Makine Öğrenmesine Giriş

Burak Tahtacı
ArGe Takım Lideri @ CRYPTTECH

AKIŞ

Makine Öğrenmesi Nedir ?
Günlük Hayatımızdaki Uygulamaları
Verilerin Sayısallaştırılması Özellik Belirleme

Özellik Seçim Metotları
- Bilgi Kazancı (Informaiton Gain-IG)
- Sinyalin Gürültüye Oranı: (S2N ratio)

Yeni Özelliklerin Çıkarımı
- Temel Bileşen Analizi (Principal Component Analysis)

Sınıflandırma Metotları
- Doğrusal Regresyon
- Karar Ağaçları (Decision Trees)
- En Yakın K Komşu Algoritması (k - Nearest Neighbor)
- Yapay Sinir Ağları

Kümeleme Algoritmaları:
- K-means
- Hiyerarşik Kümeleme

Çok Boyutlu Verilerle Çalışmak
Veri Sızıntısı
Pekiştirmeli Öğrenme

Öğrenme Nedir ?

“*Learning denotes changes in a system that enable a system to do the same task more efficiently the next time.”* - **Herbert Simon**

“*Learning is constructing or modifying representations of what is being experienced*." - **Ryszard Michalski**

“*Learning is making useful changes in our minds*." - **Marvin Minsky**

Makine Öğrenmesi Nedir ?

Çok büyük miktardaki verilerin elle işlenmesi,analizinin yapılması mümkün değildir.Bu tür problemlere çözüm bulmak amacıyla makine öğrenmesi metotları geliştirilmiştir.

Bu metotlar **geçmişteki verileri kullanarak** veriye en uygun modeli bulmaya çalışırlar.

Yeni gelen verileri de bu modele göre analiz edip sonuç üretirler.

Makine Öğrenmesi Ön Koşullar

Temel Matematik

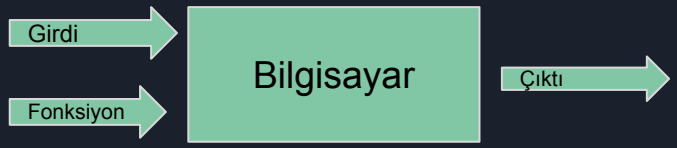
Temel İstatistik ve Olasılık Teorisi

Veri Yapıları ve Algoritmalar

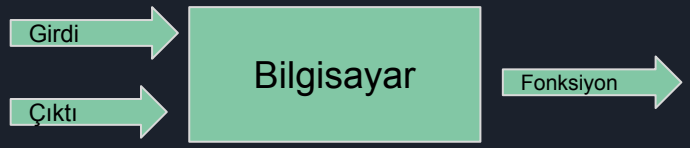
Temel Programlama Bilgisi

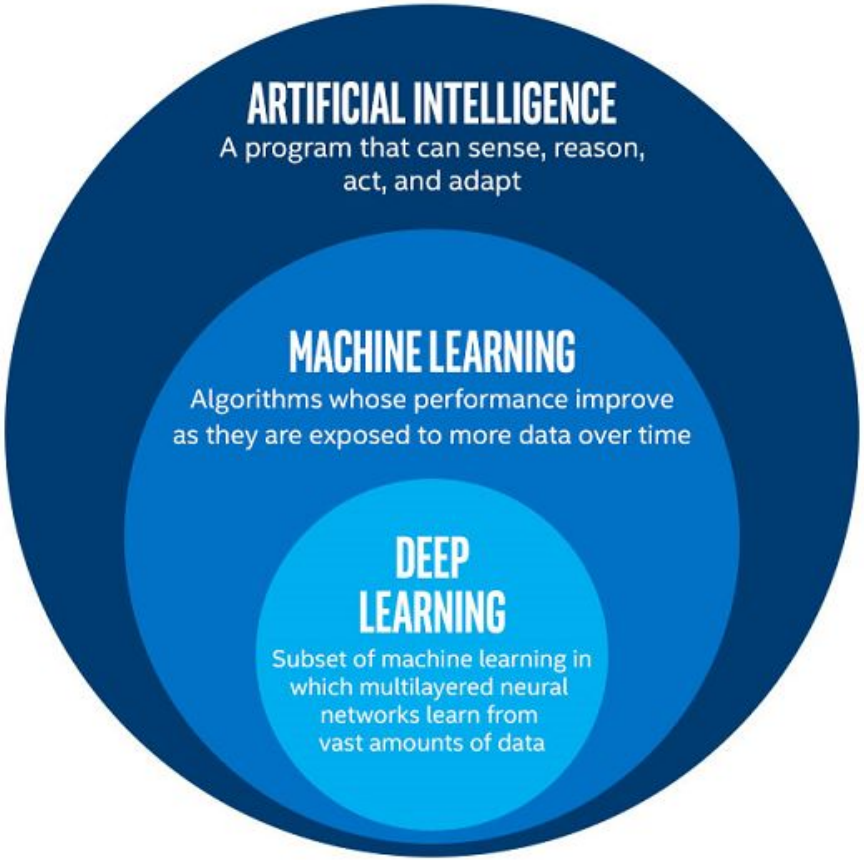
Makine Öğrenmesi Nedir ?

Klasik Programlama Yaklaşımı:



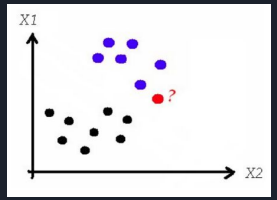
Makine Öğrenmesi Yaklaşımı:




ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

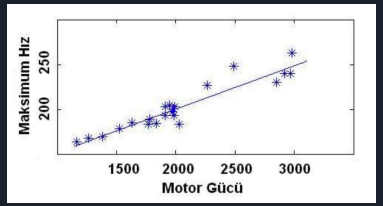
Makine Öğrenmesinden Beklentiler

1 Sınıflandırma



2 Kümeleme

3 Eğri Uydurma



4 Birliktelik Kuralları Keşfi

Makine Öğrenmesi Yöntemleri

1. Gözetimli Öğrenme (Supervised Learning)
2. Gözetimsiz Öğrenme (Unsupervised Learning)
3. Pekiştirmeli Öğrenme (Reinforcement Learning)

Makine Öğrenmesinin Gündelik Hayattaki Yeri - Kredi Talebi Değerlendirme

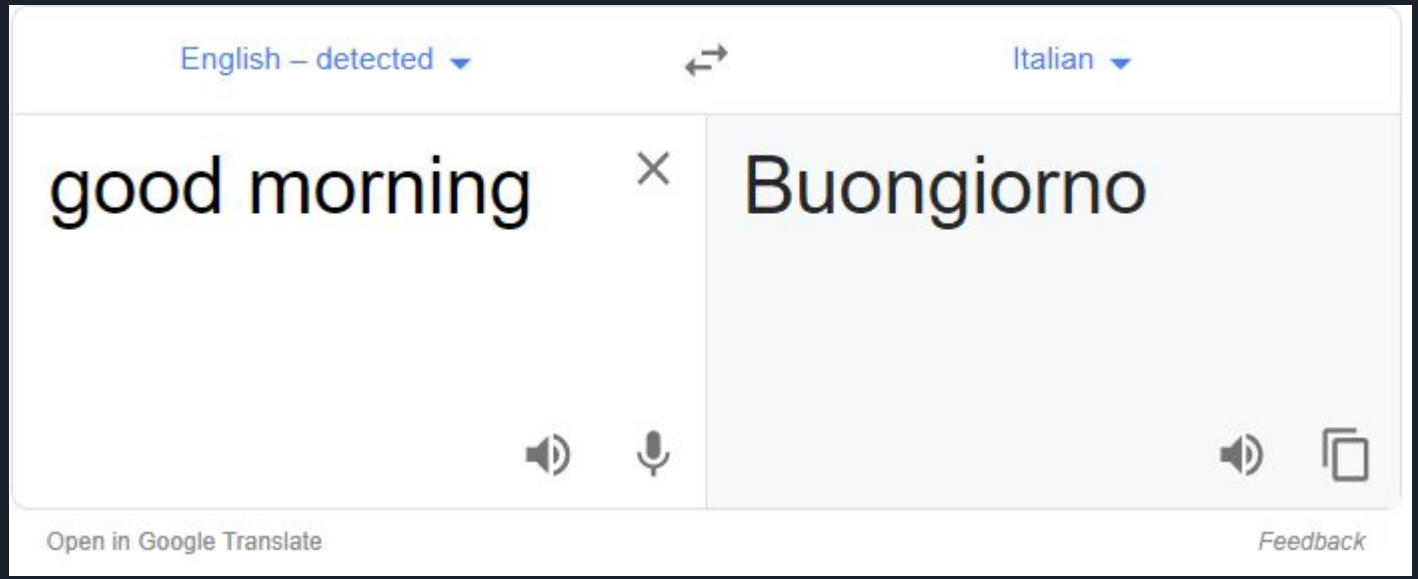
Bir X kişisi Y bankasından kredi talep eder.

Banka memuru krediyi onaylamalı mı ?

Krediyi hangi kriterlere bakarak onaylayabilir ?

Hangi kriterler kredi talebini daha çok etkiler ? (*Müşterinin göz rengi, yaşı, mesleki durumu, maaşı*...)

Makine Öğrenmesinin Gündelik Hayattaki Yeri - Doğal Dil Çevirisi



Makine Öğrenmesinin Gündelik Hayattaki Yeri - Spam Mail Tespiti

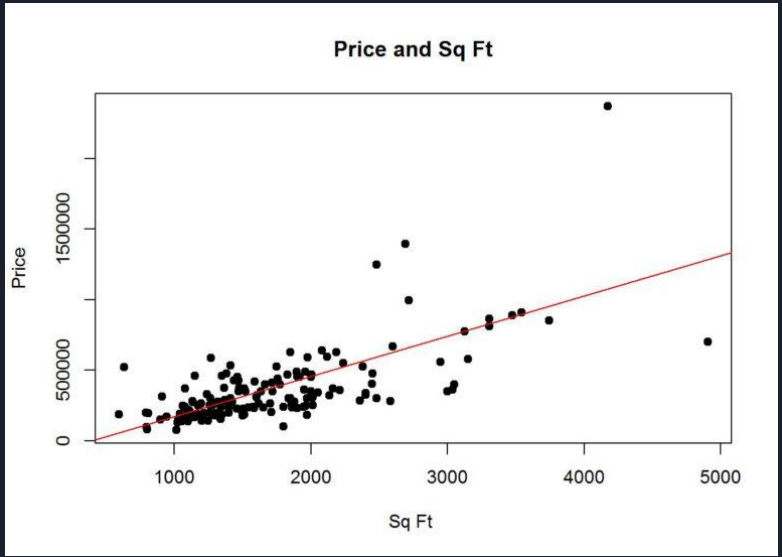
Bir A kişisi B kişisine mail gönderir.

Bu mail gereksiz spam bir mail midir ?

Mailin konusu nedir ?

Makine Öğrenmesinin
Gündelik Hayattaki Yeri - Ev fiyatlarını tahmin etmek

Bir muhitte evlerin metrekare cinsinden boyutu ve fiyat bilgisine sahipsiniz. Bilmediğiniz bir evin fiyatını tahmin edebilir misiniz ?



Makine Öğrenmesinin Gündelik Hayattaki Yeri - Market Sepeti Analizi

ABC marketinde yapılan her alışveriş bir veritabanına kaydediliyor.

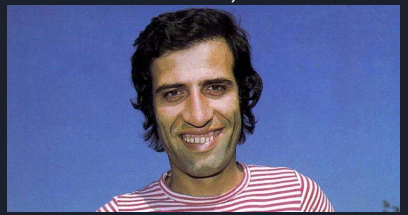
Amacımız:

- Daha fazla ürün satmak, daha fazla kazanmak ? (Nasıl yaparız ?)
- Sıklıkla birlikte satılan ürünleri bulup reyonda yakın yerlere koysak … ?

Cuma gecesi işten eve dönüyorsunuz, akşam da çok güzel bir maç var. Markete uğradınız ne alırsınız ?

Makine Öğrenmesinin Gündelik Hayattaki Yeri - Kişi Tespiti

Bu Adam;



Bu videoda geçiyor mu.

Makine Öğrenmesinin Gündelik Hayattaki Yeri - Saldırı Tespit Sistemi

HTTP Paket içeriklerinde saldırı senaryolarına dair desenler var mı ?

Web sunucunun döndürdüğü 5xx ve 4xx hatalarında anormal bir artış var mı ?

Web sunucu durduk yere kapanmış veya reboot olmuş mu ?

Makine Öğrenmesinin Gündelik Hayattaki Yeri - Zararlı Yazılım Analizi

Zararlı yazılıma dair ipuçları neler olabilir ?

Statik analiz ve dinamik analizde çıkarılabilecek özellikler neler olabilir ?

4. Gün Android Zararlı Yazılım Tespitine ilişkin WORKSHOP olacak :)

Verilerin Sayısallaştırılması

Resim	Resmin her bir pikselinin 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	Metindeki 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	Ses, genlik ve frekansın zaman içinde değişimiyle kodlanır.

SORU - CEVAP

AKIŞ

Makine Öğrenmesi Nedir ?
Günlük Hayatımızdaki Uygulamaları
Verilerin Sayısallaştırılması Özellik Belirleme
Özellik Seçim Metotları
- **Bilgi Kazancı (Informaiton Gain-IG)**
- **Sinyalin Gürültüye Oranı: (S2N ratio)**

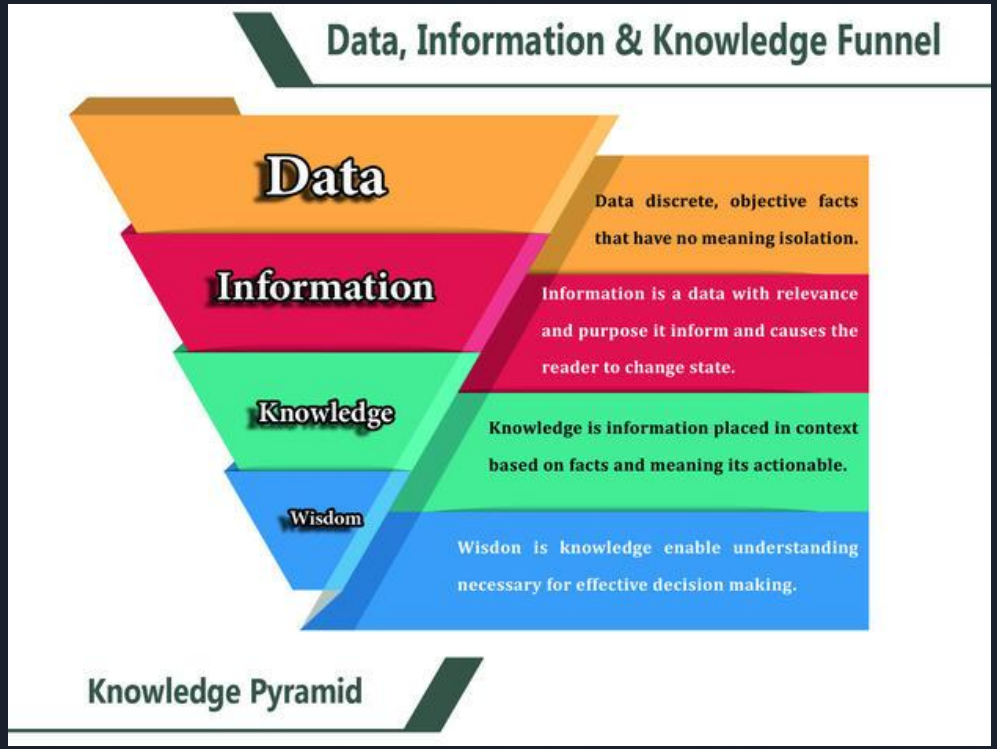
Yeni Özelliklerin Çıkarımı
- **Temel Bileşen Analizi (Principal Component Analysis)**

Sınıflandırma Metotları
- Doğrusal Regresyon
- Karar Ağaçları (Decision Trees)
- En Yakın K Komşu Algoritması (k - Nearest Neighbor)
- Yapay Sinir Ağları

Kümeleme Algoritmaları:
- K-means
- Hiyerarşik Kümeleme

Çok Boyutlu Verilerle Çalışmak
Veri Sızıntısı
Pekiştirmeli Öğrenme

Kavramlar

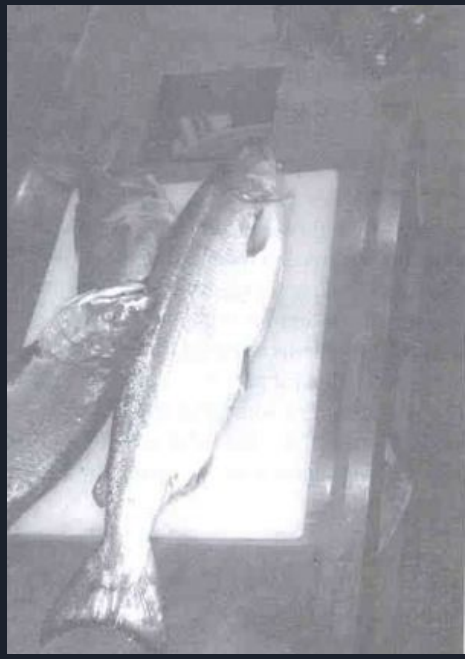

Data, Information & Knowledge Funnel
Data
Data discrete, objective facts that have no meaning isolation.
Information
Information is a data with relevance and purpose it inform and causes the reader to change state.
Knowledge
Knowledge is information placed in context based on facts and meaning its actionable.
Wisdom
Wisdon is knowledge enable understanding necessary for effective decision making.
Knowledge Pyramid

Özellik Seçimi

Görev: Kayan bir bant üzerindeki balıkların levrek mi, somon mu olduğunu ayırt etmek.

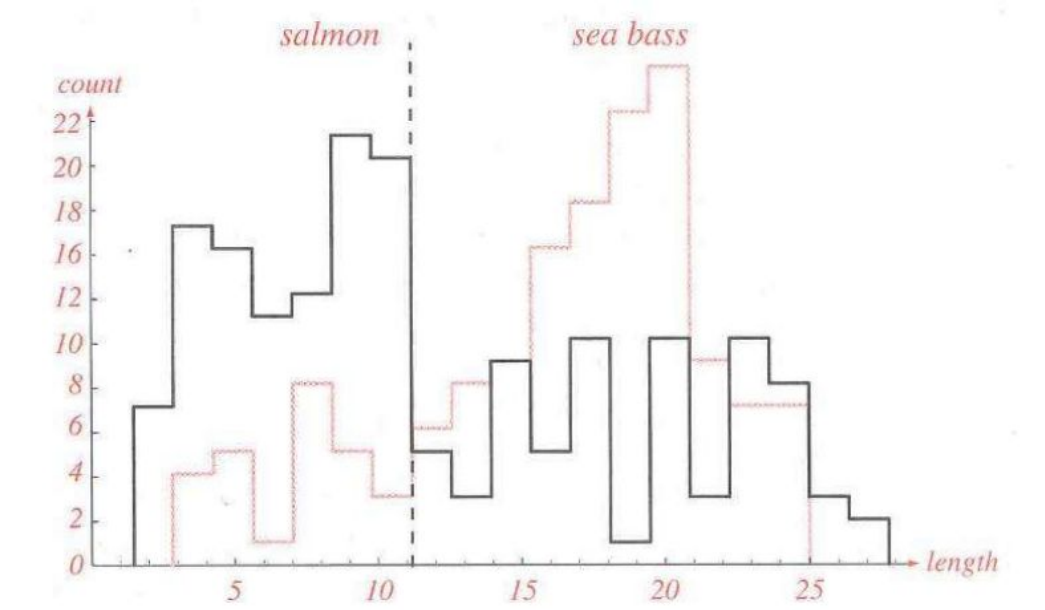
Elimizdeki veri: Kayan bant üzerindeki balıkların fotoğrafı.

Hangi özelliklerine bakarak ayırt edebiliriz ? (Uzunluk, parlaklık…….)



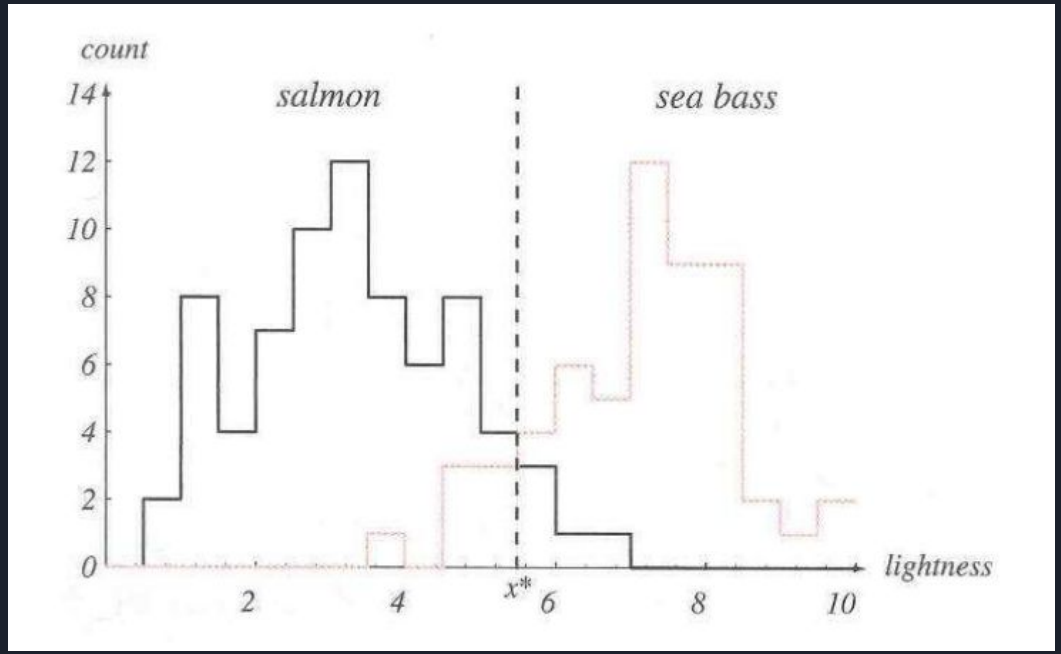
Özellik Seçimi - Uzunluk

Salmon'lar genelde SeaBass'lardan daha kısalar.



Özellik Seçimi - Parlaklık

SeaBass genelde Salmon'lardan daha parlaklar.



Özellik Seçimi

Kredi talebi ve balık örneğindeki gibi, özellik seçmek makine öğrenmesi algoritmalarının önemli bir parçasıdır.

Özellikler seçilirken iki farklı yaklaşım kullanılabilir;

1. Tüm özellik kümesinden alt kümeler oluşturup seçmek. (Feature Selection)
2. Var olan özelliklerin lineer birleşiminden yeni özellikler türetmek. (Feature Composition)

Bilgi Kazancı ile Özellik Seçimi

Olay No	Hava	Nem	Rüzgar	Su sıcaklığı	Pikniğe gidildi mi?
1	güneşli	normal	güçlü	ılık	Evet
2	güneşli	yüksek	güçlü	ılık	Evet
3	yağmurlu	yüksek	güçlü	ılık	Hayır
4	güneşli	yüksek	güçlü	soğuk	Evet

Bilgi Kazancı ile Özellik Seçimi

- Pikniğe gidildi mi? sorusunun iki cevabı vardır.
- Evet cevabının olasılığı ¾
- Hayır cevabının olasılığı ¼
- Dolayısıyla Pikniğin Entropi'si
- **E(Piknik)** = -(3/4) log2(3/4) - (1/4) log2(1/4) = **0.811 olarak** bulunur.

Bilgi Kazancı ile Özellik Seçimi

- **Gain(Piknik,Hava)**= 0.811- (3/4) (-(3/3) log2 (3/3) – 0) – (1/4) (0 – (1/1) log2 (1/1)) = **0.811**
- Hava özelliğinin IG'si hesaplanırken bulunan rakamların açıklamaları:

0.811 → Pikniğe gitme olayının Entropisi

(3/4) → havanın güneşli olma oranı

(3/3) → hava güneşli iken pikniğe gidilme oranı

0 → hava güneşli iken pikniğe gidilmeme oranı

(1/4) → havanın yağmurlu olma oranı

0 → hava yağmurlu iken pikniğe gidilme oranı

(1/1) → hava yağmurlu iken pikniğe gidilmeme oranı

Bilgi Kazancı ile Özellik Seçimi

- **Gain(Piknik,Nem)**= 0.811- (1/4) (-(1/1) log2 (1/1) - 0) – (3/4) (-(2/3) log2(2/3)-(1/3) log2(1/3))
  = 0.811 -0.688= **0.1225**
- **Gain(Piknik,Rüzgar)**= 0.811- (4/4) (-(3/4) log2(3/4) – (1/4) log2(1/4))
  = 0.811 -0.811= **0**
- **Gain(Piknik,SuSıcaklığı)**= 0.811- (3/4) (-(2/3) log2(2/3) – (1/3) log2(1/3)) – (1/4) (-(1/1) log2 (1/1))
  = 0.811 -0.688= **0.1225**
- En büyük bilgi kazancına sahip özellik ‘Hava’dır.
- Gerçek uygulamalarda ise yüzlerce özelliğin bilgi kazançları hesaplanır ve en büyük olanları seçilerek kullanılır.

Sinyal / Gürültü Oranıyla Özellik Seçimi

Sınıflar arası varyansları fazla; Sınıf içi varyansları az olan özellikler seçilir.

$$S_i = \frac{m_1 - m_2}{d_1 - d_2}$$

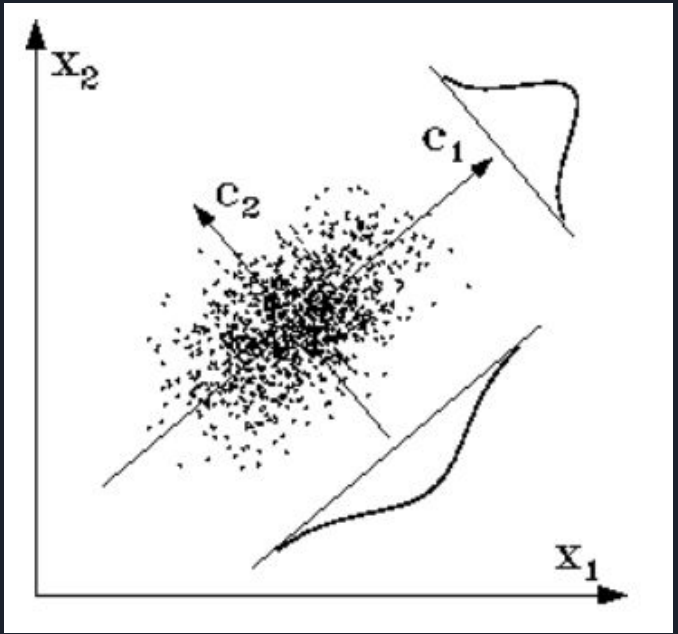
m1→ sınıf1'deki i. özelliklerin ortalaması
m2→ sınıf2'deki i. özelliklerin ortalaması
d1→ sınıf1'deki i. özelliklerin standart sapması
d2→ sınıf2'deki i. özelliklerin standart sapması
S değeri en yüksek olan özellikler
seçilerek sınıflandırmada kullanılırlar.

Principal Component Analysis

Bu metotta örneklerin en fazla değişim gösterdikleri boyutlar bulunur.

Yanda veriler c1 ve c2 eksenlerine izdüşümü yapıldığındaki dağılımları gösterilmiştir.

C1 eksenindeki değişim daha büyüktür. Dolayısıyla veriler 2 boyuttan bir boyuta C1 ekseni üzerine iz düşürülerek indirgenmiş olur.



SORU - CEVAP

AKIŞ

Makine Öğrenmesi Nedir ?
Günlük Hayatımızdaki Uygulamaları
Verilerin Sayısallaştırılması Özellik Belirleme
Özellik Seçim Metotları
- Bilgi Kazancı (Informaiton Gain-IG)
- Sinyalin Gürültüye Oranı: (S2N ratio)

Yeni Özelliklerin Çıkarımı
- Temel Bileşen Analizi (Principal Component Analysis)

Sınıflandırma Metotları
- **Doğrusal Regresyon**
- **Karar Ağaçları (Decision Trees)**
- **En Yakın K Komşu Algoritması (k - Nearest Neighbor)**
- **Yapay Sinir Ağları**

Kümeleme Algoritmaları:
- **K-means**
- **Hiyerarşik Kümeleme**

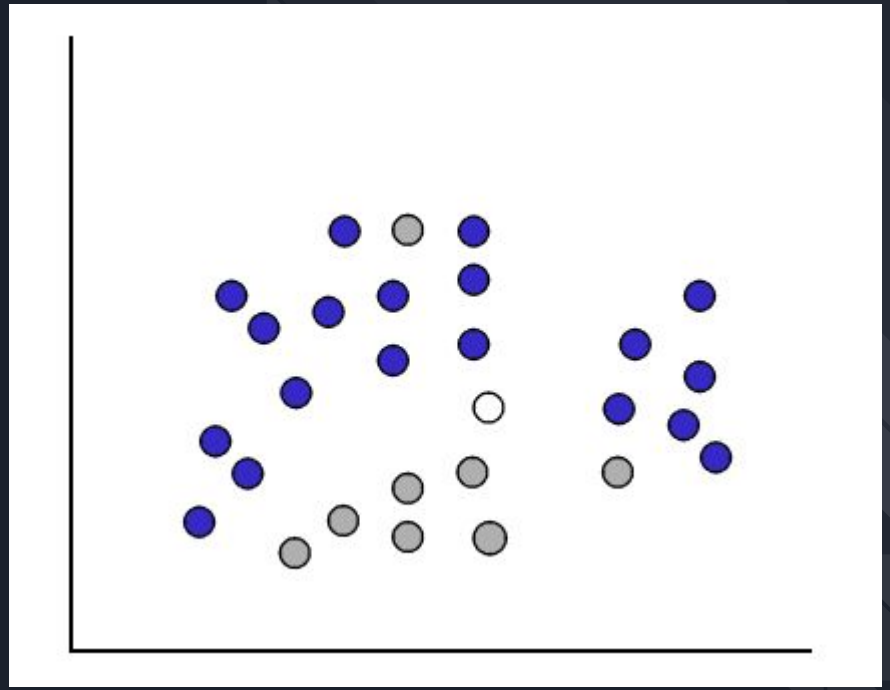
Çok Boyutlu Verilerle Çalışmak
Veri Sızıntısı
Pekiştirmeli Öğrenme

Sınıflandırma Yöntemleri

Beyaz Örnek hangi gruba dahil edilmeli ?

Mavi mi Gri mi ?

Siz nasıl ayırırdınız ?



Lineer Regresyon ile Sınıflandırma

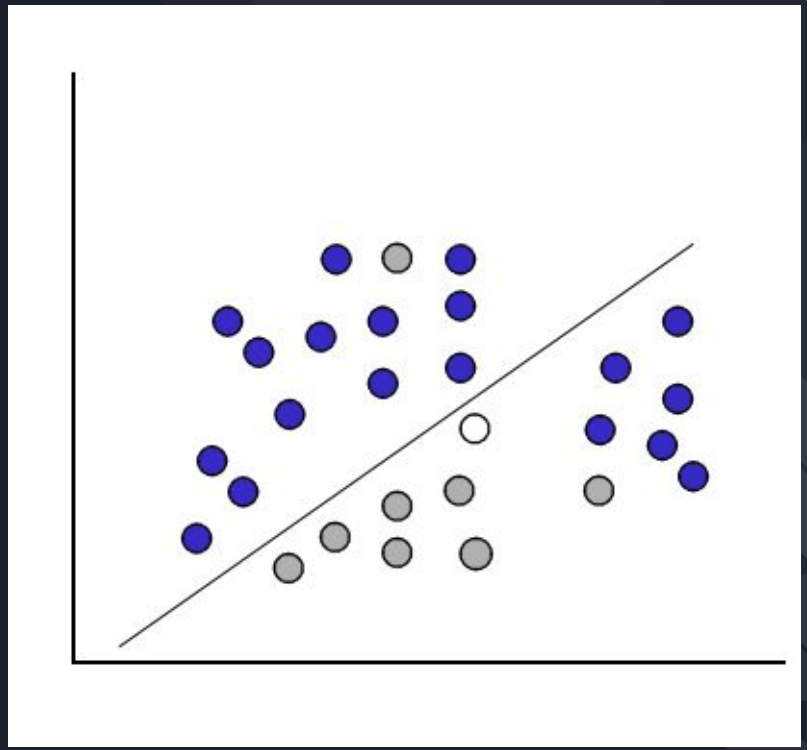
Regresyon: Eğri uydurma anlamına gelir.

Varsayalım ki; çizdiğimiz eğrinin(doğrunun) denklemi ;

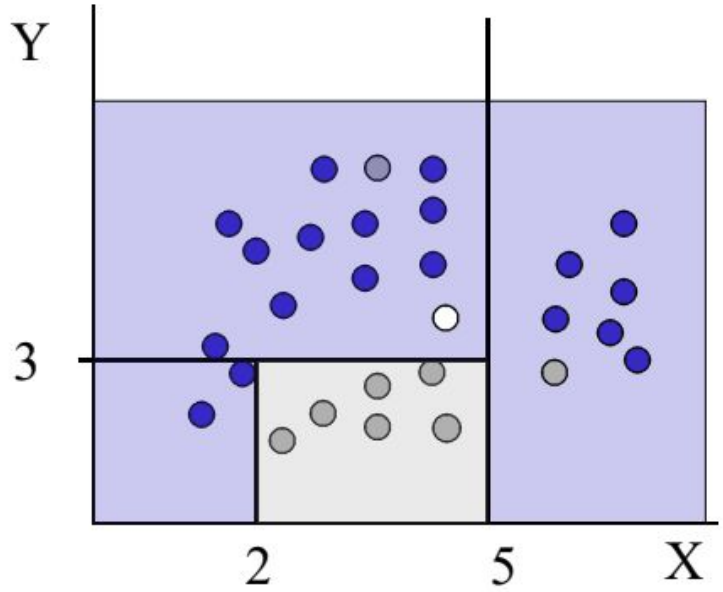
w0 + w1x + w2y >= 0 olsun.

Lineer regresyonun görevi w0, w1 ve w2'leri **en küçük kareler** yöntemiyle bulmaktır.

Eğri, uzayı iki parçaya böler beyaz örnek grilerin olduğu kısımda olduğu için GRİ gruba ait olduğunu söyleyebiliriz.



Karar Ağaçlarıyla Sınıflandırma



Böl ve yönet stratejisi

Nasıl böleceğiz?

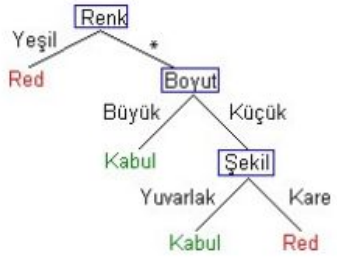
if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

Karar Ağaçlarında Karar Düğümlerinin Bulunması

Karar düğümlerinde yer alan özelliğin ve eşik değerinin belirlenmesinde genel olarak **entropi** kavramı kullanılır.

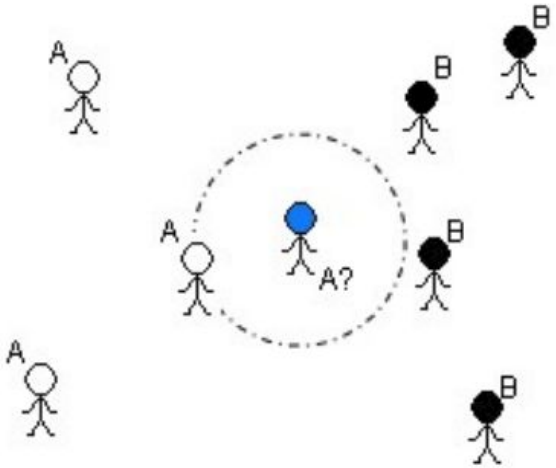
Eğitim verisi her bir özelliğin her bir değeri için ikiye bölünür. Oluşan iki alt kümenin entropileri toplanır. En düşük entropi toplamına sahip olan özellik , değer ikilisi karar düğümüne yerleştirilir.

Şekil	Renk	Boyut	Sınıf
Yuvarlak	Yeşil	Küçük	Red
Kare	Siyah	Büyük	Kabul
Kare	Sarı	Büyük	Kabul
Yuvarlak	Sarı	Küçük	Red
Kare	Yeşil	Büyük	Red
Kare	Sarı	Küçük	Kabul

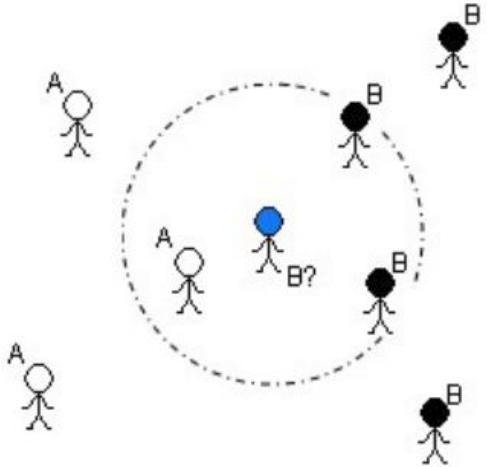


K - En Yakın Komşu

- Bana Arkadaşını söyle, sana kim olduğunu söyleyeyim.



En yakın 1 komşu



En yakın 3 komşu

SORU - CEVAP

Kümeleme Algoritmaları

Kümeleme algoritmaları eğiticisiz (unsupervised) öğrenme metodlarıdır.

Örneklere ait sınıf bilgisini kullanmazlar.

Temelde verileri en iyi temsil edecek vektörleri bulmaya çalışırlar.

Verileri temsil eden vektörler bulunduktan sonra artık tüm veriler bu yeni vektörlerle kodlanabilirler ve farklı bilgi sayısı azalır.

Bu nedenle birçok sıkıştırma algoritmasının temelinde kümeleme algoritmaları yer almaktadır.

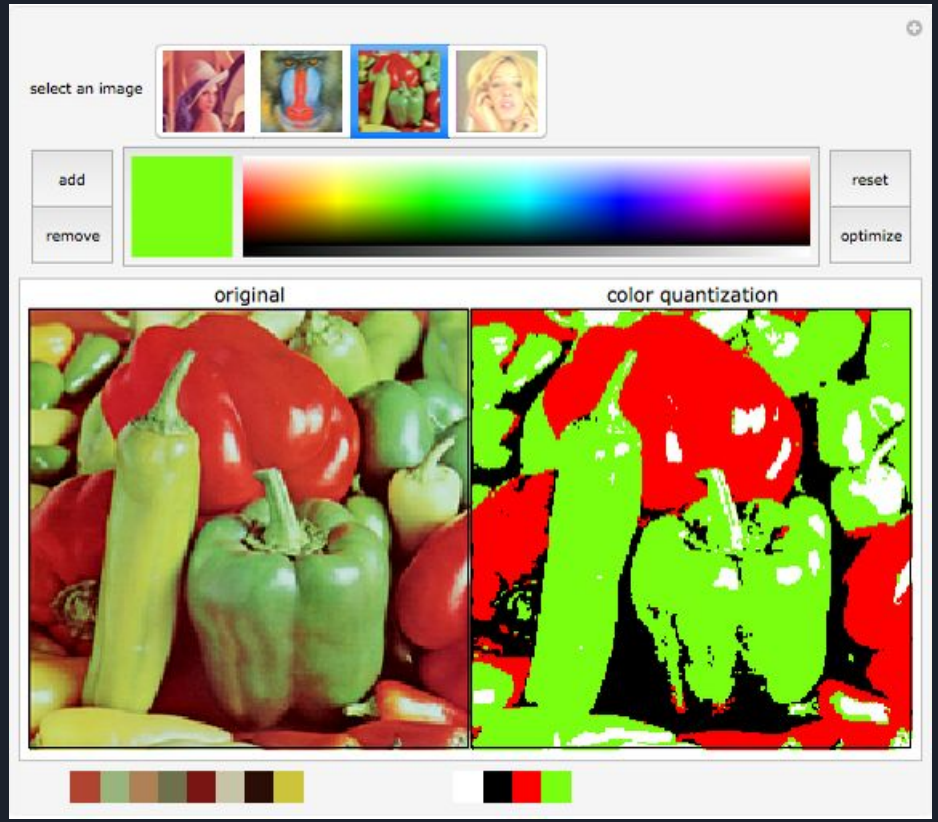
Kümeleme Algoritmaları

- Bir boyutlu (özellikli) 10 örnek içeren bir veri
  12-15-13-87-4-5-9-67-1-2
- Bu 10 farklı veriyi 3 farklı veriyle temsil etmek istersek:
  12-12-12-77-3-3-3-77-3-3
- şeklinde ifade edebiliriz.
- Kümeleme algoritmaları bu 3 farklı verinin değerlerini bulmakta kullanılırlar.
- Gerçek değerlerle temsil edilen değerler arasındaki farkları minimum yapmaya çalışırlar.

Yukarıdaki örnek için 3 küme oluşmuştur.

- 12-15-13 örnekleri 1. kümede
- 87-67 örnekleri 2. kümede
- 4-5-1-2-9 örnekleri 3. kümede yer almaktadır.

Renk Kümeleme

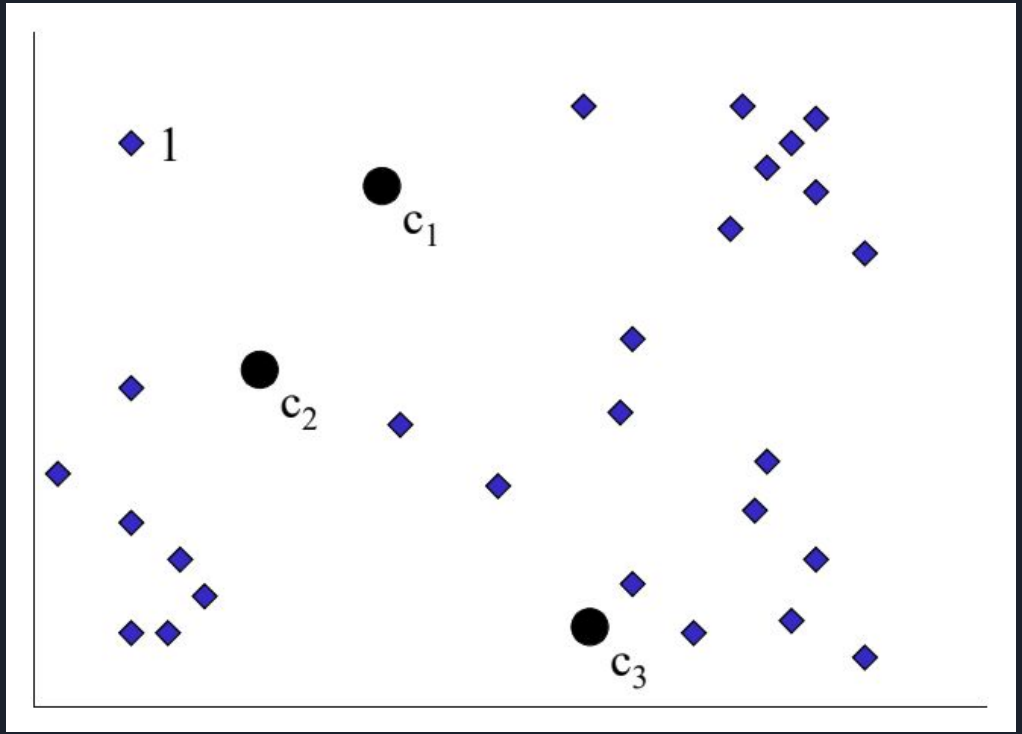


K-Means Kümeleme Algoritması

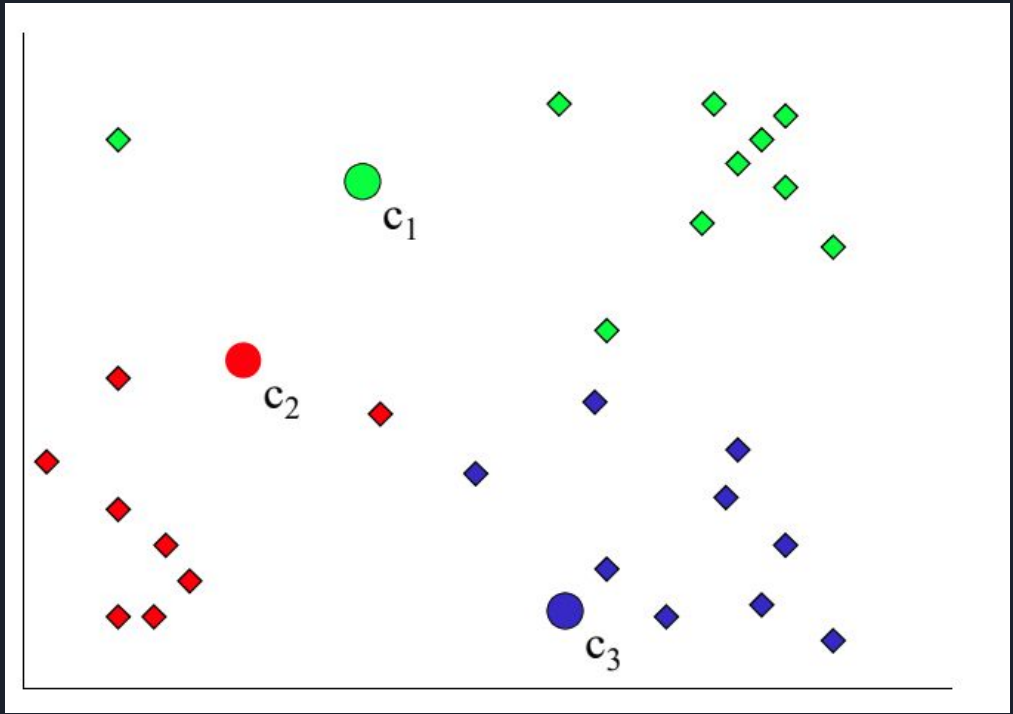
Sadece sayısal veriyle çalışır

1) Rasgele K adet küme merkezi ata
2) Her örneği en yakınındaki merkezin kümesine ata
3) Merkezleri kendi kümelerinin merkezine ata
4) 2. ve 3. adımları küme değiştiren örnek kalmayıncaya kadar tekrar et.

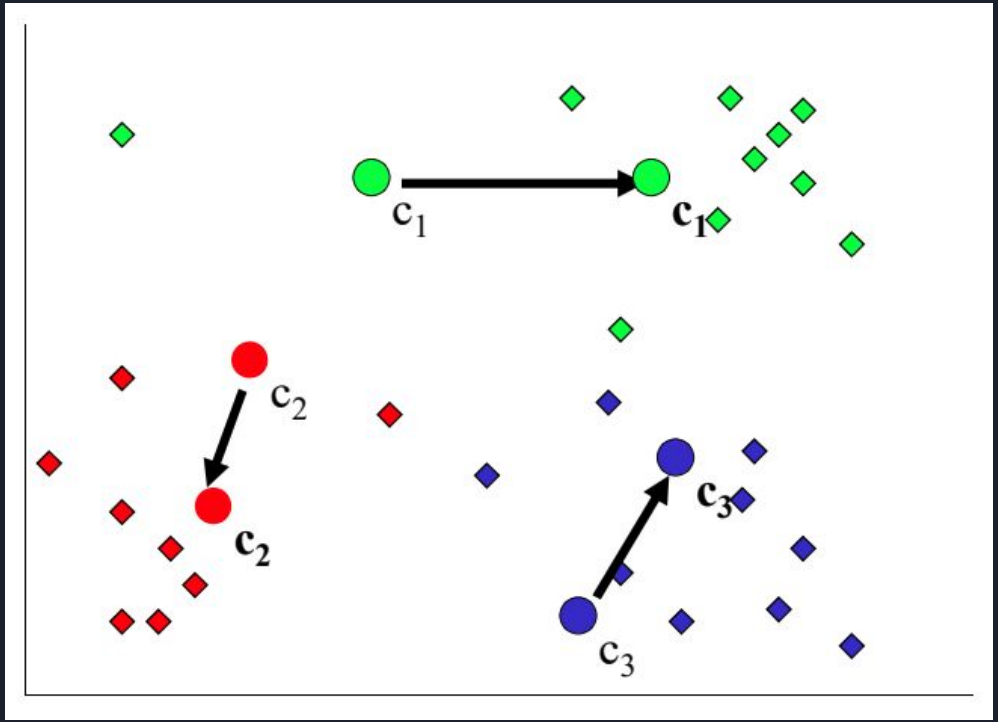
K-Means Adımları



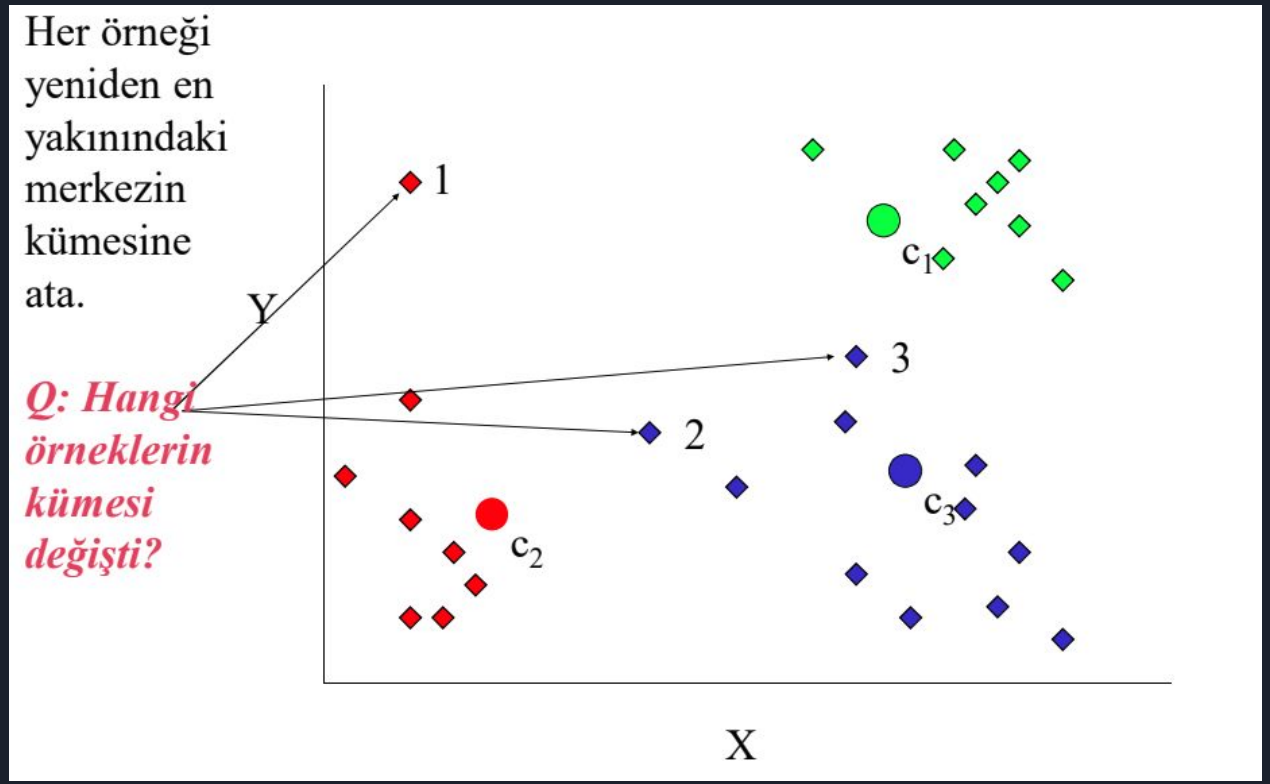
K-Means Adımları



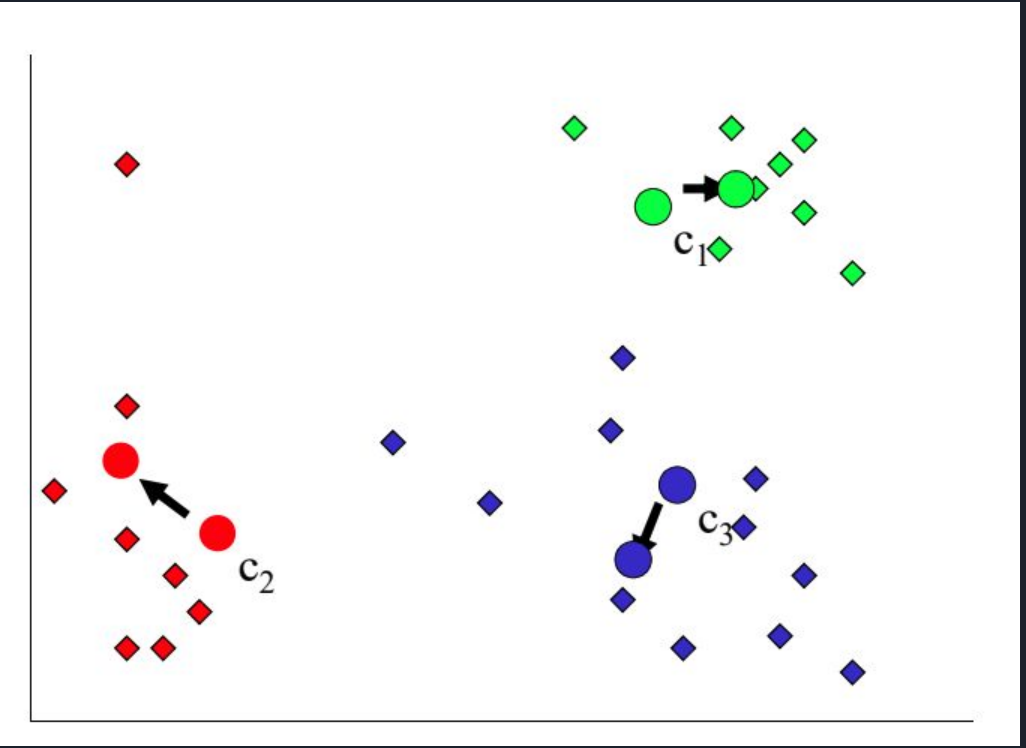
K-Means Adımları



K-Means Adımları



K-Means Adımları



Hiyerarşik Kümeleme

SORU - CEVAP

AKIŞ

Makine Öğrenmesi Nedir ?
Günlük Hayatımızdaki Uygulamaları
Verilerin Sayısallaştırılması Özellik Belirleme
Özellik Seçim Metotları
- Bilgi Kazancı (Informaiton Gain-IG)
- Sinyalin Gürültüye Oranı: (S2N ratio)

Yeni Özelliklerin Çıkarımı
- Temel Bileşen Analizi (Principal Component Analysis)

Sınıflandırma Metotları
- Doğrusal Regresyon
- Karar Ağaçları (Decision Trees)
- En Yakın K Komşu Algoritması (k - Nearest Neighbor)
- Yapay Sinir Ağları

Kümeleme Algoritmaları:
- K-means
- Hiyerarşik Kümeleme

Çok Boyutlu Verilerle Çalışmak
Veri Sızıntısı
Pekiştirmeli Öğrenme

Çok Boyutlu Verilerle Çalışmak

- Tek boyutlu uzayda [0,1] aralığı temsil eden 10 nokta
- Rastgele bir noktanın, uzayı temsil eden noktalardan en yakın olanına ortalama uzaklığı = 0.5
- İki boyutlu uzayda rasgele bir noktanın en yakın noktaya olan ortalama uzaklığının düşey ya da dikey (manhattan) 0.5 olması için gerekli temsilci nokta sayısı = 100

Boyut Sayısı	Gerekli temsil eden nokta sayısı
1	10
2	100
3	1000
…	…
p	10^p

Doğru sınıflandırma yapmak için gereken örnek sayısı artıyor.

Veri Sızıntısı

- Tahmin sonuçlarının çok iyi görünmesine sebep olabilir, ama gerçekler uygulamada ortaya çıkar
- Test kümesindeki bilgilerin eğitim sürecine karışması
  - Verileri normalize ederken, özellik seçimi yaparken test kümesini de kullanmak
- Test zamanı elde olmayacak özelliklerin kullanılması
  - $x(t)=f(x(t-1),x(t+1))$
- Zaman serisi sınıflandırmada eğitim ve test kümelerini oluşturmada hata
  - Rasgele seçim yapılmamalı, bir t anından öncesi eğitim, sonrası test olmalı ki ardışık süreçleri değil sınıfı tanısın.
- Verilerde çıkışla korelasyonu çok yüksek olan özelliklerin olması (kişi tanırken id, telefon no vb.)