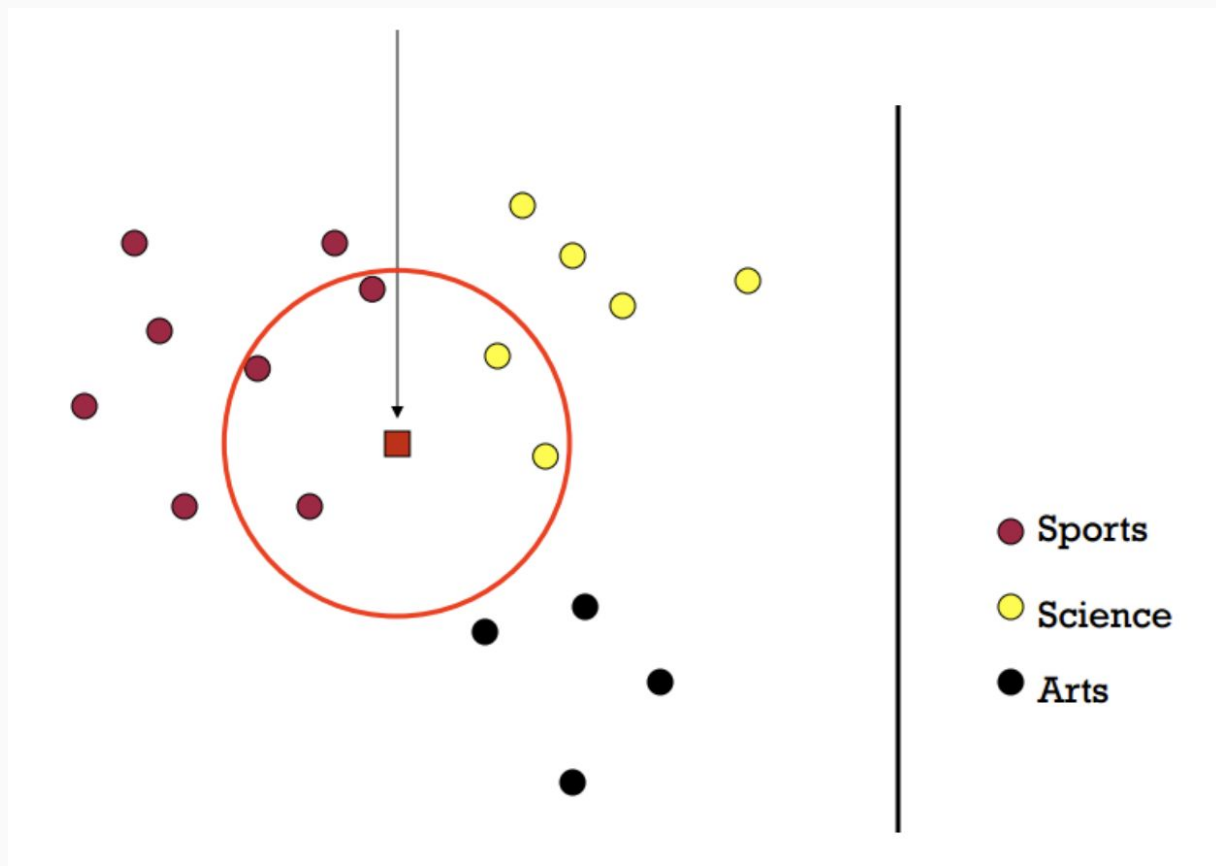
1-NN (K=1)



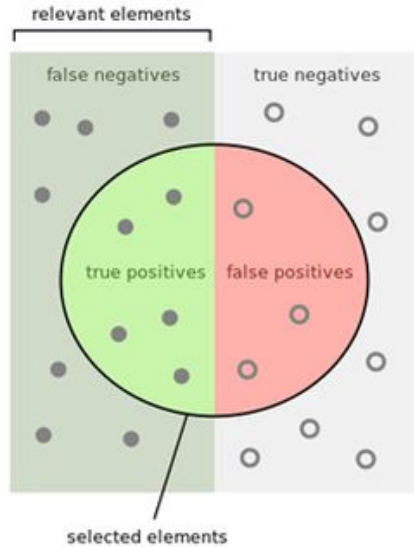
3-NN (K=3)



5-NN (K=5)



Performans Metrikleri



How many selected items are relevant?

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

		Actual Labels	
		1	0
Predicted Labels	1	True Positive	False Positive
	0	False Negative	True Negative

Performans Metrikleri

		Actual Labels	
		1	0
Predicted Labels	1	True Positive	False Positive
	0	False Negative	True Negative

(Is your prediction correct?) (What did you predict)



(Your prediction is correct)

(You predicted 0)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F1 score} = 2 \times \frac{(\text{Prec} \times \text{Rec})}{(\text{Prec} + \text{Rec})}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{False +ve rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

$$\text{Recall, Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Makine Öğrenimi - Problemlerin Modellenmesi

Fonksiyonel Temsil

Problemi modellerken amaç öznitelikler (X) ile etiketler (Y) arasında bir fonksiyon öğrenebilmektir.

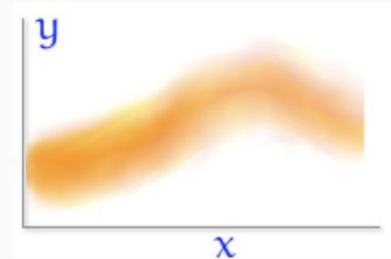
Örneğin:

$X = [\text{yaş, kilo, kan şekeri oranı, ...}]$, $y = \{\text{diyabet hastası, sağlıklı}\}$

Modelleme yapılırken X ile Y arasında bir ilişki olduğu varsayımı yapılmaktadır.

Örnek hipotez: yaşı yüksek, kilosu fazla ve kan şekeri oranı yüksek kişilerin diyabet hastası olma olasılığı daha yüksek olabilir.

Bu ilişki matematiksel olarak $P(X, Y)$ olarak ifade edilebilir.



Farklı Öğrenme Yöntemleri

Etiketin var olup olmamasına göre öğrenme biçimleri:

Supervised Learning: $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$

Unsupervised Learning: $\{X_1, X_2, \dots, X_n\}$

Semi-supervised Learning: ikisinin karışımı.

Etiket Tipleri:

- Binary Classification $Y \in \{0, 1\}$
- Multi class Classification $Y \in \{0, 1, \dots, K\}$
- Regression $Y \in \mathbb{R}$
- Structure Prediction Y kompleks bir yapıya sahip (graph, multi dim array vb.)

Regresyon problemleri

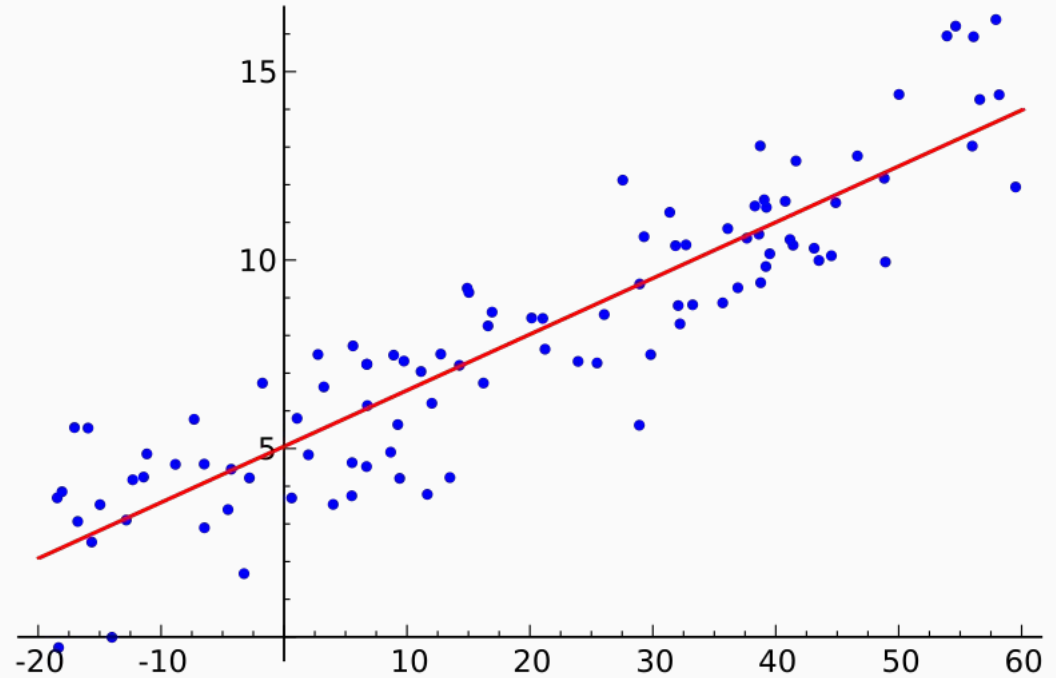
- Geçmiş verilerden yola çıkarak önümüzdeki haftaki bitcoin fiyatını tahmin etmek.
- Youtube geçmişinden yola çıkarak kişinin yaşını tahmin etmek.
- Dinlediği müziklerden yola çıkarak kişinin ne kadar mutlu olduğunu tahmin etmek.
- Bir kişinin video görüntülerinden yola çıkarak odanın sıcaklığını tahmin etmek.

Sınıflandırma problemleri

- Çalışan verisi kullanarak kişinin işten ayrılıp ayrılmayacağını tespit etmek
- Spam mail tespiti
- Fotoğrafta kedi mi var köpek mi var tespiti

Linear Regression

$$y(x) = w_0 + w_1 * x$$



Linear Regression

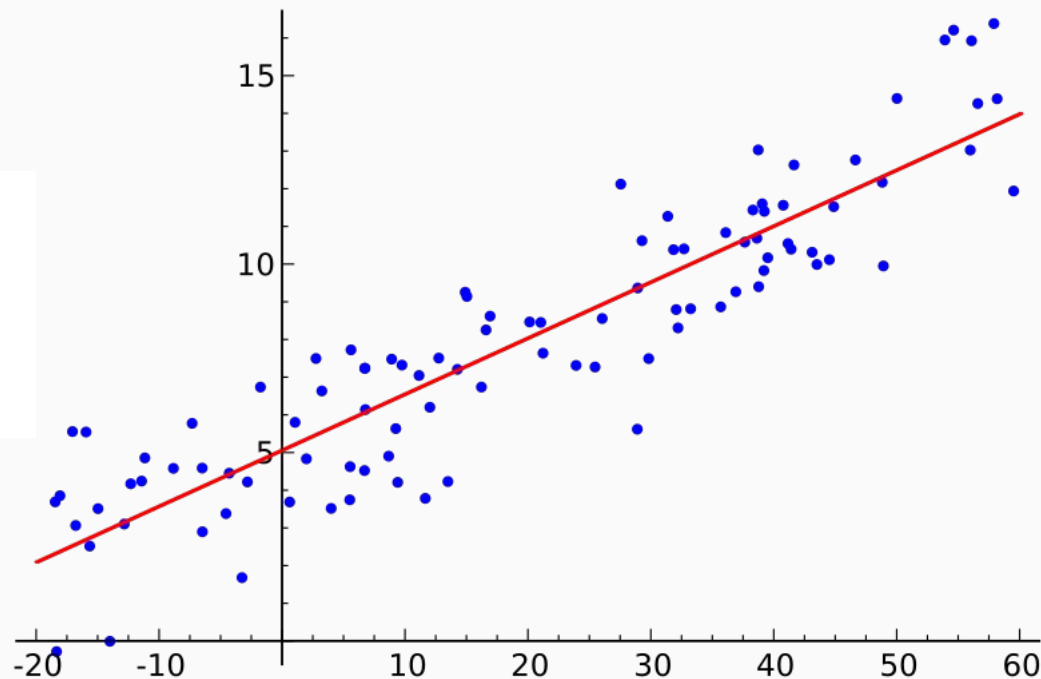
$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + \epsilon = \sum_{j=1}^D w_j x_j + \epsilon$$

Outcome variable (real-valued labels) $\rightarrow y(\mathbf{x})$

Coefficient vector $\rightarrow \mathbf{w}$

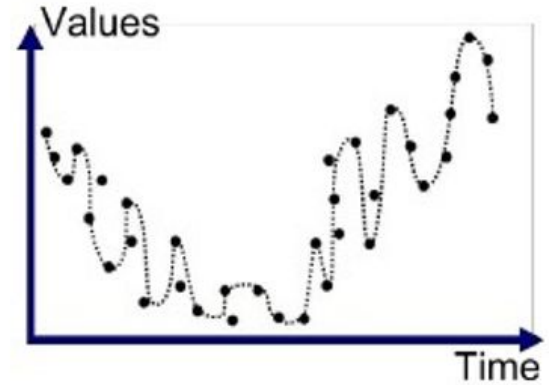
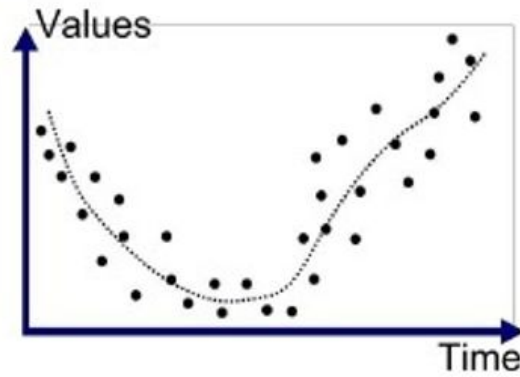
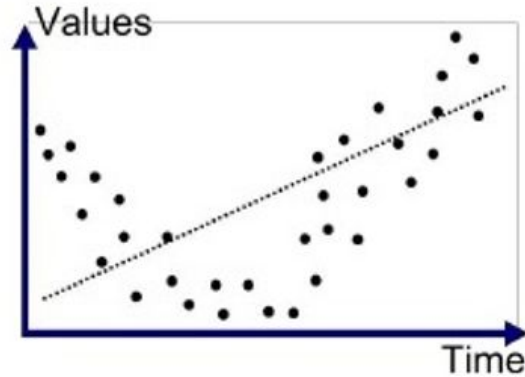
Feature vector $\rightarrow \mathbf{x}$

Residual error $\rightarrow \epsilon$



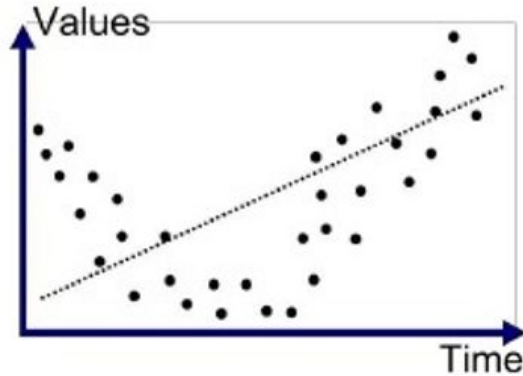
Karşılaşılabilecek Sıkıntılar

Overfitting / Underfitting: Hangi görseldekini tercih ederdiniz?

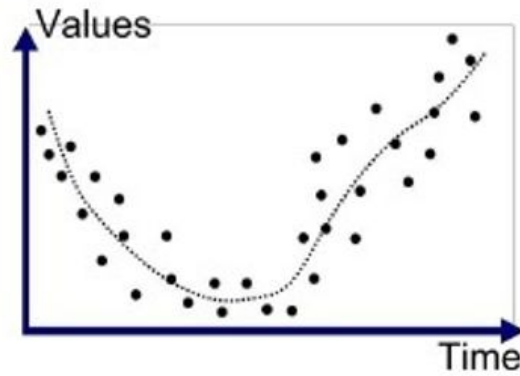


Karşılaşılabilecek Sıkıntılar

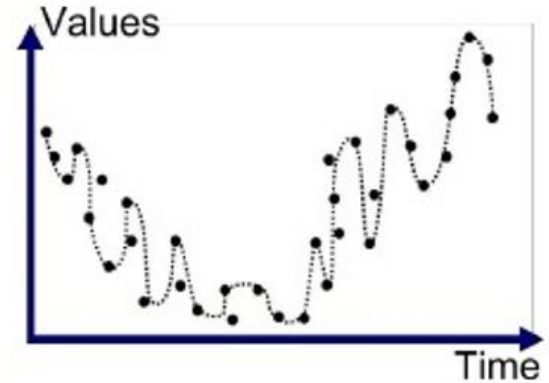
Overfitting / Underfitting: Hangi görseldekini tercih edersiniz?



Underfitted



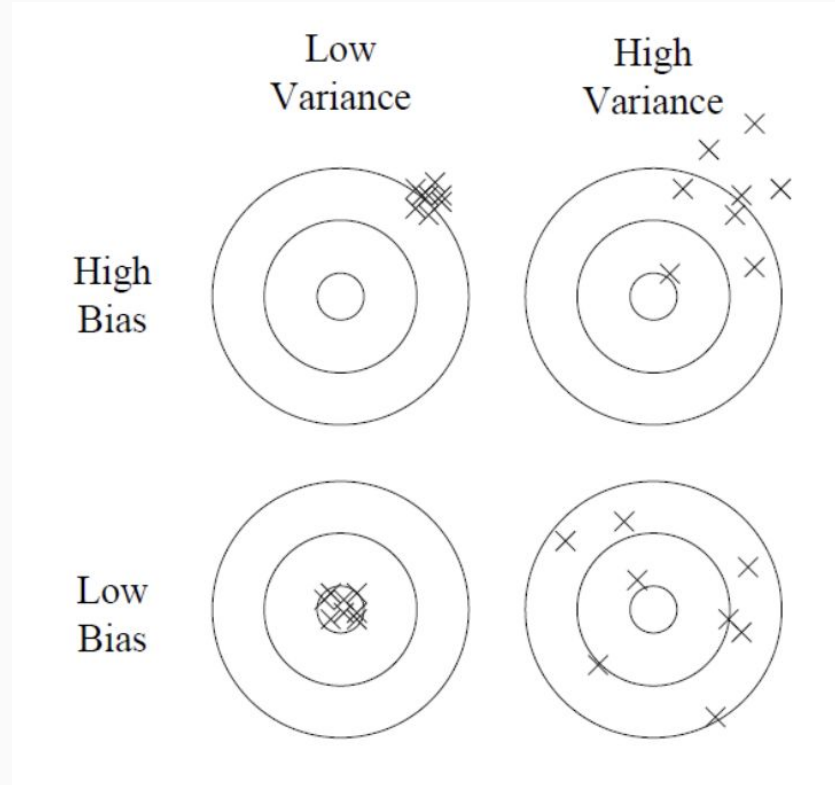
Good Fit/Robust



Overfitted

Karşılaşılabilecek Sıkıntılar

Bias / Variance

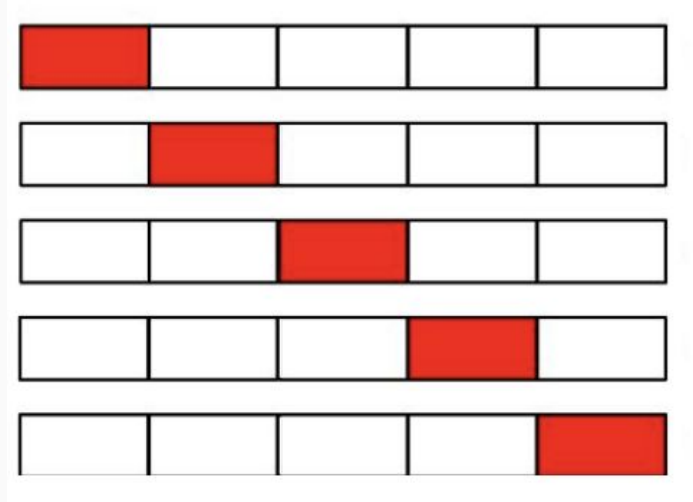


Model Seçimi ve Değerlendirmesi

Altın kural: ASLA test verisi ile model eğitme!

Test verisi farklı şekillerde seçilebilir.

- Random bir şekilde train-test diye iki sete bölmek.
- K-Fold Cross-Validation kullanmak.



Curse of Dimensionality

Büyük boyutun laneti:

- Veriye her bir yeni öz nitelik eklendiğinde vektör uzayı iki katına çıkar.
- Vektör uzayı iki katına çıkarken verideki örnek sayısı bunu karşılayamazsa veri seyrekleşir, ve performans kaybı yaşanır.

Curse of Dimensionality

Büyük boyutun laneti:

- Veriye her bir yeni öz nitelik eklendiğinde vektör uzayı iki katına çıkar.
- Vektör uzayı iki katına çıkarken verideki örnek sayısı bunu karşılayamazsa veri seyrekleşir, ve performans kaybı yaşanır.

Laneti Kırma Yöntemleri:

- **Feature Selection:** Veri içerisinde en değerli öz niteliklerin tespit edilmesi ve bunlarla bir model eğitilmesi. Yapılması uzun süren maliyetli bir işlemdir.
- **Dimensionality Reduction:** Veri içerisindeki belli desenleri tespit ederek minimum bilgi kaybıyla aynı veriyi daha az öz nitelikte temsil etme işlemidir.
 - Principal Component Analysis (PCA)
 - Singular Value Decomposition (SVD)

PCA



Mean

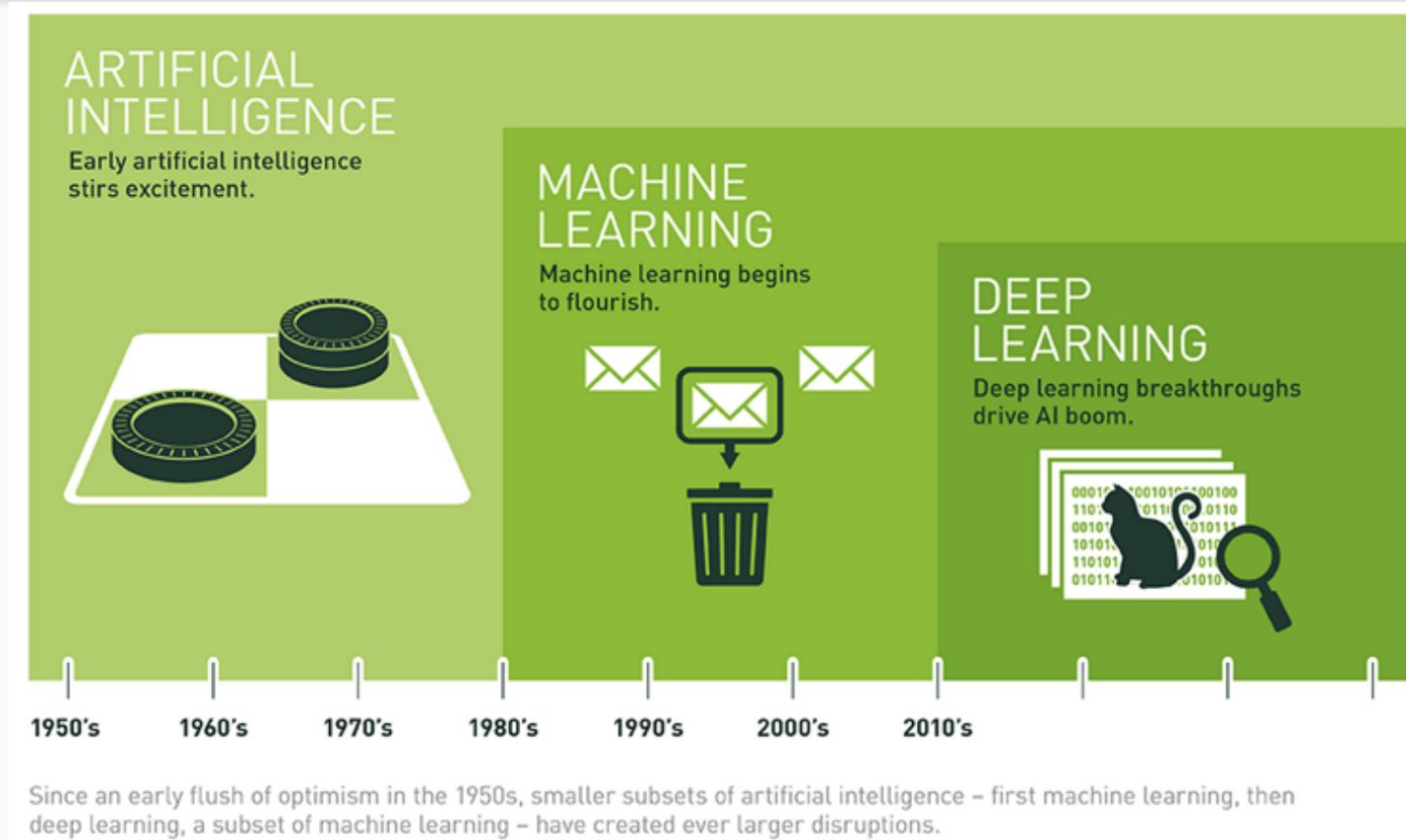


Principal Component 1



Principal Component 2

Yapay Zeka vs Makine Öğrenimi vs Derin Öğrenme



Kurs ne içeriyor?

- Keşifçi Veri Analizi
 - Python ile veri analizi, veri görselleştirme
 - Veri içerisindeki ilişkilerin tespiti
- Veri Mühendisliği
 - Feature Extraction (Öznitelik Çıkarma Teknikleri)
 - Feature Selection (Öznitelik Seçme Teknikleri)
- Yapay Zeka Modelleri Eğitimi Pratik Bilgiler
 - Az teori ve çok pratik ile makine öğrenimi uygulamaları,
 - Farklı makine öğrenimi ve derin öğrenme modelleri hakkında genel bilgi
 - Model başarımların metrikleri, model başarımların ölçümü
 - Model optimizasyonu
- Model Serving
 - FastAPI aracılığıyla eğitilen modelin kullanıcılara sunulması

Kurs ne içermiyor?

- Python dersleri
- Makine öğrenimi modelleriyle ilgili derinlemesine teorik bilgiler
- Unsupervised learning, semi-supervised learning, reinforcement learning
- CNN, RNN, LSTM, BERT, GPT gibi derin öğrenme modellerinin teknik anlatımları (genel bilgi olarak anlatılacak)

Q&A