Doğal Dil İşlemede Eğilimler

Önceden: Yapay Zeka Tabanlı, tam olarak anlama

Şimdiki: Külliyat(Corpus)-tabanlı, İstatistiki, makine öğrenmesi içeren

Akış

- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- Verilerin Sayısallaştırılması
- Özellik Belirleme
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

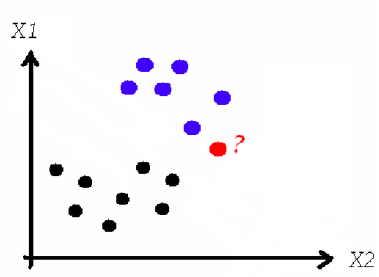
Makine Öğrenmesi

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 **model**i bulmaya çalışırlar.
Yeni gelen verileri de bu modele göre analiz edip sonuç üretirler.

Metot türleri

- Farklı uygulamaların analizlerden farklı beklentileri olmaktadır.
- Makine öğrenmesi metotlarını bu beklentilere göre sınıflandırmak mümkündür.

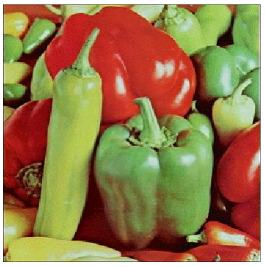
Sınıflandırma



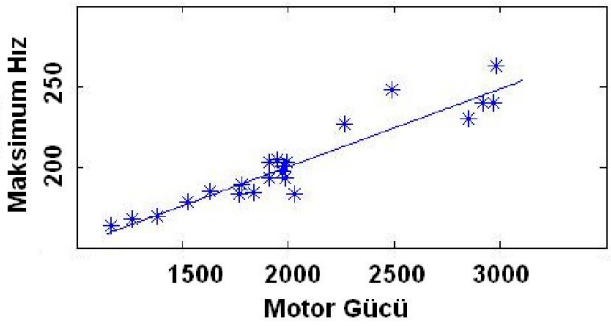
Kırmızı hangi sınıftan?

Kümeleme

- 256 rengi 16 renge nasıl indiririz?



Regresyon
Eğri Uydurma



Birliktelik Kuralları Keşfi

- Sepet analizi
  - hangi ürünler birlikte satılıyor?
- Raf düzenlemesi
  - hangi ürünler yan yana konmalı?
- Promosyonlar
  - neyin yanında ne verilmeli?

Akış

- Makine Öğrenmesi nedir?
- **Günlük Hayatımızdaki Uygulamaları**
- Verilerin Sayısallaştırılması
- Özellik Belirleme
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

Makine Öğrenmesinin

Günlük Hayatımızdaki Uygulamaları

El yazısı / Kitap Yazısı Tanıma
HCR /OCR

İşlem: Şekillerin hangi harf olduğunu tahmin etme

Kredi Taleplerini Değerlendirme

- Birisi bankadan borç ister.
- Banka borcu versin/vermesin.
- Nasıl?

e-ticaret

- Birisi Amazon.com dan bir kitap yada ürün alıyor.

Görev ne olabilir?

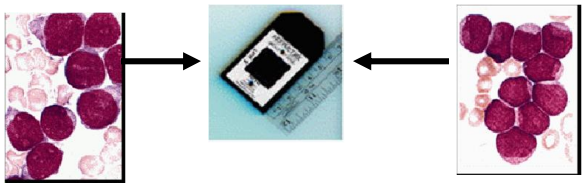
e-ticaret

- Birisi Amazon.com dan bir kitap yada ürün alıyor.

Görev ne olabilir?

Müşteriye alması muhtemel kitapları önerelim.
Ama nasıl?
Kitapları
- konularına
- yazarlarına
- birlikte satılışlarına

göre kümelemek?

ALL  AML

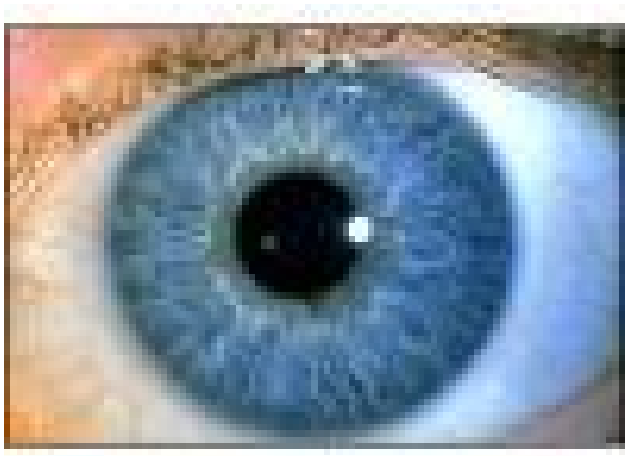
Gen Mikrodizilimleri

100 kişinin (hasta/sağlam) elimizde gen dizilimleri var. Bu dizilimleri analiz ederek hasta olup olmadığı bilinmeyen birisinin hasta olup olmadığını yada hastalığının türünü öğrenebilir miyiz?

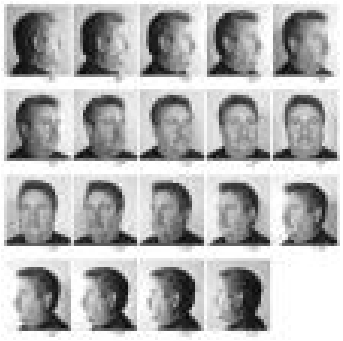
En iyi tedaviyi önerebilir miyiz?

Nasıl? Elimizde hangi bilgiler olmalı?

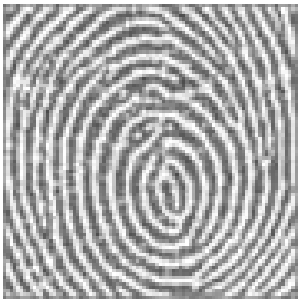
Bu adam kim? İçeri girsin mi?



Bu adam kim?
Bu adam havaalanında mı?



Bu parmak izi kimin?
Bu adamı tutuklayalım mı?



Bu ses kimin?
Bu ses ne diyor?

Bu imza kimin? Yoksa taklit mi?

Taklit olup olmadığını nasıl anlarız?
Zaman bilgisi ?

Bu metnin konusu nedir?
Bu mail spam mi?

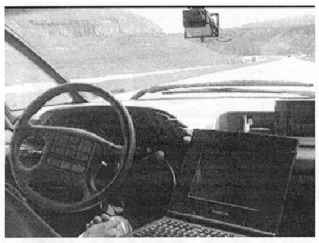
Anti spam yazılımları nasıl çalışır?
Spamciler nasıl çalışıyor?
Yeni nesil spam mailler: Mesaj resimde,
metinde ise anti spamlardan kaçmak için gereken kelimeler var.
Makine öğrenmesi metotlarını hem spamciler hem anti spamciler kullanıyor.

Olağan dışı bir durum var mı?
Güvenlik kamerası kayıtları

Kamera kaydındaki kişi ne anlatıyor?

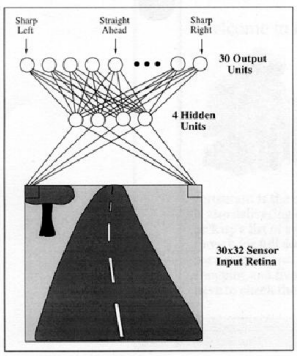
Sonuç:İletişimin artması

ALVIN



Otobanda saatte 70 mil hızla **sürücüsüz** gidebilen bir otomobil

Bütün denemeler trafiğe kapalı alanlarda gerçekleştirilmiştir ☺



Neden şehiriçi değil?
Neden otoban?
Neden diğer arabalar yok?
Araba birine çarparsa suçlu kim?

Adalet

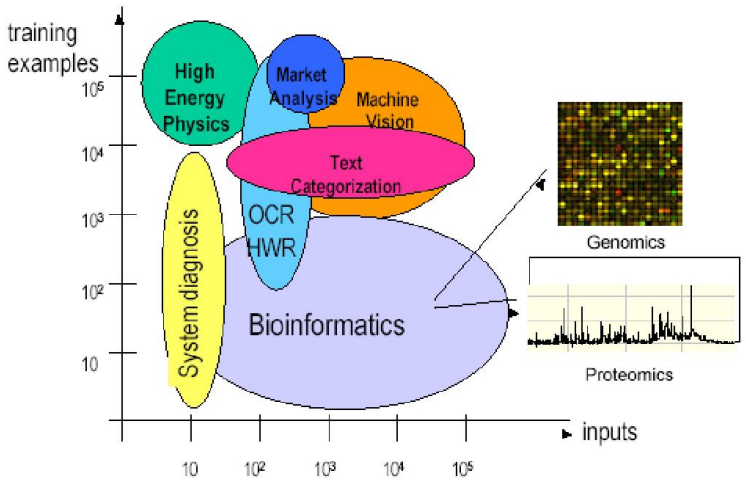
- Çin'de pilot uygulama:
  - bir şehrin mahkeme hakimleri bir bilgisayar programı
  - Amaç: Daha adil bir dünya
    - Aynı özelliklere sahip davalarda aynı kararların alınması
  - Sistemin eğitimi için neler gerekli?
    - Milyonlarca/Milyarlarca (orası Çin) davaya ait verilerin kategorilenmesi

Beyin Aktiviteleri

- İnsanların
  - değişik şeyler düşünürkenki,
  - değişik duygulara sahipkenki,
  - problem çözerken ki

  beyin aktiviteleri kaydedilir.
- Görev?

Uygulamalardaki boyut örnek sayıları



Learning is the future

- Learning techniques will be a basis for every application that involves a connection to a real world
- Basic learning algorithms are ready for use in limited applications today
- Prospects for broader future application make for exciting fundamental research and development opportunities
- Many unresolved issues – Theory and Systems

Akış

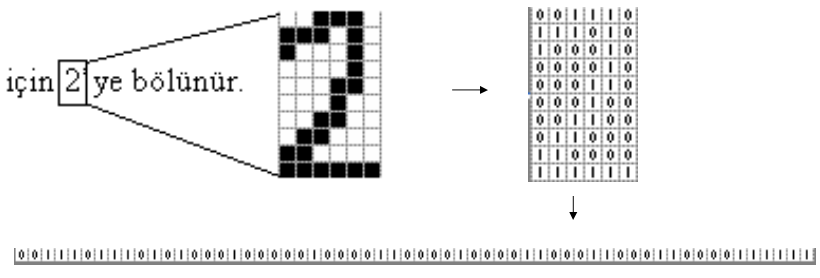
- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- **Verilerin Sayısallaştırılması**
- Özellik Belirleme
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

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.

Örnek sınıflandırma uygulaması

- Sistem: bir kitap fotokopisinin içindeki yazılarının metne dönüştürülmesi
- Öncelikle metindeki satırlar bulunur.
- Her bir satırdaki harfler bulunur. Her harfe ait onlarca örnek resimden etiketlenmiş bir veritabanı oluşturulur. Her bir resim için



- Bu şekilde tanınmak istenen harf için çeşitli fontlarla yazılmış birçok örneği temsil eden 60 boyutlu vektörler elde edilir.
- Bu uygulamamız için özellik sayımız 60'tır. Diğer bir deyişle örneklerimiz 60 boyutlu bir uzayda temsil edilmektedirler.
- Elimizde 10 rakama ait farklı fontlarla yazılmış 10'ar resim olursa veri kümemiz 100 örnek* 60 boyutluk bir matris olacaktır.
- Elimizde her örneğin hangi harf olduğunu gösteren sınıf bilgiside bulunmaktadır.
- Bu matris eğitim ve test kümesi oluşturmak için 2'ye bölünür.
- Eğitim kümesi bir sınıflandırıcıya verilir.
- Modellenir.
- Modelin başarısını ölçmek için sınıflandırıcının daha önce görmediği, modelini oluşturmakta kullanmadığı test kümesi için tahminde bulunması istenir.
- Bu tahminlerle gerçek sınıfların ayınılığının ölçüsü sınıflandırıcının başarı ölçüsüdür.

Akış

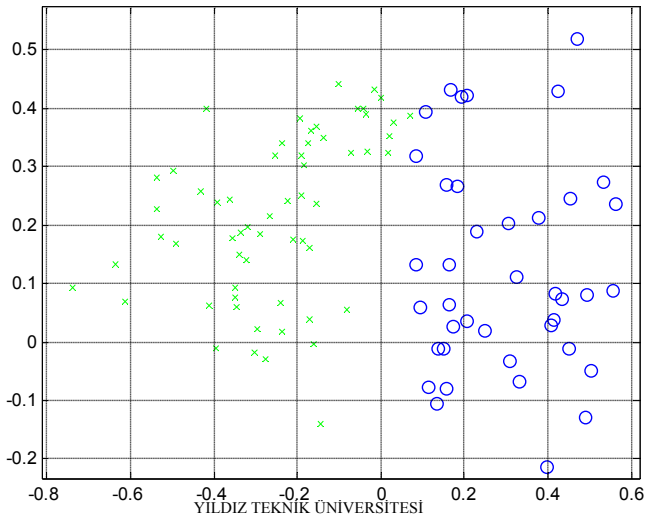
- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- Verilerin Sayısallaştırılması
- **Özellik Belirleme**
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

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.

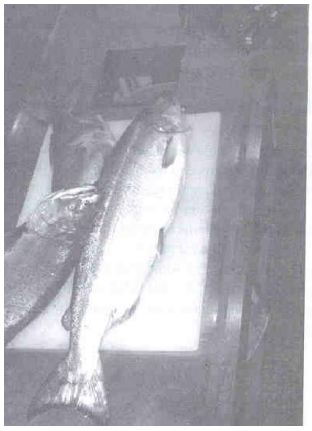
1. Özellik	2. Özellik	Sınıf
1	3	A
2	3	B
1	4	A
2	3	B

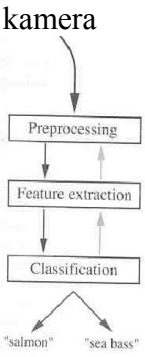
Hangi boyut?



Balık Hali

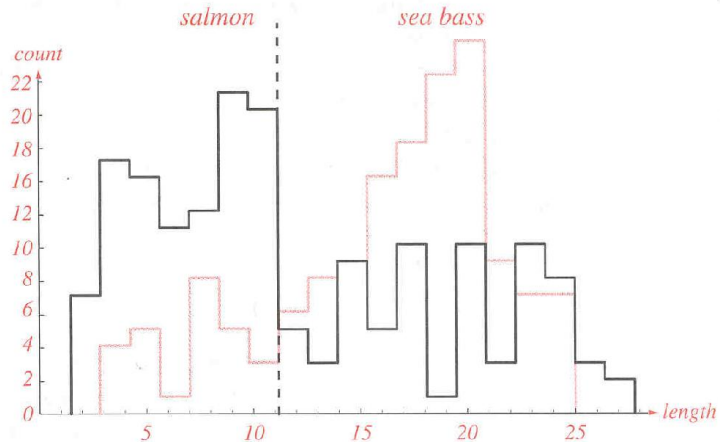
- Kayan bant üzerindeki balığın türünü belirlemek(Salmon? Sea Bass?)





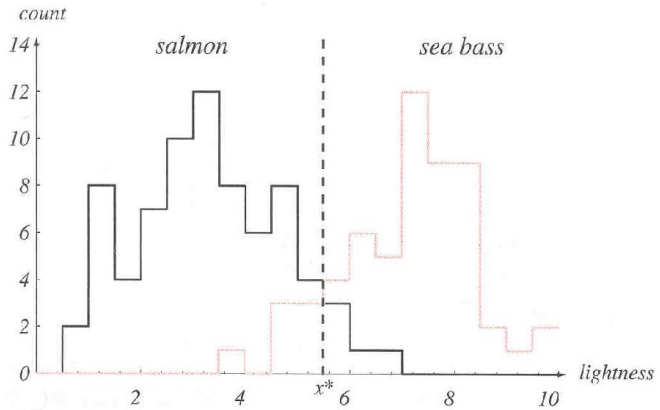
Balık Özellikleri. Uzunluk.

- Salmon lar genelde Sez Bass lardan daha kısalar.

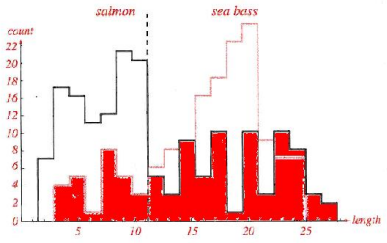


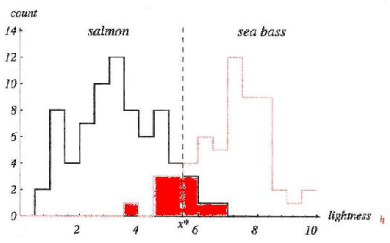
Balık Özellikleri. Parlaklık.

- Sea Bass genelde Salmon lardan daha parlaklar.



Hangi Özellik?





Kırmızı bölgeler yapılan hataları gösteriyor.

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 gözükmekte.
- 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 - Informaiton Gain

S eğitim seti içindeki A özelliğinin

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

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

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

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 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.

- **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

- **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.

S2N

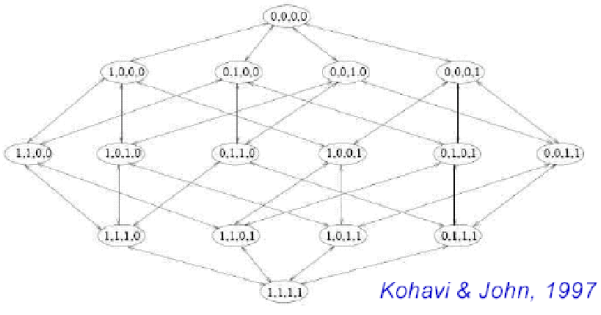
- sınıflar arası ayrılıkların fazla sınıf içi ayrılıkların az olan özellikler seçilir.

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

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.

Özellik altkümesi seçiciler (Wrappers)

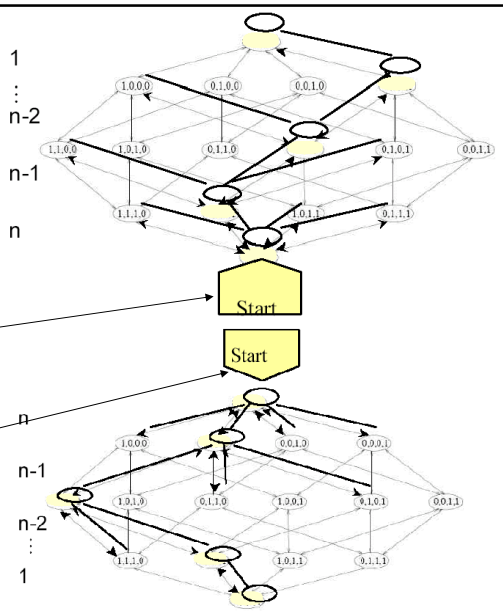
N özellik için olası 2^N özellik alt kümesi = 2^N eğitim



Kohavi & John, 1997

Özellik altkümesi seçiciler

- *Hızlandırmak için tüm olasılıkları denemek yerine*
  - Hepsiyle başlayıp her seferinde bir tane elemek
  - Tek özellikle başlayıp her seferinde bir tane eklemek



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

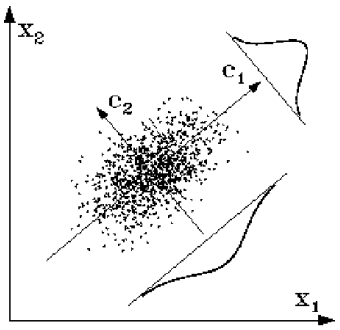
Yeni Özelliklerin Çıkarımı

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

- PCA
- LDA

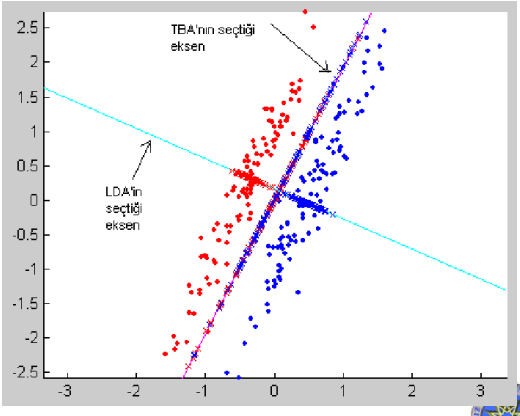
Temel Bileşen Analizi-TBA (Principle Component Analysis - PCA)

- Bu metotta örneklerin en fazla değişim gösterdikleri boyutlar bulunur. Yansa 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.



Doğrusal Ayırteden 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.



Hangisi

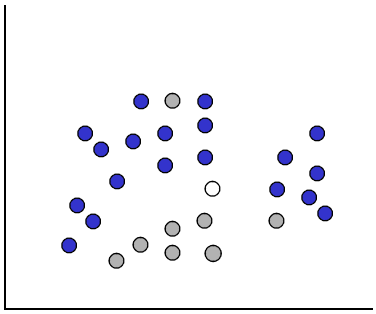
- Niye bu kadar çok metot var?
- Ne zaman hangisini kullanacağız?

Akış

- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- Verilerin Sayısallaştırılması
- Özellik Belirleme
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

Sınıflandırma Metotları

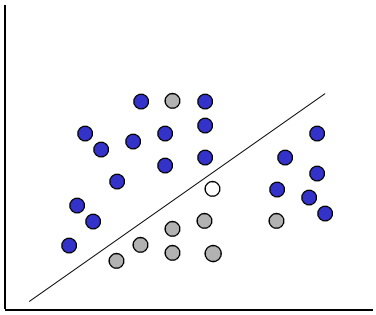
Görev: Önceden etiketlenmiş örnekleri kullanarak yeni örneklerin sınıflarını bulmak



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

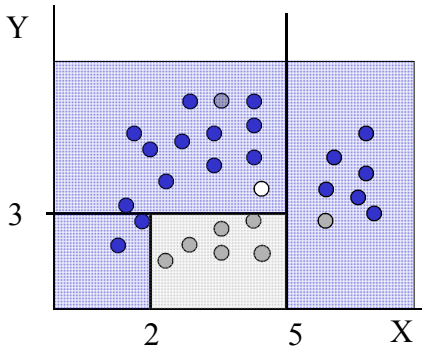
Mavi ve gri sınıftan örnekler ● ●
Beyaz, mavi mi gri mi? ○

Doğrusal Regresyon



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

Karar Ağaçları



Böl ve yönet stratejisi

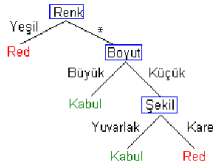
Nasıl böleceğiz?

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

Karar Ağaçları

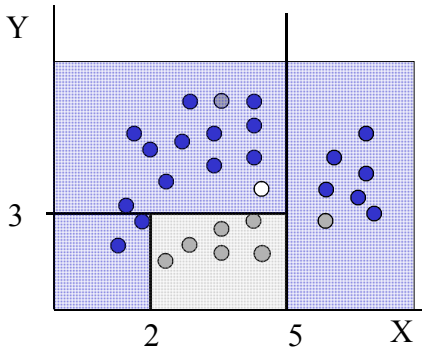
- Ürettikleri kurallar anlaşılır.
- Karar düğümleri ve yapraklardan oluşan hiyerarşik bir yapı.

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.



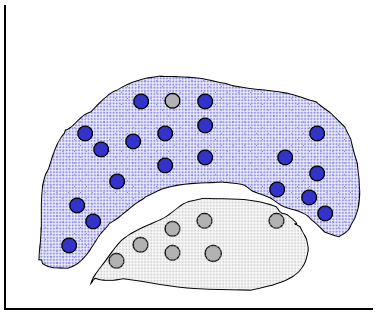
Karar Düğümleri Nasıl Bulunur?

- 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 ve değeri karar düğümüne yerleştirilir.

Karar Ağaçlarıyla Sınıflandırma

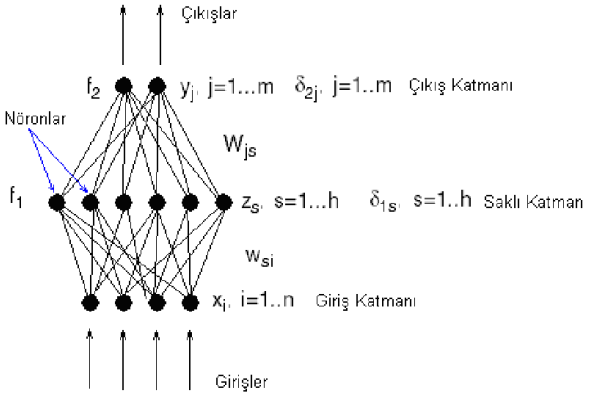
- En tepedeki kök karar düğümünden başla.
- Bir yaprağa gelinceye kadar karar düğümlerindeki yönlendirmelere göre dallarda ilerle. (Karar düğümlerinde tek bir özelliğin adı ve bir eşik değeri yer alır. O düğüme gelen verinin hangi dala gideceğine verinin o düğümdeki özelliğinin eşik değerinden büyük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

Yapay Sinir Ağları



- Daha kompleks karar sınırlar üretebilirler.
- Can be more accurate
- Also can overfit the data – find patterns in random noise

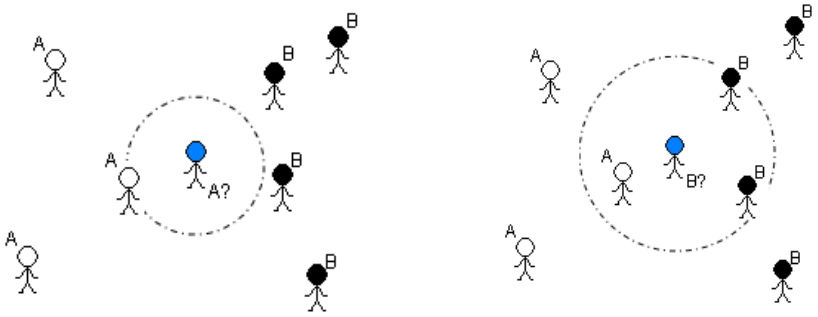
Yapay Sinir Ağları



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

En Yakın K Komşu

- Bana Arkadaşını söyle, sana kim olduğunu söyleyeyim.



En yakın 1 komşu En yakın 3 komşu

En Yakın K Komşu

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

Öğrenmeli Vektör Kuantalama (Learning Vector Quantization)

[η] öğrenme oranı
[n] maximum eğitim sayısı
[c] betimleyici vector sayısı
[$\mu_1,\ldots,\mu_c$] betimleyici vektörler (centroids)
[x] eğitim datasından bir örnek
[S(x)] x vektörünün ait olduğu yada betimlediği sınıf
olmak üzere

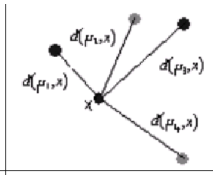
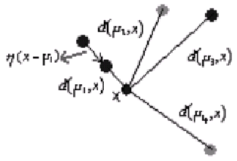
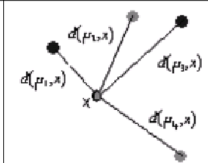
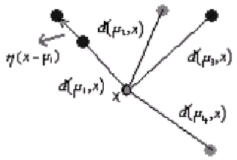
1. η,n, $\mu_1,\ldots,\mu_c$ için ilk değer atamalarını gerçekleştir
2. Aşağıdaki işlemleri n defa tekrar et
2.1 X eğitim datasını al
2.2 X e en yakın betimleyici vektörü bul

$(\mu_k): k \leftarrow \text{argmin}_j \left\| x-\mu_j \right\|$ j=1..c

2.3 μ_k nın güncellenmesi:

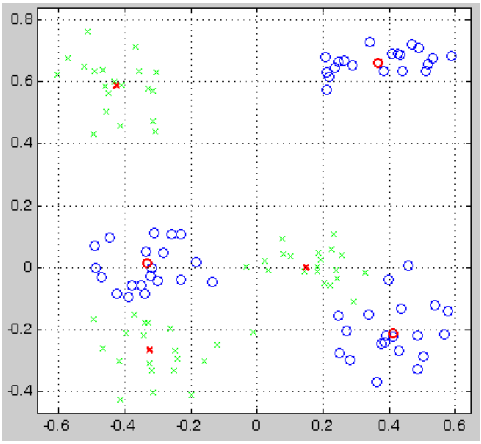
Eğer x doğru sınıfsa (s(x)=s(μ_k) sınıfları aynı ise)
$\mu_k \leftarrow \mu_k + \eta(x-\mu_k)$ ödüllendir x'e yaklaştır
değilse
$\mu_k \leftarrow \mu_k - \eta(x-\mu_k)$ cezalandır x'den uzaklaştır

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)		

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.

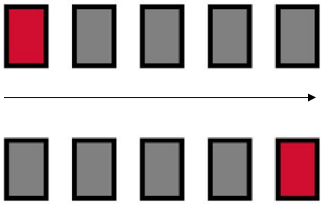


Çapraz Geçerleme

—Tüm dataseti eşit boyutlu N gruba böl
—
—

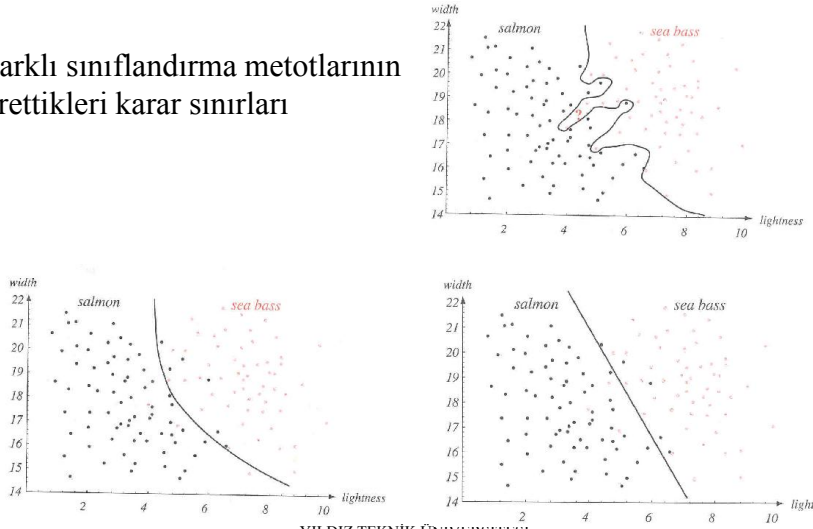
—Bir grubu test için geriye kalanların hepsini eğitim için kullan

—
Test
—N defa tekrar et



Sınıflandırma Metotları- Sonuç

Farklı sınıflandırma metotlarının ürettikleri karar sınırları



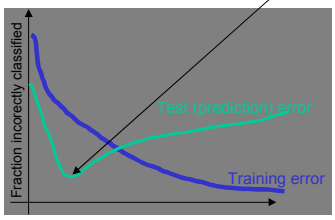
Sınıflandırma Metotları- Sonuç

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

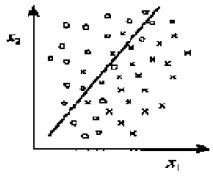
dataset	amlall	ann	bi75ds3	derma	gkanser	Hava
Özellik sayısı	7129	21	470	34	30	34
Sınıf sayısı	2	3	9	6	2	2
Örnek sayısı	72	3772	315	286	456	281
NB	**97,14**	95,55	68,49	77,97	94,29	89,31
SVM	92,86	93,74	62,11	**79,37**	**96,26**	86,48
1NN	94,29	93,4	63,19	76,26	**96,26**	89,72
C45	83,39	**99,58**	65,01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

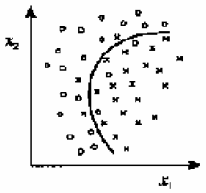
Modelim karmaşıklığı arttığında eğitim kümesindeki hata düşerken test kümesindeki hata yükselir.

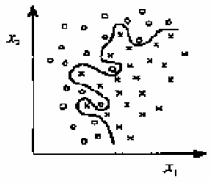
Her veri kümesi için optimum nokta (optimum karmaşıklık) farklıdır.



Model karmaşıklığı







Akış

- Makine Öğrenmesi nedir?
- Günlük Hayatımızdaki Uygulamaları
- Verilerin Sayısallaştırılması
- Özellik Belirleme
  - Özellik Seçim Metotları
    - Bilgi Kazancı (Informaiton Gain-IG)
    - Sinyalin Gürültüye Oranı: (S2N ratio)
    - Alt küme seçiciler (Wrappers)
  - Yeni Özelliklerin Çıkarımı
    - Temel Bileşen Analizi (Principal Component Analysis)
    - Doğrusal Ayırteden 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:**
  - K-Ortalama (K-Means)
  - Kendi Kendini Düzenleyen Haritalar (Self Organizing Map -SOM)

Kümeleme Algoritmaları

- Kümeleme algoritmaları eğiticisiz öğrenme metotlarıdır.
- Örneklere ait sınıf bilgisini kullanmazlar.
- Temelde verileri en iyi temsil edecek vektörleri bulmaya çalışırlar.
- Verileri temsil eden vektörler bulunduktan sonar artık tüm veriler bu yeni vektörlerle kodlanabilirler ve farklı bilgi sayısı azalır.
- Bu nedenle birçok sıkıştırma algoritmasının temelinde kümeleme algoritmaları yer almaktadır.

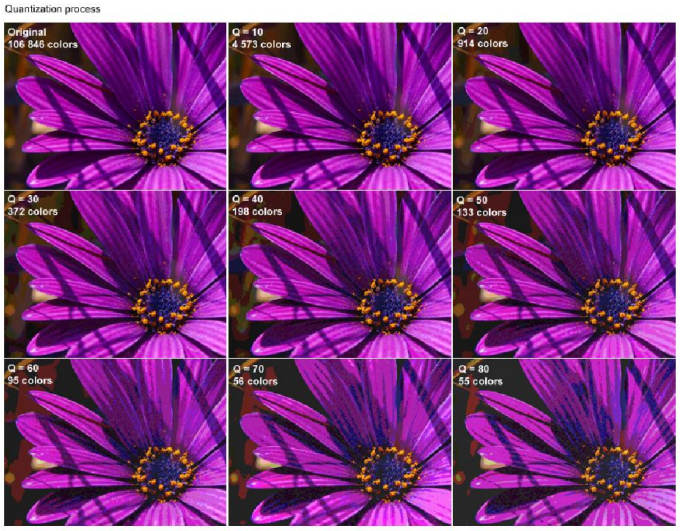
Kümeleme Algoritmaları

- 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



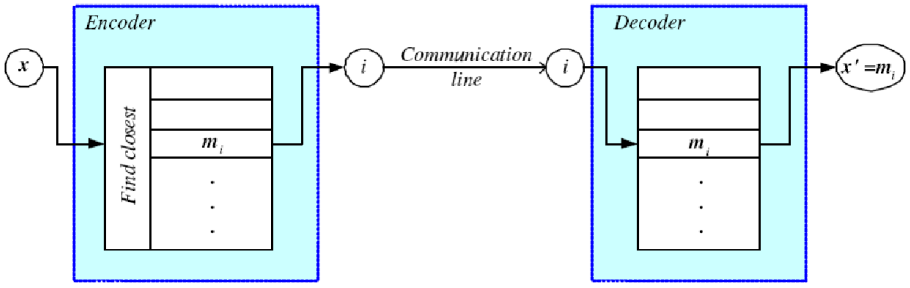
Resimdeki farklı renk sayısı 106846'dan 55'e indirilmiş.

Resim Kümeleme

10*10 luk blokları ifade eden vektörler kümelenmiş

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 arttırır.
Bu yüzden farklı vektör sayısının az olması istenir.



ETHEM ALPAYDIN © The MIT Press, 2004

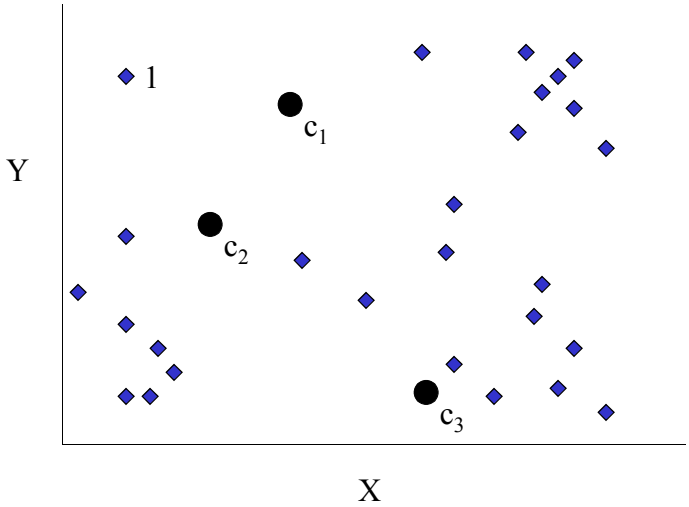
K-means

Works with numeric data only

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.

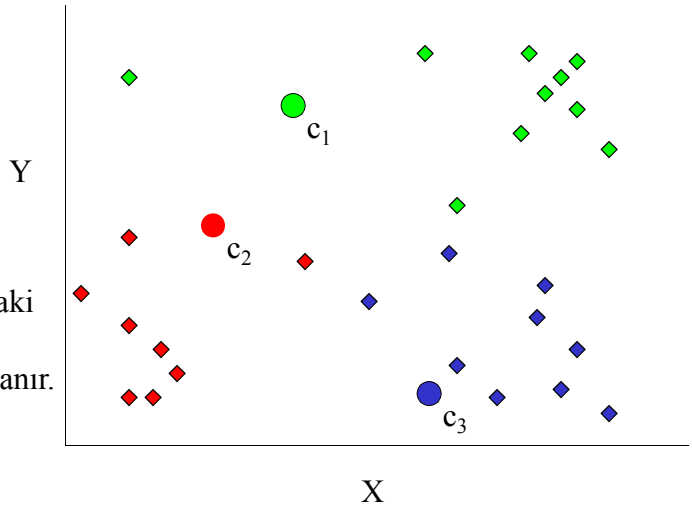
K-means örnek adım 1

Rasgele 3 küme merkezi ata.



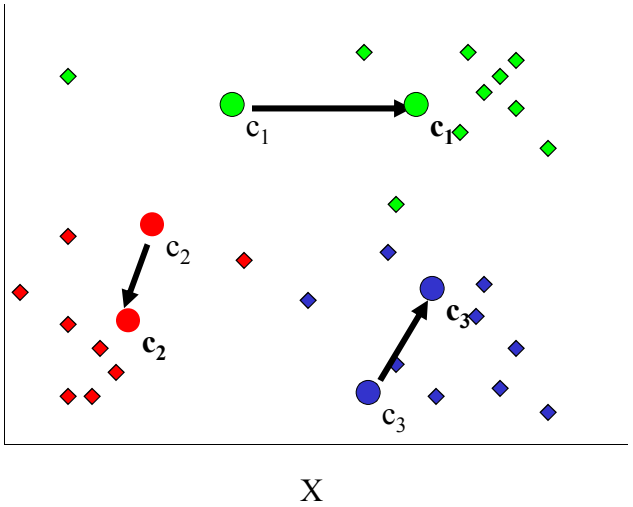
K-means örnek adım 2

Her örnek en yakınındaki merkezin kümesine atanır.



K-means örnek adım 3

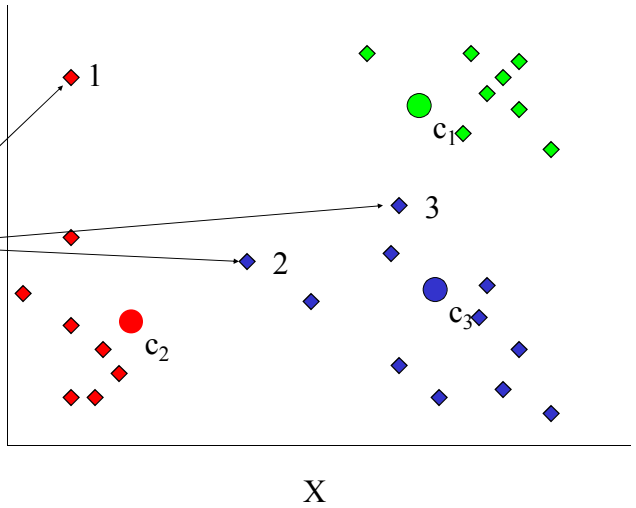
Merkezleri kendi kümelerinin merkezine götür.



K-means örnek adım 4

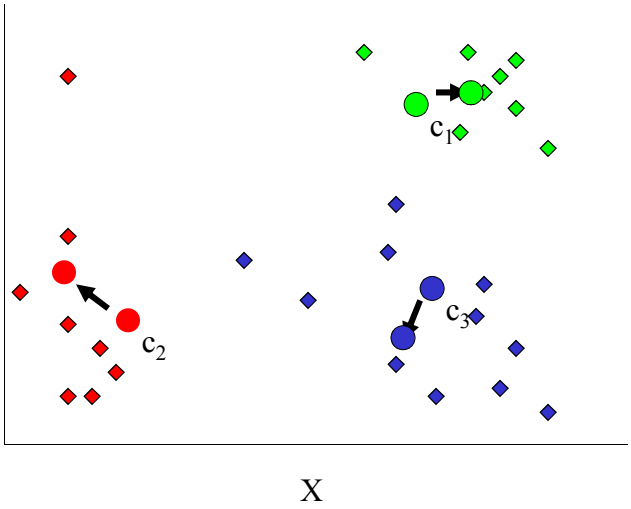
Her örneği yeniden en yakınındaki merkezin kümesine ata.

Q: Hangi örneklerin kümesi değişti?



K-means örnek adım 5

Merkezleri kendi kümelerinin merkezine götür.

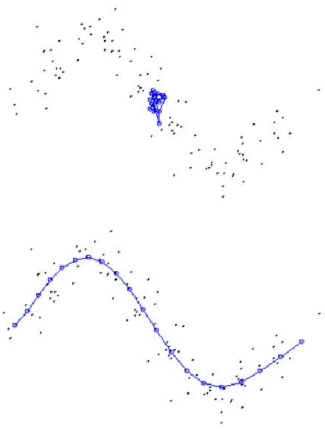


Kendi Kendini Düzenleyen Haritalar Self Organizing Maps

- 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

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



Sonuç olarak

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

Bir gün bilgisayarlar bizi anlarlarsa?

Ve 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/hse/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
- .....

YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Weka

Copyright: Martin Kramer (mkramer@wxs.nl)