

Makine Öğrenmesi-2

M.Fatih AMASYALI
Yapay Zeka Ders Notları

Akış

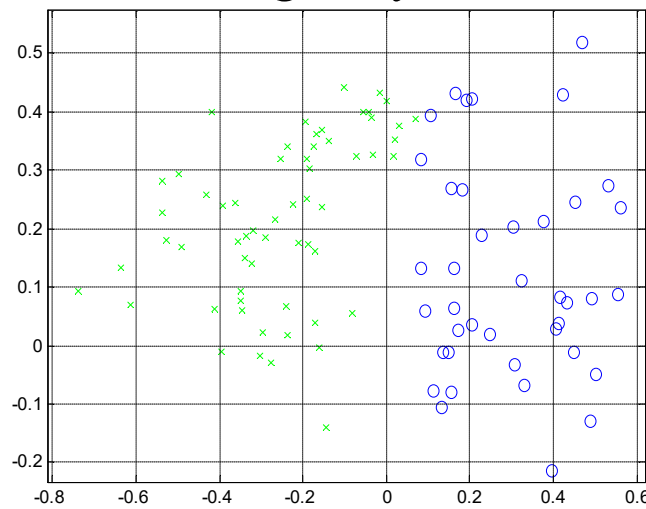
- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- Verilerin Sayısallaştırılması
- Özellik Belirleme
 - Özellik Seçim Metotları
 - Bilgi Kazancı (Information Gain-IG)
 - Sinyalin Gürültüye Oranı: (S2N ratio)
 - Alt küme seçiciler (Wrappers)
 - Yeni Özelliklerin Çıkarımı
 - Temel Bileşen Analizi (Principal Component Analysis)
 - Doğrusal Ayırtıcı Analizi (Linear Discriminant Analysis)
- Sınıflandırma Metotları
 - Doğrusal Regresyon
 - Karar Ağaçları (Decision Trees)
 - Yapay Sinir Ağları
 - En Yakın K Komşu Algoritması (k - Nearest Neighbor)
 - Öğrenmeli Vektör Kuantalama (Learning Vector Quantization)
- Kümeleme Algoritmaları:
 - Hiyerarşik Kümeleme
 - K-means
 - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)
 - DBscan
- Regresyon Algoritmaları
- Çok Boyutlu Verilerle Çalışmak
- Veri Sızıntısı
- Pekiştirmeli Öğrenme

Özellik Belirleme

- Bir doktor
- Veri: Kişi bilgilerini içeren dosyalar
- Görev: Kimler hasta bul.
- Hangi bilgilere bakılır?
 - Ad soyad
 - Doğum yeri
 - Cinsiyet
 - Kan tahlili sonuçları
 - Röntgen sonuçları
 - vs.

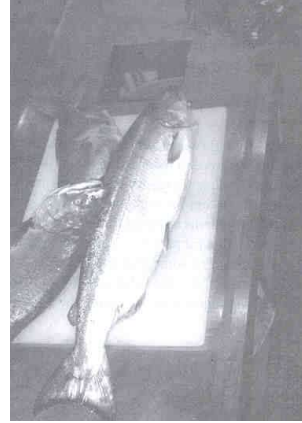
1. Özellik	2. Özellik	Sınıf
1	3	A
2	3	B
1	4	A
2	3	B

Hangi boyut?

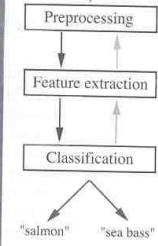


Balık Hali

- Kayan bant üzerindeki balığın türünü belirlemek(Salmon? Sea Bass?)



kamera

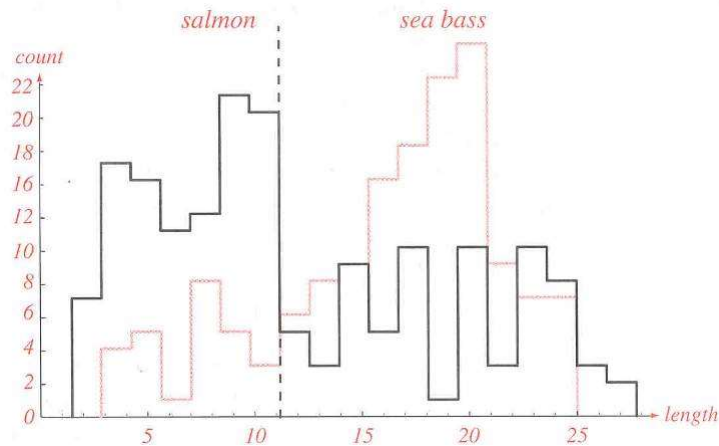


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Balık Özellikleri. Uzunluk.

- Salmon lar genelde Sea Bass lardan daha kısalar.

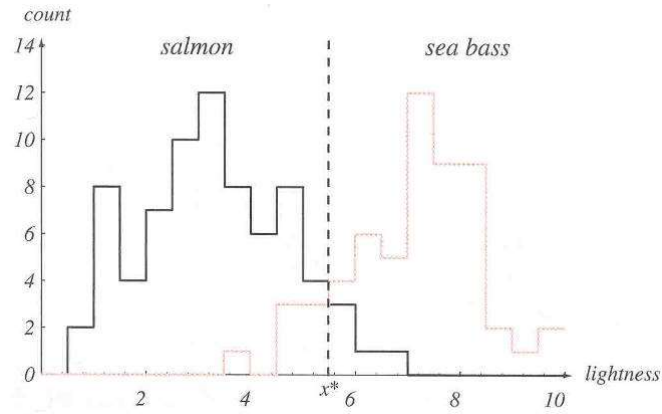


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Balık Özellikleri. Parlaklık.

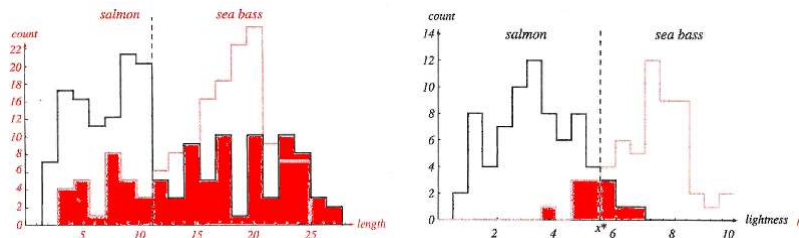
- Sea Bass genelde Salmon lardan daha parlaklar.



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Hangi Özellik?



Kırmızı bölgeler yapılan hataları gösteriyor.

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Doktoru yoralım ☺

- Hastalık dosyasında 5000 adet özellik olsaydı? Örneğin kişinin DNA dizisine bakarak hasta olup olmadığına karar verecek olsaydık ne yapardık? Nerelere bakacağımıza nasıl karar verirdik.
- Burada devreye makineleri sokmamız gerekiyor gibi gözükmemekte.
- Bu olay bir insanın hesap yapma kabiliyetiyle, bir hesap makinesininkini karşılaştırmaya benziyor.

Özellik seçimi

- Bu problem makinelerle iki farklı metotla çözülebilir.
 - Var olan özelliklerden bazılarını seçmek
 - Özellikleri tek tek değerlendirmek (Filter)
 - Özellik alt kümeleri oluşturup, sınıflandırıcılar kullanıp performanslarını ölçüp, bu alt kümeleri en iyilemek için değiştirerek (Wrapper)
 - Var olan özelliklerin lineer birleşimlerinden yeni özelliklerin çıkarımı

Özellikleri birer birer inceleme (Filters)

- Eğitim bilgilerindeki her bir özellik teker teker ele alınır.
- Örnek ile ilgili sadece o özellik elimizde olsaydı ne olurdu sorusunun cevabı bulunmaya çalışılır.
- Seçilen özellikle sınıf ya da sonucun birlikte değişimleri incelenir.
- Özellik değiştiğinde sınıf ya da sonuç ne kadar değişiyorsa o özelliğin sonuca o kadar etkisi vardır denilir.

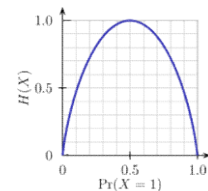
Bilgi Kazancı - Information Gain

S eğitim seti içindeki A özelliğinin

$$Gain(S, A) \equiv Entrophy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entrophy(S_v)$$

N kavramının c farklı değeri varsa N'in entropisi, N'in aldığı her değer olasılıkları kullanılarak

$$Entropy(N) = \sum_{i=1}^c -p_i \log_2 p_i$$



daha önceki hava, nem, rüzgar, su sıcaklığı gibi değerlere göre pikniğe gidip gitmeme kararı verilmiş 4 olay

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

Her bir özelliğin piknik kavramı için bilgi kazancını bulalım

- Pikniğe gidildi mi? sorusunun iki cevabı vardır.
- Evet cevabının olasılığı $\frac{3}{4}$
- Hayır cevabının olasılığı $\frac{1}{4}$
- Dolayısıyla Pikniğin Entropi'si
- $E(\text{Piknik}) = -(\frac{3}{4}) \log_2(\frac{3}{4}) - (\frac{1}{4}) \log_2(\frac{1}{4})$
= **0.811** olarak bulunur.

- **Gain(Piknik,Hava)**= $0.811 - (3/4) (- (3/3) \log_2 (3/3) - 0) - (1/4) (0 - (1/1) \log_2 (1/1)) = 0.811$
- Hava özelliğinin IG'si hesaplanırken bulunan rakamların açıklamaları:
 $0.811 \rightarrow$ Pikniğe gitme olayının Entropisi
 $(3/4) \rightarrow$ havanın güneşli olma oranı
 $(3/3) \rightarrow$ hava güneşli iken pikniğe gidilme oranı
 $0 \rightarrow$ hava güneşli iken pikniğe gidilmeme oranı
 $(1/4) \rightarrow$ havanın yağmurlu olma oranı
 $0 \rightarrow$ hava yağmurlu iken pikniğe gidilme oranı
 $(1/1) \rightarrow$ hava yağmurlu iken pikniğe gidilmeme oranı

- **Gain(Piknik,Nem)**= $0.811 - (1/4) (- (1/1) \log_2 (1/1) - 0) - (3/4) (- (2/3) \log_2 (2/3) - (1/3) \log_2 (1/3))$
 $= 0.811 - 0.688 = 0.1225$
- **Gain(Piknik,Rüzgar)**= $0.811 - (4/4) (- (3/4) \log_2 (3/4) - (1/4) \log_2 (1/4))$
 $= 0.811 - 0.811 = 0$
- **Gain(Piknik,Su Sıcaklığı)**= $0.811 - (3/4) (- (2/3) \log_2 (2/3) - (1/3) \log_2 (1/3)) - (1/4) (- (1/1) \log_2 (1/1))$
 $= 0.811 - 0.688 = 0.1225$
- En büyük bilgi kazancına sahip özellik 'Hava'dır.
- Gerçek uygulamalarda ise yüzlerce özelliğin bilgi kazançları hesaplanır ve en büyük olanları seçilerek kullanılır.

S2N

- Sınıflar arası ayrılıkların fazla sınıf içi ayrılıkların az olan özellikler seçilir.

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

$m_1 \rightarrow$ sınıf1'deki i. özelliklerin ortalaması

$m_2 \rightarrow$ sınıf2'deki i. özelliklerin ortalaması

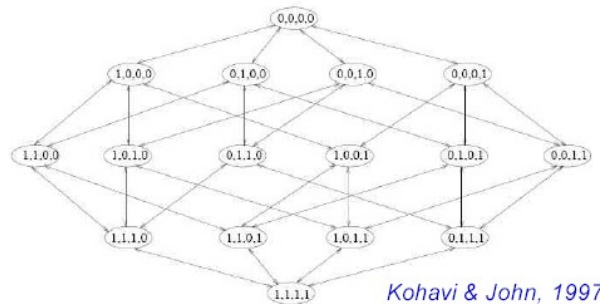
$d_1 \rightarrow$ sınıf1'deki i. özelliklerin standart sapması

$d_2 \rightarrow$ sınıf2'deki i. özelliklerin standart sapması

S değeri en yüksek olan özellikler seçilerek sınıflandırmada kullanılırlar.

Özellik altkümesi seçiciler (Wrappers)

N özellik için olası 2^N özellik alt kümesi = 2^N eğitim

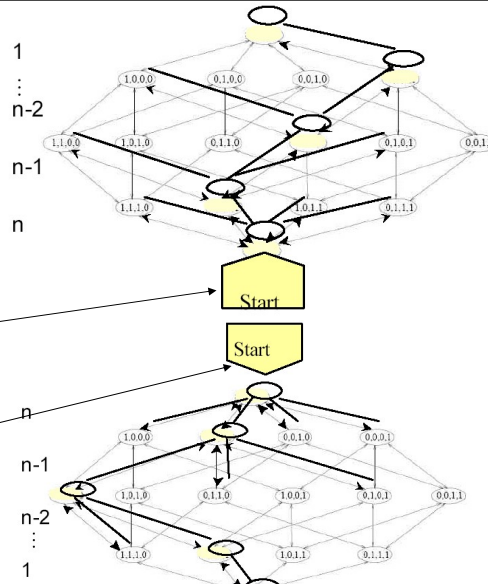


Kohavi & John, 1997

Özellik altkümesi seçiciler

- *Hızlandırmak için tüm olasılıkları denemek yerine*

- Hepsiyi başlayıp her seferinde bir tane elemek
- Tek özelle başlayıp her seferinde bir tane eklemek



Hangi yoldan gidileceğine o özellik kümesinin sınıflandırmadaki performansına bakılarak karar verilir.

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Yeni Özelliklerin Çıkarımı

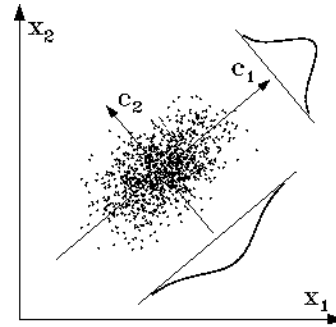
- Var olan özelliklerin lineer birleşimlerinden yeni bir özellik uzayı oluşturulur ve veriler bu uzayda ifade edilirler. Yaygın olarak kullanılan 2 metot vardır.
- PCA
- LDA

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Temel Bileşen Analizi-TBA (Principle Component Analysis - PCA)

- Bu metotta örneklerin en fazla değişim gösterdikleri boyutlar bulunur. Yansa veriler c_1 ve c_2 eksenlerine izdüşümü yapıldığındaki dağılımları gösterilmiştir.
- C_1 eksenindeki değişim daha büyüktür. Dolayısıyla veriler 2 boyuttan bir boyuta C_1 eksenine iz düşürülerek indirgenmiş olur.

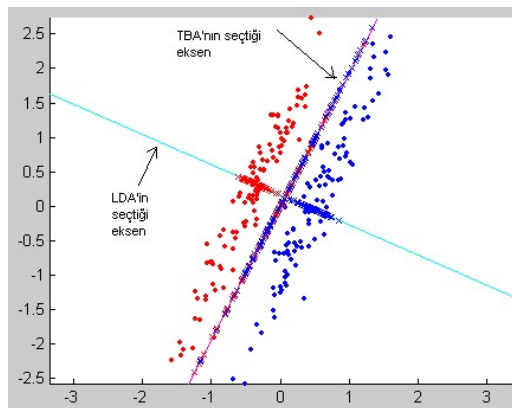


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Doğrusal Ayırte den Analizi (Linear Discriminant Analysis - LDA)

Yandaki gibi durumlar için LDA önerilmiştir. LDA varyanslara ek olarak sınıf bilgisini de kullanarak boyut indirgene yapar. Sadece varyansa değil sınıflandırabilmeye de bakar.



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

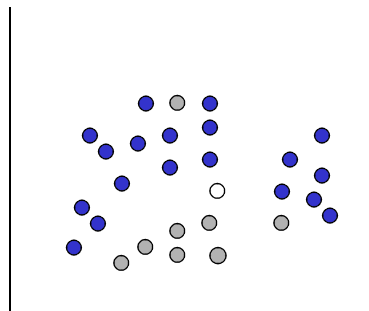
YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Hangisi

- Niye bu kadar çok metot var?
- Ne zaman hangisini kullanacağız?

Sınıflandırma Metotları

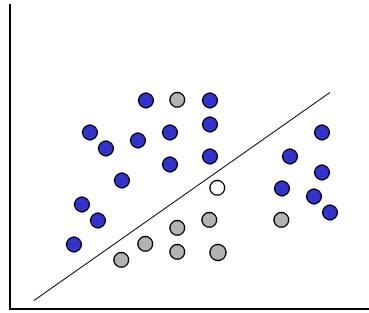
Görev: Önceden etiketlenmiş örnekleri kullanarak yeni örneklerin sınıflarını bulmak



Metotlar:
Regresyon,
Karar Ağaçları,
LVQ,
Yapay Sinir Ağları,
...

Mavi ve gri sınıftan örnekler ● ○
Beyaz, mavi mi gri mi? ○

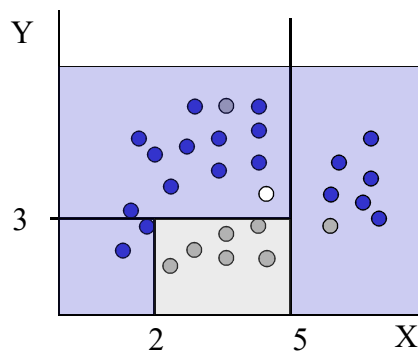
Doğrusal Regresyon



- $w_0 + w_1 x + w_2 y \geq 0$
- Regresyon en az hata yapan w_i leri bulmaya çalışır.
- Basit bir model
- Yeterince esnek değil

Karar Ağaçları

Böl ve yönet stratejisi



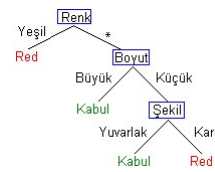
Nasıl böleceğiz?

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

Karar Ağaçları

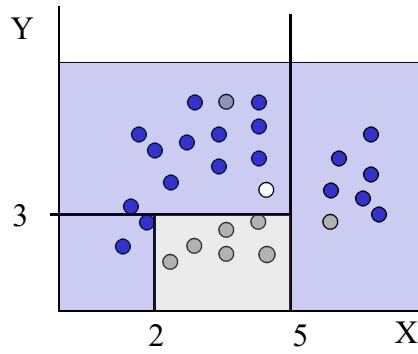
- Ürettikleri kurallar anlaşılır.
- Karar düğümleri ve yapraklardan oluşan hiyerarşik bir yapı.

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



Karar Ağaçları Oluşturma

- Tüm veri kümesiyle başla.
- Bir özelliğin bir değerlerine göre veri kümesi iki alt kümeye böl. Bölmede kullanılan özellikler ve değerleri karar düğüme yerleştir.
- Her alt küme için aynı prosedür her alt kümede sadece tek bir sınıfa ait örnekler kalıncaya kadar uygula.



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Karar Düğümleri Nasıl Bulunur?

- 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 düğüme yerleştirilir.

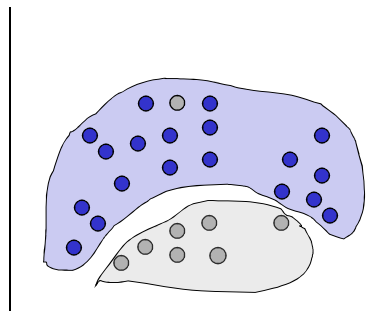
Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Karar Ağaçlarıyla Sınıflandırma

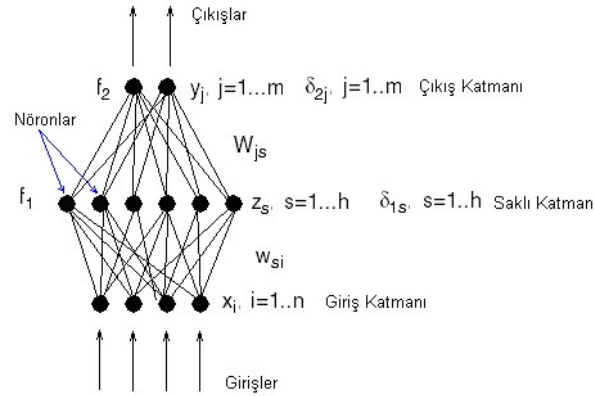
- En tepedeki kök karar düğümünden başla.
- Bir yaprağa gelinceye kadar karar düğümlerindeki yönlendirmelere göre dallarda ilerle. (Karar düğümlerinde tek bir özelliğin adı ve bir eşik değeri yer alır. O düğüme gelen verinin hangi dala gideceğine verinin o düğümdeki özelliğinin eşik değerinden büyük ya da küçük olmasına göre karar verilir.)
- Verinin sınıfı, yaprağın temsil ettiği sınıf olarak belirlir.

Yapay Sinir Ağları



- Daha kompleks karar sınırlar üretebilirler.
- Daha doğru sınırlar üretebilirler.

Yapay Sinir Ağları



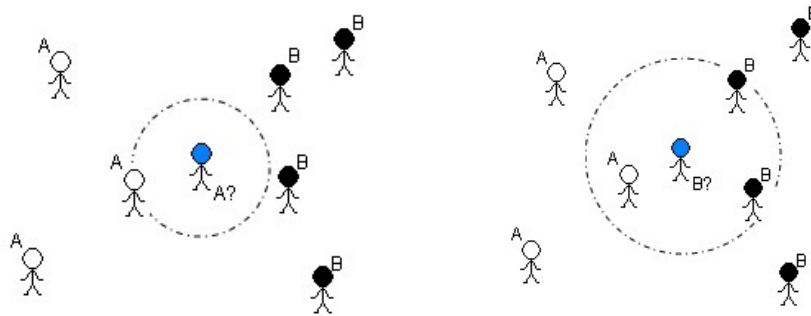
Elimizdeki Eğitim seti Girişler ve Çıkışlar ı içerir.
Bu girişler verildiğinde bu çıkışları verecek
Ağırlık değerlerini (W) bulmaya çalışır.

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

En Yakın K Komşu

- Bana Arkadaşını söyle, sana kim olduğunu söyleyeyim.



En yakın 1 komşu

En yakın 3 komşu

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

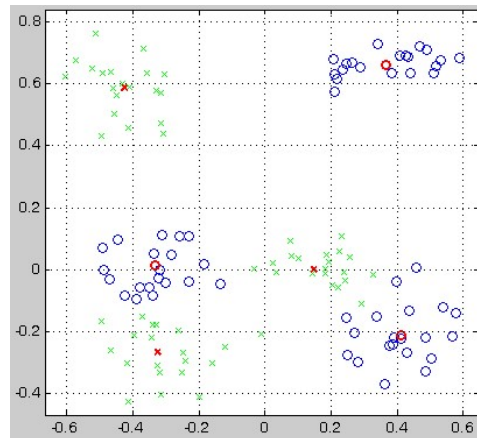
YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

En Yakın K Komşu

- Eğitim yok.
- Test verileri en yakınlarındaki K adet komşularının sınıf değerlerine bakılarak sınıflandırılırlar.

1NN'deki Tüm Noktaları kullanmasak?

- Tüm noktalar yerine yeni/ seçilmiş noktalar
- Bu noktalar nasıl belirlenebilir?



Öğrenmeli Vektör Kuantalama (Learning Vector Quantization)

[η] öğrenme oranı

[n] maximum eğitim sayısı

[c] betimleyici vektör sayısı

[μ_1, \dots, μ_c] betimleyici vektörler (centroids)

[x] eğitim datasından bir örnek

[$S(x)$] x vektörünün ait olduğu yada betimlediği sınıf olmak üzere

1. $\eta, n, \mu_1, \dots, \mu_c$ için ilk değer atamalarını gerçekleştir

2. Aşağıdaki işlemleri n defa tekrar et

2.1 X eğitim datasını al

2.2 X e en yakın betimleyici vektörü bul

$(\mu_k) : k \leftarrow \arg \min_j \|x - \mu_j\| \quad j=1..c$

2.3 μ_k nın güncellenmesi:

Eğer x doğru sınıfsa ($S(x)=S(\mu_k)$ sınıfları aynı ise)

$\mu_k \leftarrow \mu_k + \eta(x - \mu_k)$ ödüllendir x 'e yaklaştır

değilse

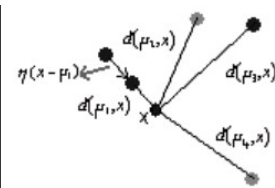
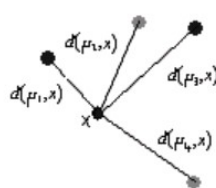
$\mu_k \leftarrow \mu_k - \eta(x - \mu_k)$ cezalandır x 'den uzaklaştır

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

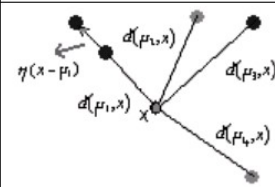
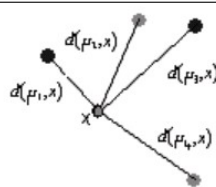
YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

LVQ'da eğitim

LVQ'da ödüllendirme
Kazanan vektörle, örnek
aynı sınıftan (ikisi de siyah
sınıftan)



LVQ'da cezalandırma
Kazanan vektörle, örnek
farklı sınıflardan (kazanan
siyah, örnek gri sınıftan)

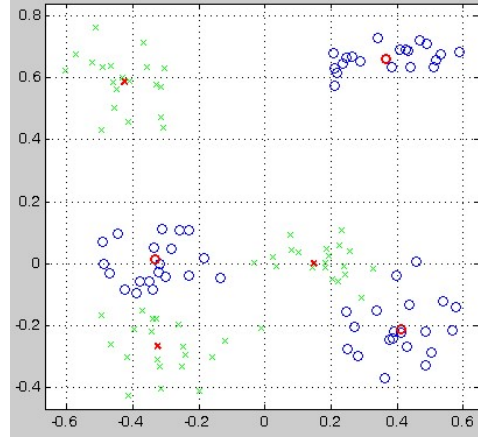


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

LVQ- Test İşlemi

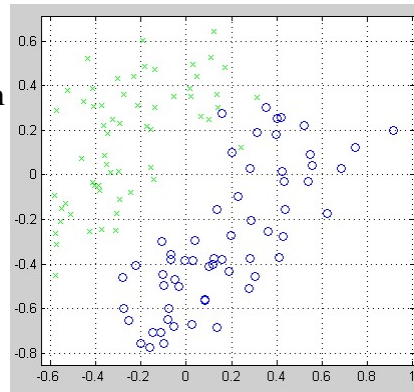
- Eğitim sonucu bulunan 2 sınıfa ait 3'er betimleyici vektör.
- Test işlemi, test örneğinin bu 6 vektörden en yakın olanının sınıfına atanmasıdır.



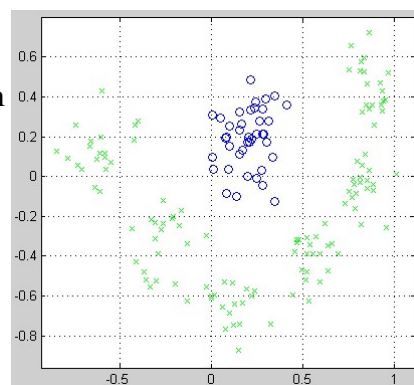
- LVQ ne zaman lineer/doğrusal karar sınırı üretir?

Hangisi?

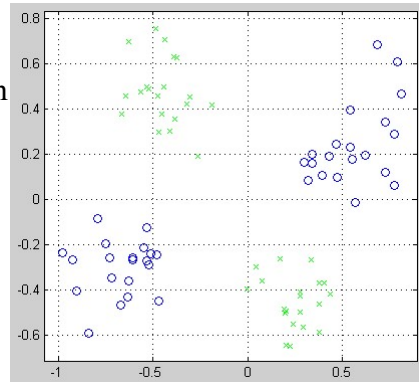
Lineer Regresyon
Karar Ağaçları
KNN
YSA
LVQ



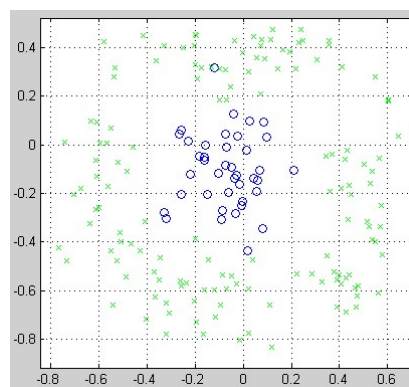
Lineer Regresyon
Karar Ağaçları
KNN
YSA
LVQ



Lineer Regresyon
Karar Ağaçları
KNN
YSA
LVQ



Başka bir model?



Eğitim kümesi üzerindeki performans neyi ifade eder?

- Borsa oynayan maymunlar*
- 1000 kişiye hisse senedi tahminleri göndermek*
- * Hatasız Düşünme Sanatı'ndan

YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

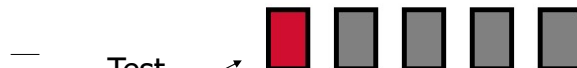


Hata nasıl ölçülür? Çapraz Geçerleme

—Tüm dataseti eşit boyutlu N gruba böl



—Bir grubu test için geriye kalanların hepsini eğitim için kullan

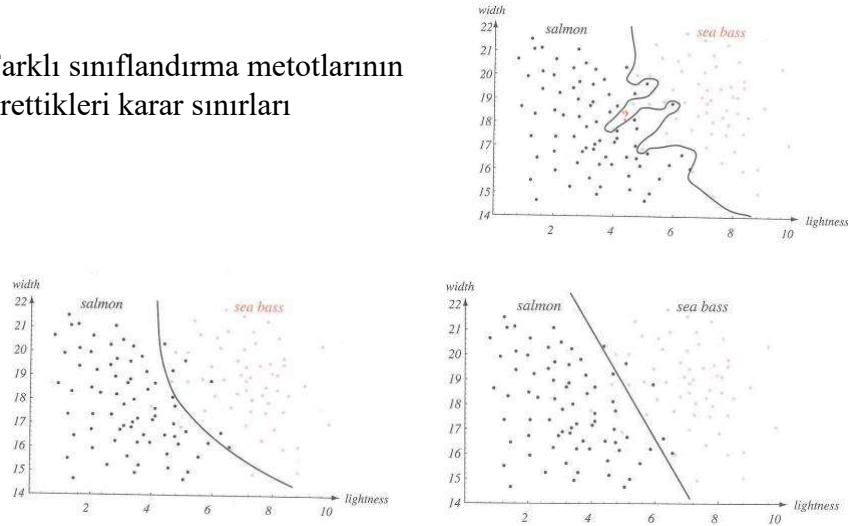


—N defa tekrar et



Sınıflandırma Metotları- Sonuç

Farklı sınıflandırma metotlarının ürettikleri karar sınırları



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Sınıflandırma Metotları- Sonuç

- Neden bu kadar çok algoritma var?
- Ne zaman hangisini seçeceğiz?

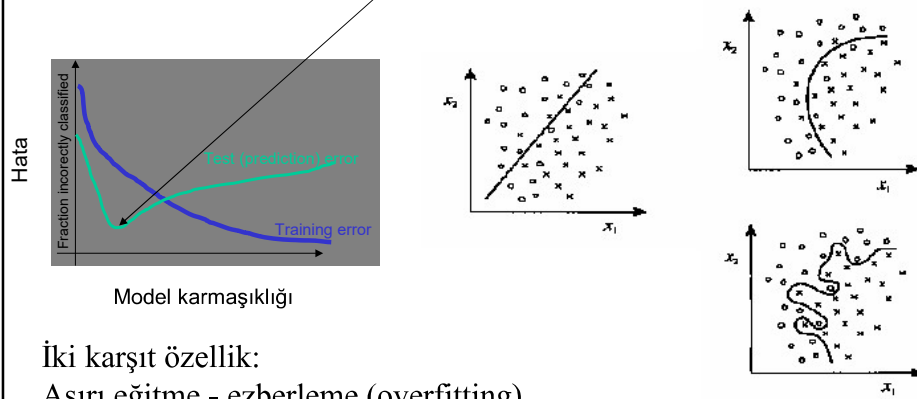
dataset	amlall	ann	bi75ds3	derma	gkanser	Hava
Özellik sayısı	7129	21	470	34	30	34
Sınıf sayısı	2	3	9	6	2	2
Örnek sayısı	72	3772	315	286	456	281
NB	97,14	95,55	68,49	77,97	94,29	89,31
SVM	92,86	93,74	62,11	79,37	96,26	86,48
1NN	94,29	93,4	63,19	76,26	96,26	89,72
C45	83,39	99,58	65,01	75,2	93,62	91,82
RF	95,71	99,5	72	76,96	95,38	95,02

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Modelin karmaşıklığı arttığında eğitim kümesindeki hata sürekli düşerken, test kümesindeki hata bir noktadan sonra yükselir.

Her veri kümesi için optimum nokta (optimum karmaşıklık) farklıdır.



İki karşıt özellik:

Aşırı eğitme - ezberleme (overfitting)

Genelleştirebilme

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Kümeleme Algoritmaları

- Kümeleme algoritmaları eğiticişiz öğrenme metotlarıdır.
- Örneklerle 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.

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

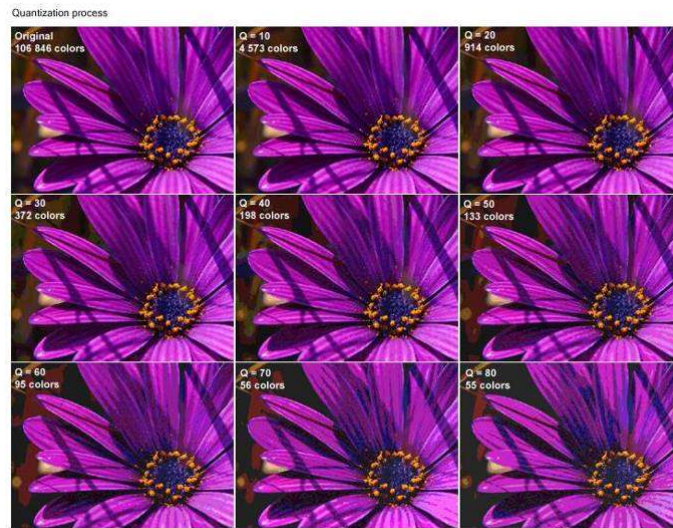
Kümeleme Algoritmaları

- 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



Resimdeki farklı renk sayısı 106846'dan 55'e indirilmiştir.

Resim Kümeleme



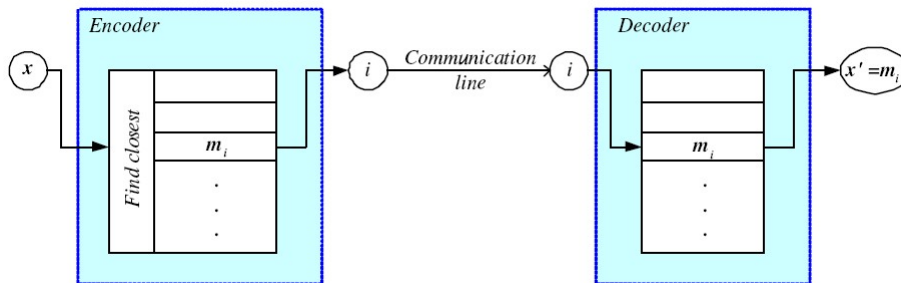
10*10 luk blokları ifade eden vektörler kümelenmiş

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Nasıl Kullanılır?

Bulunan (renkleri yada blokları temsil eden) küme merkezlerinden bir kod kitabı oluşturulur. Bu kitap her iki merkeze verilir. Vektörlerin kendileri yerine sadece indisler kullanılır. İndisin maximum büyüklüğü kodlanması için gereken bit sayısını artırır. Bu yüzden farklı vektör sayısının az olması istenir.



ETHEM ALPAYDIN © The MIT Press, 2004

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Hiyerarşik Kümeleme

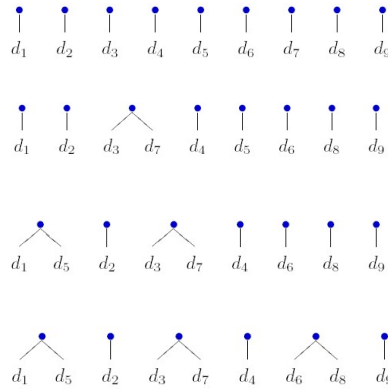
- Çok kümeden tek kümeye (eklemeli)
- Tek kümeden tek kümeye (bölmeli)

Hiyerarşik Eklemeli Kümeleme

- Birbirine en benzeyen iki kümeyi birleştir
- Tekrar et

Hiyerarşik Eklemeli kümeleme

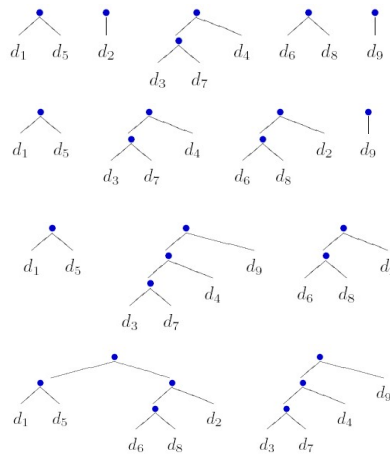
Başlangıçta küme sayısı= örnek sayısı



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Hiyerarşik Eklemeli kümeleme



Sonuçta küme sayısı=2

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Kümelerin birbirine benzerliği

- İki kümenin benzerliği
 - En benzer elemanları (Single link)
 - En benzemeyen elemanları (Compete link)
 - Ortalamaları (Group average)

kullanılarak bulunabilir.

Bölmeli Hiyerarşik Kümeleme

- Tek bir kümeyle başla
- Küme içinde birbirine en az benzeyen iki elemanı bul.
- Kümeyi bu iki elemana yakınlığa göre böl.
- Oluşan her alt küme için bu işlemi tekrar et.

Ne zaman duracağız

- İstenen küme sayısına ulaşıncaya
- Önceden belirlenmiş bir toplam benzerlik eşik değerine göre

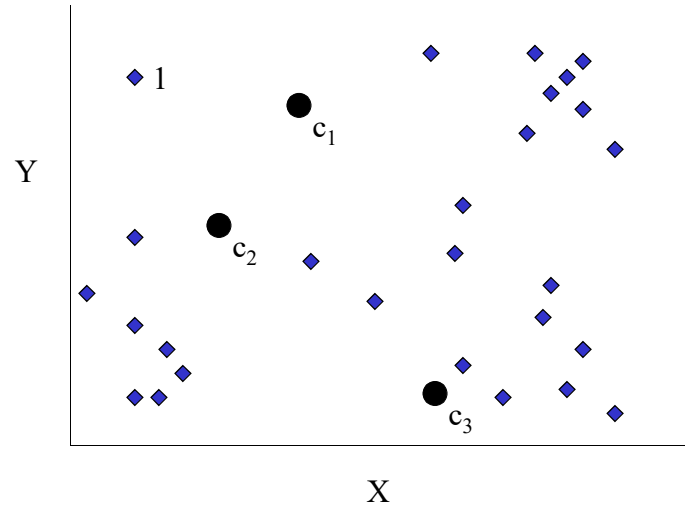
K-means

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 örnek adım 1

Rasgele
3 küme
merkezi
ata.

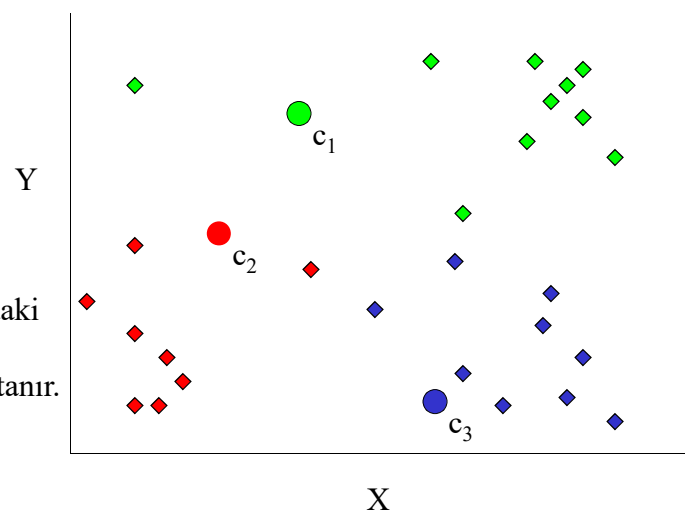


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

K-means örnek adım 2

Her örnek
en yakınındaki
merkezin
kümesine atanır.

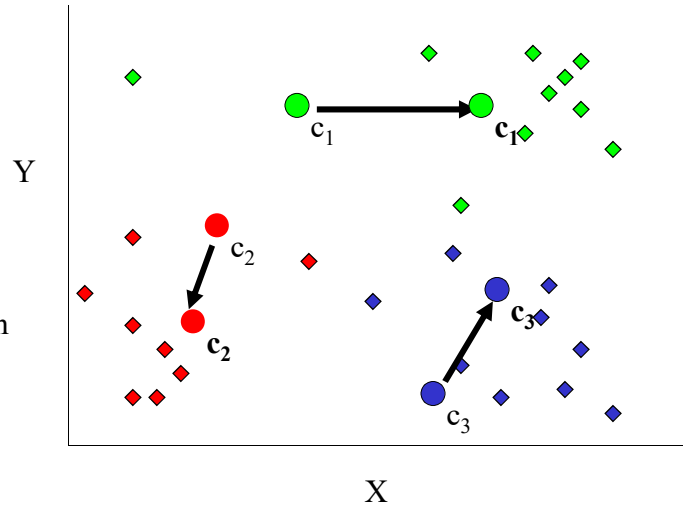


Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

K-means örnek adım 3

Merkezleri
kendi
kümelerinin
merkezine
götür.



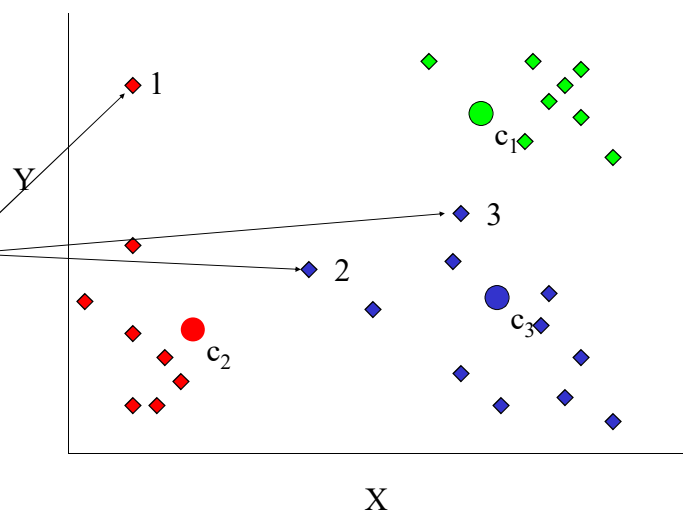
Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

K-means örnek adım 4

Her örneği
yeniden en
yakınındaki
merkezin
kümesine
ata.

*Q: Hangi
örneklerin
kümesi
değişti?*



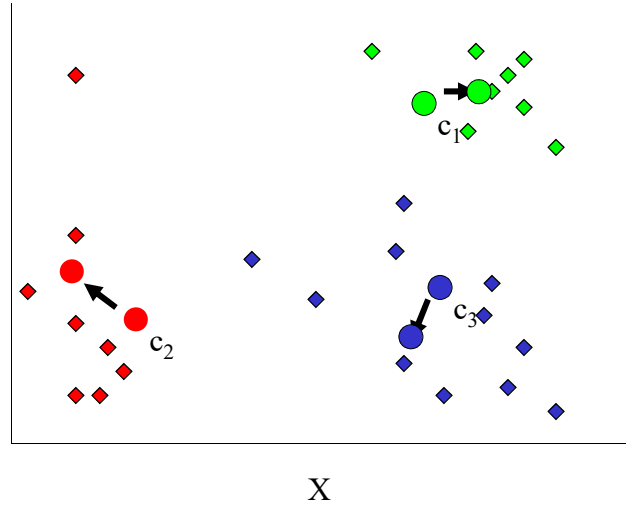
Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

K-means örnek adım 5

Y

Merkezleri
kendi
kümelerinin
merkezine
götür.



Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Tekli (Online) vs. Toplu (Batch) K-means

- Önceki örnekte yapılan işlem toplu (batch) k-means
- Eğer her bir örnek için mean'ler güncellenirse → tekli k-means

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

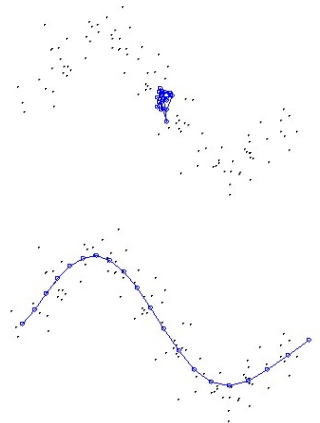
YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Kendi Kendini Düzenleyen Haritalar Self Organizing Maps

- Kmeans algoritmasında merkez noktalar arasında herhangi bir ilişki yoktur. SOM'da ise merkez noktalar 1 ya da 2 boyutlu bir dizi içinde yer alırlar. Buna göre birbirlerine 1 ya da 2 boyutlu uzayda komşudurlar.
- Kmeans algoritmasında sadece kazanan (en yakın) merkez güncellenirken SOM'da bütün merkezler kazanan nörona komşuluklarına göre güncellenir. Yakın komşular uzak komşulara göre daha fazla hareket ederler (güncellenirler).
- Merkezlerin birbirlerine bağlı oluşu verinin 1 ya da 2 boyutlu uzaydaki yansımasının da elde edilmesini sağlar.

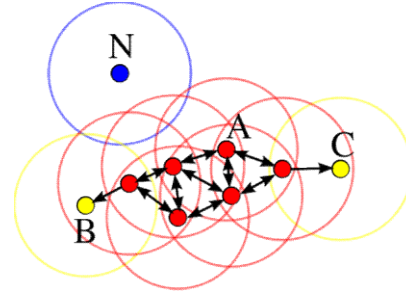
SOM

- Örnekte SOM merkezleri 1 boyutlu bir dizide birbirlerine komşudurlar. Başlangıçtaki durumları rasgele atandığı için bir yumak şeklindedirler. Eğitim tamamlandığında ise SOM merkezleri verinin şeklini almıştır.



DBscan

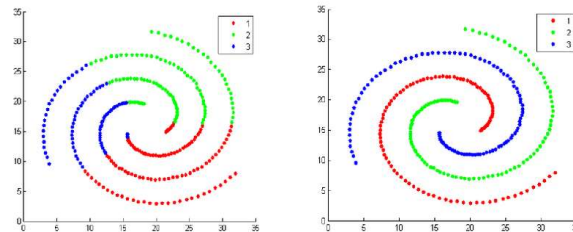
- Yoğunluk tabanlı kümeleme
- Hiper-parametreleri: eps, min_p
- Önce tüm noktaları çekirdek, sınır, gürültü olarak belirle
- Çekirdek (kırmızı): eps komşuluğunda en az min_p adet nokta bulunan noktalar
- Gürültü (mavi): eps komşuluğunda hiç nokta bulunmayan noktalar
- Sınır (sarı): eps komşuluğunda min_p'den az nokta bulunan noktalar



[*] <https://en.wikipedia.org/wiki/DBSCAN>

DBscan

- Noktalar etiketlendikten sonra gürültüler elenir
- Çekirdeklerden birbirine eps mesafede olanlar birleştirilir.
- X (hiper parametrelere bağlı) adet küme oluşur.
- Sınır noktalar en yakın çekirdeğin kümesine atanır



[*] https://www.researchgate.net/figure/Results-of-a-k-means-b-DBSCAN-on-Jain-dataset_fig2_291831757

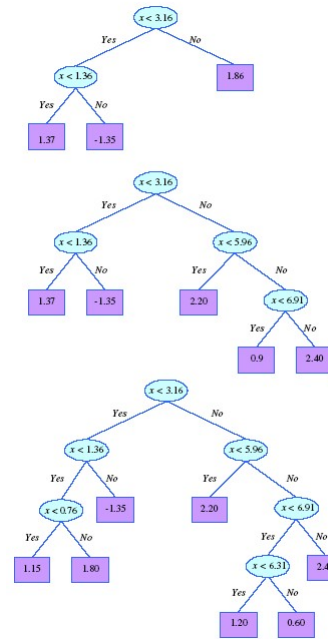
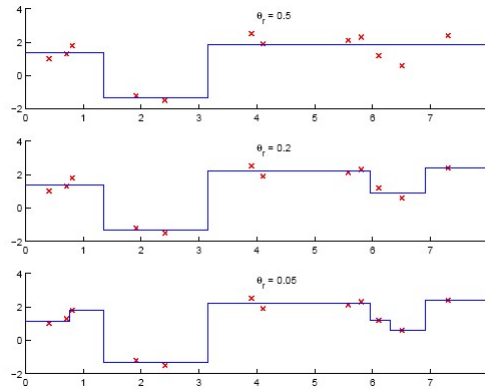
Kümeleme algoritmalarında

- Ne zaman dururuz?
- Ölü nöron kavramı (SOM, K-means)
- Başarı ölçütü nedir?
- **Küme sayısı nasıl belirlenir?**
 - Mecburiyet (verilmiştir)
 - ?

Regresyon Algoritmaları

- Basit lineer regresyon ($y=w_1x+w_0$)
- kNN
- Regresyon ağaçları
- YSA

Regresyon ağaçlarında kabul edilen hata değerinin etkisi

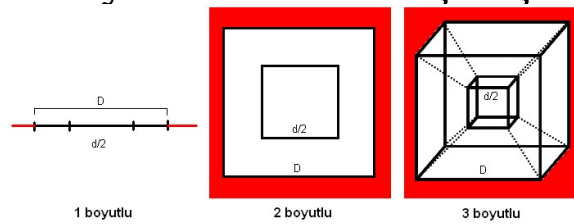


ETHEM ALPAYDIN © The MIT Press, 2004

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Çok Boyutlu Verilerle Çalışmak-1



Boyut Sayısı	Merkeze daha yakın noktaların oranı (%)
1	50
2	25
3	12,50
...	...
P	$(\frac{1}{2})^P$

Boyut sayısı arttığında verilerin çok büyük bir kısmı sınıfları ayıran sınırlara çok yakın yerlerde bulunacağından sınıflandırma yapmak zorlaşmaktadır.

Mehmet Fatih AMASYALI Yapay Zeka Ders Notları

YILDIZ TEKNİK ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Çok Boyutlu Verilerle Çalışmak-2

- 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 (Data Leakage)

- 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 tanısin.
- Verilerde çıkışla korelasyonu çok yüksek olan özelliklerin olması (kişi tanırken id, telefon no vb.)

Sonuç olarak

- Makineler insanlığın işgücüne sağladıkları katkıyı, makine öğrenmesi metotları sayesinde insanlığın beyin gücüne de sağlamaya başlamışlardır.

Bir gün bilgisayarlar
bütün bunları mükemmel bir şekilde
yaparlarsa Nasıl bir dünya

- Bir sürü işsiz bilgisayar mühendisi ☺
- Bir sürü işsiz insan
- ???

Kaynaklar

- Alpaydın E. (2004) "Introduction to Machine Learning", The MIT Press, 3-6
- <http://www.autonlab.org/tutorials/infogain11.pdf>
- http://www.kdnuggets.com/dmcourse/data_mining_course/assignments/assignment-4.html
- http://pespmc1.vub.ac.be/asc/SENSIT_ANALY.html
- http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- http://www.cavs.msstate.edu/hsc/ies/publications/reports/isip_internal/1998/linear_discrim_analysis/lda_theory.pdf
- <http://www.kernel-machines.org>
- T. Kohonen, " Self-Organization and associative Memory", 3d ed, 1989, Berlin :Springer-Verlag.
- <http://www.willamette.edu/~gorr/classes/cs449/Classification/perceptron.html>
- O. T. Yıldız, E. Alpaydın, Univariate and Multivariate Decision Trees, Tainn 2000
- <http://www.ph.tn.tudelft.nl/PHDTheses/AHoekstra/html/node45.html>
- <http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html>
-

Weka



Copyright: Martin Kramer (mkramer@wxs.nl)