

VERİ MADENCİLİĞİ

Temel Sınıflandırma Yöntemleri

Yrd. Doç. Dr. Şule Gündüz Öğüdücü
<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195/>

1

Konular

- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcılar
 - Bayes ağıları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

2

Sınıflandırma

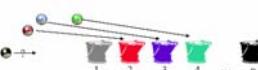
- Sınıflandırma (classification) problemi:
 - nesnelerden oluşan veri kümesi (**öğrenme kümesi**):
 $D=\{t_1, t_2, \dots, t_n\}$
 - her nesne niteliklerden oluşuyor, niteliklerden biri **sınıf** bilgisi
- Sınıf niteliğini belirlemek için diğer nitelikleri kullanarak bir **model** bulma
- Öğrenme kümesinde yer almayan nesneleri (**sinama kümesi**) mümkün olan en iyi şekilde doğru sınıflara atamak
- **sınıflandırma**=ayrık değişkenler için öngörüde (prediction) bulunma

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

3

Sınıflandırma

- Amaç: Bir niteliğin değerini diğer nitelikleri kullanarak belirlemek
 - verinin dağılımına göre bir model bulunur
 - bulunan model, başarımı belirlendikten sonra niteliğin gelecekteki ya da bilinmeyen değerini tahmin etmek için kullanılabilir
 - model başarımı: doğru sınıflandırılmış sinama kümesi örneklerinin oranı
- Veri madenciliği uygulamasında:
 - ayrıntılı değerlerini tahmin etmek: sınıflandırma
 - sürekli nitelik değerlerini tahmin etmek: öngörü



- Sınıflandırma: hangi topun hangi sepete koyulabileceği
- Öngörü: Topun ağırlığı

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

4

Gözetimli & Gözetimsiz Sınıflandırma

- Gözetimli (Supervised) sınıflandırma = sınıflandırma
 - Sınıfların sayısı ve hangi nesnenin hangi sınıfı olduğu biliniyor.
- Gözetimsiz (Unsupervised) sınıflandırma = demetleme (clustering)
 - Hangi nesnenin hangi sınıfı olduğu bilinmiyor. Genelde sınıf sayısı bilinmiyor.



<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

5

Konular

- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcılar
 - Bayes ağıları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

6

Sınıflandırma Uygulamaları

- Kredi başvurusu değerlendirme
- Kredi kartı harcamasının sahtekarlık olup olmadığına karar verme
- Hastalık teşhisisi
- Ses tanıma
- Karakter tanıma
- Gazete haberlerini konularına göre ayırma
- Kullanıcı davranışları belirleme



<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

7

Sınıflandırma için Veri Hazırlama

- Veri dönüşümü:
 - Sürekli nitelik değeri ayrık hale getirilir
 - Normalizasyon ($[-1, \dots, 1]$, $[0, \dots, 1]$)
- Veri temizleme:
 - gürültüyü azaltma
 - gereksiz nitelikleri silme

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

8

Sınıflandırma İşlemi

- Sınıflandırma işlemi üç aşamadan oluşur:
 1. Model oluşturma
 2. Model değerlendirme
 3. Modeli kullanma

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

9

Sınıflandırma İşlemi: Model Oluşturma

1. Model Oluşturma:
 - Her nesnenin sınıf etiketi olarak tanımlanan niteliğinin belirdiği bir sınıfta olduğu varsayılar
 - Model oluşturmak için kullanılan nesnelerin oluşturduğu veri kümesi öğrenme kümesi olarak tanımlanır
 - Model farklı biçimlerde ifade edilebilir
 - IF – THEN – ELSE kuralları ile
 - Karar ağaçları ile
 - Matematiksel formüller ile

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

10

Sınıflandırma İşlemi: Model Değerlendirme

2. Model Değerlendirme:
 - Modelin başarımı (doğruluğu) sınama kümesi örnekleri kullanılarak belirlenir
 - Sınıf etiketi bilinen bir sınama kümesi örneği model kullanılarak belirlenen sınıf etiketileyi karşılaştırılır
 - Modelin doğruluğu, doğru sınıflandırılmış sınama kümesi örneklerinin toplam sınama kümesi örneklerine oranı olarak belirlenir
 - Sınama kümesi model öğrenirken kullanılmaz

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

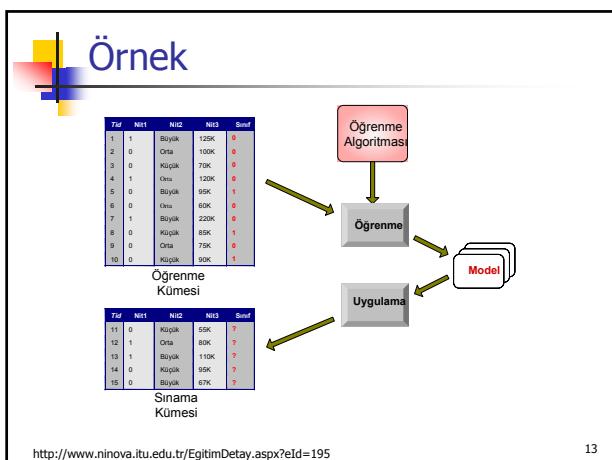
11

Sınıflandırma İşlemi: Modeli Kullanma

3. Modeli kullanma:
 - Model daha önce görülmemiş örnekleri sınıflandırmak için kullanılır
 - Örneklerin sınıf etiketlerini tahmin etme
 - Bir niteliğin değerini tahmin etme

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

12

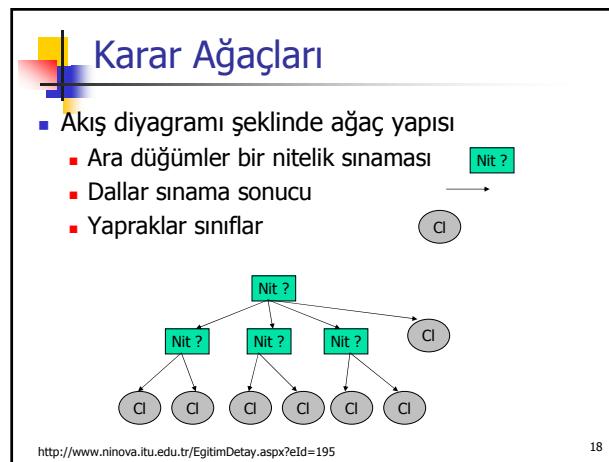


- ## Sınıflandırıcı Başarımını Değerlendirme
- Doğru sınıflandırma başarısı
 - Hız
 - modeli oluşturmak için gerekli süre
 - sınıflandırma yapmak için gerekli süre
 - Kararlı olması
 - veri kümesinde gürültülü ve eksik nitelik değerleri olduğu durumlarda da iyi sonuç vermesi
 - Ölçeklenebilirlik
 - büyük miktarda veri kümesi ile çalışabilmesi
 - Anlaşılabilir olması
 - kullanıcı tarafından yorumlanabilir olması
 - Kuralların yapısı
 - birbirine örtüşmeyen kurallar
- 14
- http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195

- ## Konular
- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
 - Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcıları
 - Bayes ağları
- 15
- http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195

- ## Sınıflandırma Yöntemleri
- Karar ağaçları (decision trees)
 - Yapay sinir ağları (artificial neural networks)
 - Bayes sınıflandırıcıları (Bayes classifier)
 - İlişki tabanlı sınıflandırıcılar (association-based classifier)
 - k-en yakın komşu yöntemi (k- nearest neighbor method)
 - Destek vektör makineleri (support vector machines)
 - Genetik algoritmalar (genetic algorithms)
 - ...
- 16
- http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195

- ## Konular
- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
 - Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcıları
 - Bayes ağları
- 17
- http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195



Örnek: Karar Ağacı

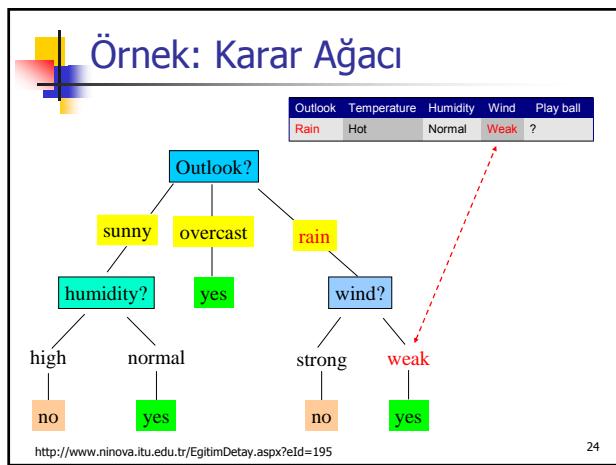
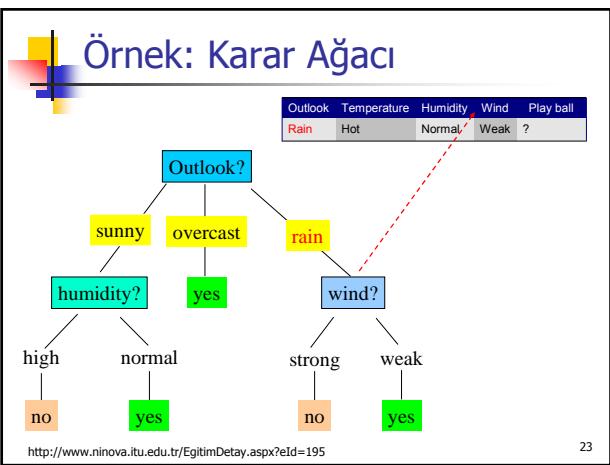
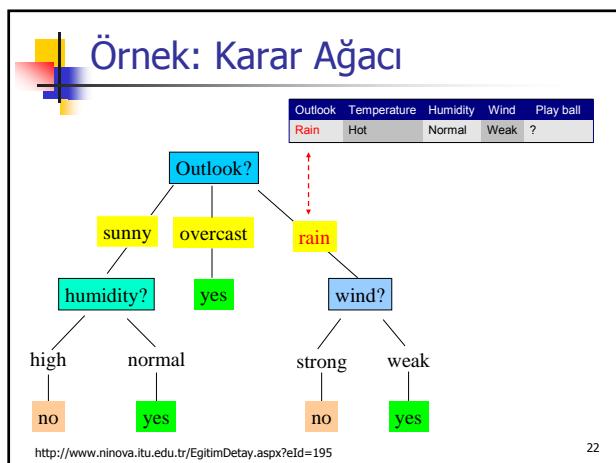
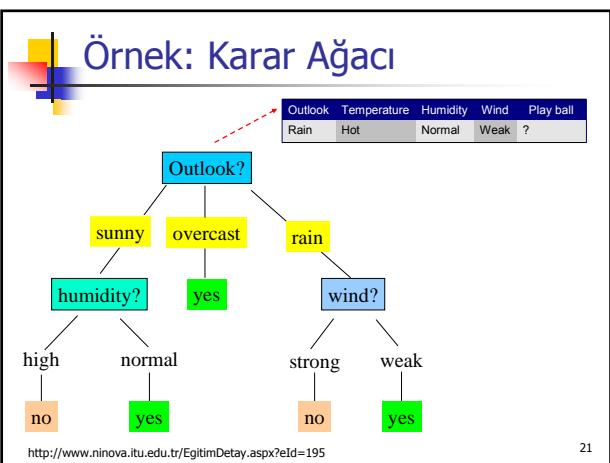
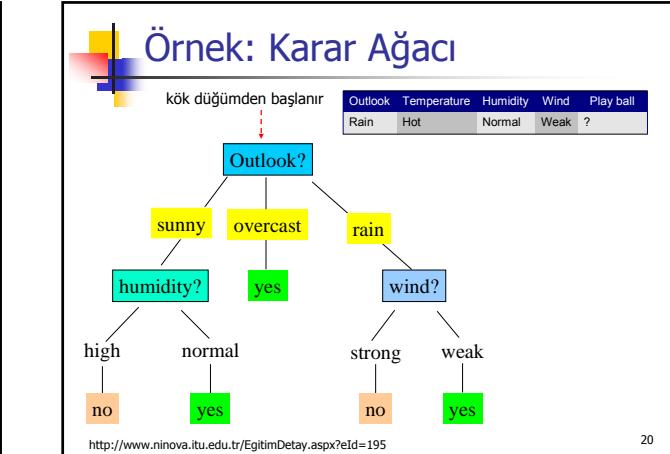
J. Ross Quinlan'ın geliştirdiği ID3 modeline uyarlanmıştır:

- hava tenis oynamaya uygun mu?

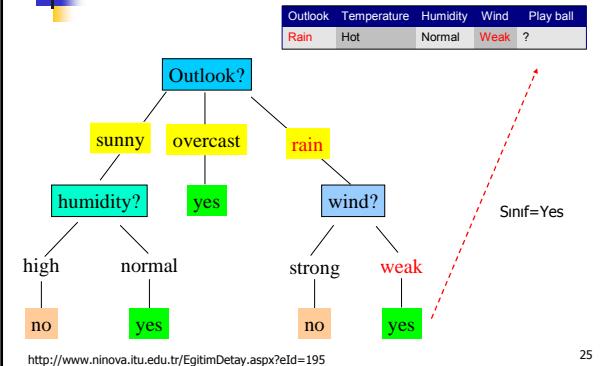
Day	Outlook	Temperature	Humidity	Wind	Play ball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Hava durumu Verisi

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>



Örnek: Karar Ağacı



<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

25

Karar Ağacı Yöntemleri

- Karar ağıacı oluşturma yöntemleri genel olarak iki aşamadan oluşur:
 - ağaç oluşturma
 - en başta bütün öğrenme kümesi örnekleri kökte
 - seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor
 - ağaç budama
 - öğrenme kümelerindeki görültülü verilerden oluşan ve sinama kümelerinde hataya neden olan dalları silme (sınıflandırma başarısını artırır)

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

26

Karar Ağacı Oluşturma

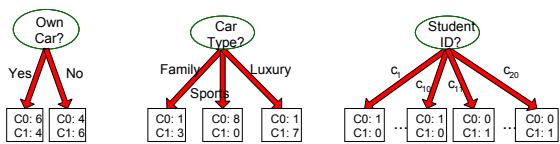
- Yinelemeli işlem
 - ağaç bütün verinin oluşturduğu tek bir düğümle başlıyor
 - CL ■ eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor
 - NIT? ■ eğer değilse örnekleri sınıflara en iyi bölecek olan nitelik seçiliyor
 - işlem sona eriyor
 - örneklerin hepsi (çoğunluğu) aynı sınıfa ait
 - örnekleri bölecek nitelik kalmamış
 - kalan niteliklerin değerini taşıyan örnek yok

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

27

Örnekleri En İyi Bölen Nitelik Hangisi?

- Bölmeden önce:
 - 10 örnek C0 sınıfında
 - 10 örnek C1 sınıfında



Hangisi daha iyi?

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

28

En iyi Bölme Nasıl Belirlenir?

- "Greedy" yaklaşım
 - çoğunlukla aynı sınıfı ait örneklerin bulunduğu düğümler tercih edilir
 - Düğümlün kalitesini ölçmek için bir yöntem

C0: 5
C1: 5

kalitesi düşük

C0: 9
C1: 1

kalitesi yüksek

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

29

En İyi Bölen Nitelik Nasıl Belirlenir?

- İyilik Fonksiyonu (Goodness Function)
 - Farklı algoritmalar farklı iyilik fonksiyonları kullanabilir:
 - bilgi kazancı (information gain): ID3, C4.5
 - bütün niteliklerin ayrık değerler aldığı varsayıiyor
 - sürekli değişkenlere uygulamak için değişiklik yapılabilir
 - gini index (IBM IntelligentMiner)
 - her nitelik ikiye bölünüyor
 - her nitelik için olası bütün ikiye bölünmeler sınanıyor

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

30



Bilgi Kazancı



Bilgi / Entropi

$\log(p)$

$H(p,1-p)$



Örnek

- S veri kümesinde 14 örnek: C0 sınıfına ait 9, C1 sınıfına ait 5 örnek.
- Entropi
$$H(p_1, p_2, \dots, p_s) = -\sum_{i=1}^s p_i \log(p_i)$$
- $$\begin{aligned} H(p_1, p_2) &= - (9/14) \text{ Log}_2 (9/14) - (5/14) \text{ Log}_2 (5/14) \\ &= 0.940 \end{aligned}$$

Bilgi Kazancı (ID3 / C4.5)

Örnek

- Bilgi kazancına göre nitelik seçme
toplam örnek sayısı = 14, iki sınıfı ayrılmış
 $s_1=9$ (yes), $s_2=5$ (no)

Entropy(S) = $-(9/14) \log_2(9/14) - (5/14) \log_2(5/14) = 0.940$

wind için: weak=8, strong=6

weak: no=2, yes=6

strong: no=3, yes=3

Entropy(S_{weak}) = $-(6/8)*\log_2(6/8) - (2/8)*\log_2(2/8) = 0.811$

Entropy(S_{strong}) = $-(3/6)*\log_2(3/6) - (3/6)*\log_2(3/6) = 1.00$

Entropy_{wind}(S) = $(8/14)*0.811 + (6/14)*1.00$

Gain(wind)=0.940 - (8/14)*0.811 - (6/14)*1.00

$$Gain(S, A) = Entropy(S) - \sum_{s \in \text{values}(A)} \frac{|S_s|}{|S|} Entropy(S_s)$$

$$H(p_1, p_2, \dots, p_n) = -\sum_{i=1}^n p_i \log(p_i)$$

ID	Outlook	Temperature	Humidity	Wind	Play
101	Sunny	Hot	High	Weak	No
102	Overcast	Hot	High	Strong	Yes
103	Rain	Mild	High	Weak	Yes
104	Rain	Cool	Normal	Weak	Yes
105	Rain	Cool	Normal	Strong	No
106	Overcast	Cool	Normal	Strong	Yes
107	Sunny	Mild	High	Weak	No
108	Sunny	Cool	Normal	Weak	No
109	Rain	Mild	Normal	Weak	Yes
110	Sunny	Mild	Normal	Strong	Yes
111	Overcast	Mild	High	Strong	Yes
112	Overcast	Hot	Normal	Weak	Yes
113	Rain	Mild	High	Strong	No

Gain(Outlook) = 0.246

Gain(Humidity) = 0.151

Gain(wind)=0.048

Gain(Temperature) = 0.029

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?Id=195>

36

Örnek

$S_{\text{sunny}} = \{D_1, D_2, D_8, D_9, D_{11}\}$, $\text{Entropy}(S_{\text{sunny}}) = 0.970$
humidity için: high=3, normal=2
high: no=3, yes=0
normal: no=0, yes=2
 $\text{Entropy}(S_{\text{high}}) = 0$
 $\text{Entropy}(S_{\text{normal}}) = 0$
 $\text{Gain}(S_{\text{sunny}}, \text{Humidity}) = 0.970 - (3/5)0.0 - (3/5)0.0 = 0.970$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Gini Index (IBM IntelligentMiner)

- Veri kümesi S içinde n sınıf varsa ve p_j C_j sınıfının olasılığı ise

$$gini(S) = 1 - \sum_{j=1}^n p_j^2$$

- Eğer veri kümesi S_1 ve S_2 altkümelere bölündüğse ve her altkümede sırasıyla N_1 ve N_2 örnek varsa:

$$gini_{\text{split}}(S) = \frac{N_1}{N} gini(S_1) + \frac{N_2}{N} gini(S_2)$$

- Gini Index değeri en küçük olan nitelik seçilir.

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

38

Örnek

$$GINI(S) = 1 - \sum_j [p_j]^2$$

C1	0
C2	6

$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$
 $\text{Gini} = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$P(C1) = 1/6 \quad P(C2) = 5/6$
 $\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$P(C1) = 2/6 \quad P(C2) = 4/6$
 $\text{Gini} = 1 - (2/6)^2 - (4/6)^2 = 0.444$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Sürekli Nitelikleri Bölme

- Bölmeleme:
 - Statik: En başta bölmelenir
 - Bölmeler eşit genişlik, eşit derinlik veya demetleme yöntemi ile bulunur.
 - Dinamik:
 - Sürekli nitelik A sıralanır. Birbirini izleyen ancak sınıf etiketi farklı olan nitelik değerleri bulunur. En fazla kazanç sağlayan bölme seçilir.

Temperature	40	48	60	72	80	90
Play tennis	No	No	Yes	Yes	Yes	No

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

40

Ağaç Oluşturmada Temel Yaklaşımlar

- Bölme kriteri:
 - ağacın bir düğümünde karşılaştırma yapılacak niteliğin seçilmesi
 - farklı algoritmalar farklı iyilik fonksiyonları kullanabilir: bilgi kazancı, gini index, ...
- Dallanma kriteri:
 - bir örneğin hangi dala ait olduğunu belirleme
 - ikiye dallanma (gini index), çoklu dallanma (bilgi kazancı)
- Durma kararı:
 - dallanma işleminin devam edip etmeyeceğine karar verme
- Etiketleme kuralı:
 - yaprak düğüm en çok örneği olan sınıfı etiketleniyor

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

41

Ağaç Oluşturma:

- Parçala ve çöz (divide and conquer)
 - kökten yapraklara
 - düğümü dallara ayı
 - 'Greedy' algoritma
 - her adında en iyi çözümü bul: her düğümde dallanmak için en iyi niteliği bul
 - her dal için algoritmayı uygula

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

42

Örnek Algoritma: ID3

- Bütün nitelikler ayrık
- Bir düğüm oluştur N:
 - Eğer örneklerin hepsi C sınıfına ait ise, N düşümü C etiketli yaprak
 - Eğer karşılaştırma yapılacak nitelik yoksa N düşümü en çok örnegi olan sınıf
- En büyük bilgi kazancı olan niteliği bölmek için seç
 - N_i yi seçilen nitelik ile etiketle
 - niteliğin her A_i değeri için bir dal oluştur
 - S_i , örneklerin hepsinin A_i değeri aldığı dal
 - S_i boş \rightarrow bir yaprak oluşturup en çok örnegi olan sınıfı etiketle
 - S_i boş değil \rightarrow algoritmayı S_i düşümü üzerinde yinele
- Yaprak düğümlere kadar

Ayrıntılı bilgiler: http://dms.irb.hr/tutorial/tut_dtrees.php

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

43

Örnek Algoritma: C4.5

- Kökten yapraklara doğru ağaç oluşturma
- Bilgi kazancı yöntemini kullanıyor
- Bütün veri kümelerini bellekte tutuyor
 - Büyük veri kümeleri için uygun değil

<http://www.rulequest.com/Personal/c4.5r8.tar.gz>

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

44

Karar Ağacı Kullanarak Sınıflandırma

- Doğrudan
 - sınıflandırmak istenilen örnegin nitelikleri ağaç boyunca sinanır
 - ulaşılan yaprağın etiketi sınıf bilgisini verir
- Doyayı
 - karar ağacı sınıflandırma kurallarına dönüştürülür
 - Kökten yaprakların herbirene giden yollar için ayrı bir kural oluşturular.
 - IF-THEN şeklinde kuralları insanlar daha kolay anlıyor
 - Örnek: IF Outlook="sunny" AND humidity="normal" THEN play tennis

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

45

Karar Ağacı Kullanarak Sınıflandırma

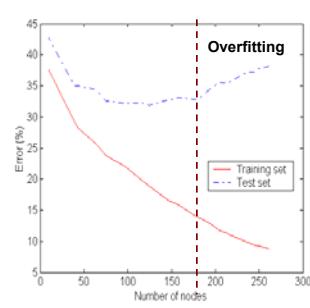
- Avantajları:
 - Karar ağacı oluşturmak zahmetsziz
 - Küçük ağaçları yorumlamak kolay
 - Anlaşılabılır kurallar oluşturulabilir
 - Sürekli ve ayrik nitelik değerleri için kullanılabilir
- Dezavantajları:
 - Sürekli nitelik değerlerini tahmin etmekte çok başarılı değil
 - Sınıf sayısı fazla ve öğrenme kümlesi örnekleri sayısı az olduğunda model oluşturma çok başarılı değil
 - Zaman ve yer karmaşaklılığı öğrenme kümlesi örnekleri sayısına (q), nitelik sayısına (h) ve oluşan ağaçın yapısına bağlı.
 - Ağaç oluşturma karmaşaklılığı fazla, ağaç будама karmaşaklılığı fazla
 - ağaç oluşturmak için zaman karmaşaklılığı: $O(h \log q)$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

46

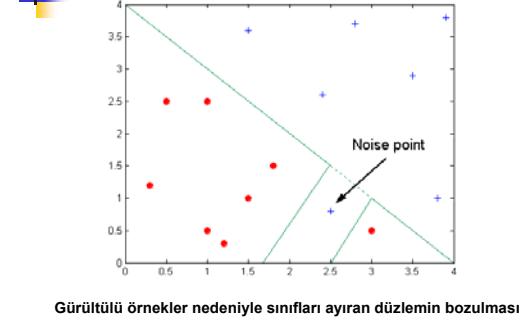
Karar Ağaçlarında Aşırı Öğrenme

- Öğrenme kümelerindeki örneklerin azlığı veya gürültülü olması
- Aşırı öğrenmeye engelleyen iki yaklaşım
 - işlemi erken sona erdirme
 - işlemi sona erdirmek için eşik değeri belirlemek gerekiyor
 - karar ağacı oluşturuktan sonra ağaç küçültme



47

Aşırı Öğrenme: Gürültülü Örnekler



<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

48

Aşırı Öğrenme: Yetersiz Öğrenme Kümesi

Öğrenme kümelerindeki örnek sayısının yetersiz olması nedeniyle sınıma kümelerindeki örneklerin yanlış sınıflandırılması

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

49

Aşırı Öğrenme

- Gereğinden fazla karmaşık karar ağaçları aşırı öğrenmeye neden oluyor.
- Karar ağacının yeni örnekler üzerindeki başarımını tahmin etmek için öğrenme kümeleri yeterli olmuyor.
- Hatayı tahmin etmek için farklı yöntemler gereklidir.

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

50

Genel Hatayı Tahmin Etme

- Yerine koyma (Resubstitution) Hatası: öğrenme kümesi kullanılarak hesaplanan hata ($\sum e(t)$)
- Genel (Generalization) hata: sınıma kümeleri kullanılarak hesaplanan hata ($\sum e'(t)$)
- Genel hatayı tahmin etme yöntemleri:
 - İyimser yaklaşım: $e'(t) = e(t)$
 - Kötümer yaklaşım:
 - Her yaprak düğüm için: $e'(t) = e(t)+0.5$
 - Toplam hata: $e'(T) = e(T) + N \times 0.5$ (N : yaprak düğüm sayısı)
 - 30 yaprak düğümü olan bir karar ağıacı, 1000 öğrenme kümeleri örneğinden 10 örneği yanlış sınıflandırırsa
 - Yerine koyma hatası: $10/1000 = \%1$
 - Genel hata: $(10 + 30 \times 0.5)/1000 = \% 2.5$
 - Ağacı budama: Genel hatayı tahmin etmek için geçerleme kümesi kullanılır

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

51

Occam's Razor

- Genel hatası aynı olan iki modelden karmaşıklığı daha az olan seçilmeli
- Karmaşık modellerin veri içindeki gürültüyü öğrenme ihtimaleri daha fazla

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

52

Karar Ağacı Boyutunu Belirleme

- Veri kümesi öğrenme ve sınıma kümeleri olarak ayrılır
- Çapraz geçerleme kullanılır.
- Veri kümelerinin tümü ağıacı oluşturmak için kullanılır
 - İstatistiksel bir test ile (chi-square) düğüm eklemenin ya da ağıacı küçültmenin katkısı sınanır
 - MDL (Minimum Description Length) yöntemi kullanılır: kodlama en aza indirildiğinde ağıacın büyümesi durdurulur

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

53

Karar Ağacı: Erken Durdurma

- Karar ağacını tam oluşturmadan işlemi bitirme
 - İşlemi sona erdirme için ek kurallar
 - Örneklerin sayısı kullanıcı tarafından belirlenen bir eşik değerinden daha az ise
 - Örneklerin sınıf dağılımı niteliklerden bağımsız ise (chi-square testi ile belirlenebilir)
 - Ağaca yeni bir düğüm ekleyince iyilik fonksiyonu yeterince artmıyorsa

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

54

Konular

- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - **Yapay sinir ağları**
 - Bayes sınıflandırıcılar
 - Bayes ağaçları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

55

Yapay Sinir Ağları ile Sınıflandırma

- İnsan beynindeki sinir hücrelerinin işlevini modelleyen bir yapı
- Birbiri ile bağlantılı katmanlardan oluşur.
 - katmanlar hücrelerden oluşur
- Katmanlar arasında iletişim
- İleti katmanlar arasındaki bağın ağırlığına ve her hücrenin değerine bağlı olarak değişimelidir

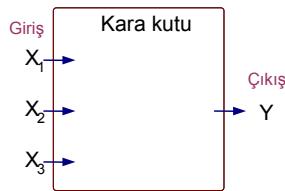


<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

56

Örnek:

X ₁	X ₂	X ₃	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	1
0	1	0	0
0	1	1	1
0	0	0	0

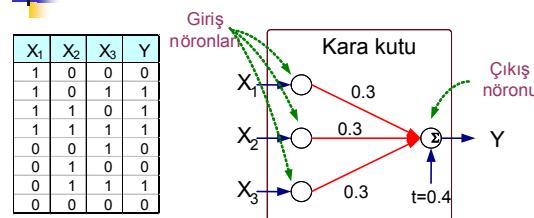


- En az iki giriş 1 ise çıkış 1, diğer durumlarda çıkış 0

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

57

Örnek



$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

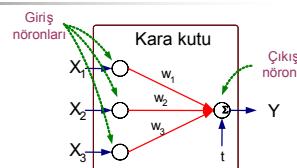
$$I(z) = \begin{cases} 1 & \text{eğer } z > 0 \\ 0 & \text{diğer} \end{cases}$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

58

Yapay Sinir Ağları

- Birbiri ile bağlantılı nöronlar ve ağırlıklar
- Çıkış nöronu kendisine gelen girişleri ağırlıklı olarak topluyor
- Çıkış nöronu bir eşik değeri ile karşılaştırılıyor



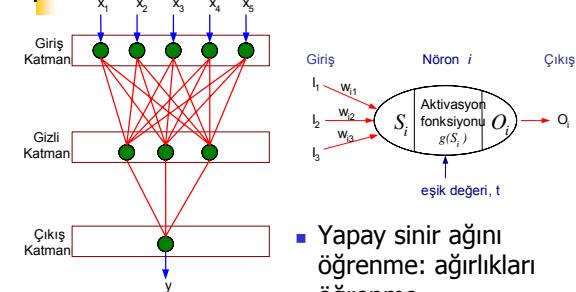
$$Y = I(\sum_i w_i X_i - t)$$

$$Y = \text{sign}(\sum_i w_i X_i - t)$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

59

Çok Katmanlı



- Yapay sinir ağını öğrenme: ağırlıkları öğrenme

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

60

Yapay Sinir Ağı ile Öğrenme

- Yapay sinir ağı oluşturma
 - giriş verisini modelleme
 - gizli katman sayısını, gizli katmanlardaki nöron sayısını belirleme
- Yapay sinir ağını eğitme
- Sinir ağını küçültme
- Sonucu yorumlama

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

61

Yapay Sinir Ağını Oluşturma

- Giriş nöron sayısı
 - Öğrenme kümelerindeki verilerin nitelik sayısı
- Gizli nöron sayısı
 - öğrenme sırasında ayarlanır
- Çıkış nöron sayısı
 - sınıf sayısı

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

62

Yapay Sinir Ağını Eğitme

- Amaç: Veri kümelerindeki örneklerin hepsini doğru sınıflandıracak ağırlıkları belirlemek
 - ağırlıklara rasgele değerler ata
 - öğrenme kümelerindeki giriş değerlerini teker teker sinir ağına uygula
 - çıkış hesapla
 - hata değerini hesapla $E = \sum_i [Y_i - f(w_i, X_i)]^2$
 - ağırlıkları hata fonksiyonunu enküçültucek şekilde düzelt

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

63

Yapay Sinir Ağını Küçültme

- Tam bağlı ağın anlaşılması çok güç
- n giriş nöron, h gizli nöron, m çıkış nöronu
 $h(m+n)$ ağırlık
- Küçültme: ağırlıklardan bazıları sınıflandırma sonucunu etkilemeyecek şekilde silinir

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

64

Yapay Sinir Ağları

- Avantajları
 - doğru sınıflandırma oranı genelde yüksek
 - kararlı – öğrenme kümelerinde hata olduğu durumda da çalışıyor
 - çıkış ayrık, sürekli ya da ayrık veya sürekli değişkenlerden oluşan bir vektör olabilir
- Dezavantajları
 - öğrenme süresi uzun
 - öğrenilen fonksiyonun anlaşılması zor

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

65

Konular

- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcılar
 - Bayes ağları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

66

Bayes (İstatistiksel) Modelleme

- Bayes teoremini kullanan istatistiksel sınıflandırıcı
- Örneklerin hangi sınıfı hangi olasılıkla ait oldukları
- Naïve Bayes sınıflandırıcı
 - niteliklerin hepsi aynı derecede önemli
 - nitelikler birbirinden bağımsız
 - bir niteliğin değeri başka bir nitelik değeri hakkında bilgi içermiyor
 - sınıflandırma ve öğrenme problemleri

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

67

Bayes Teoremi

- X sınıflandırılacak örnek. Hipotez h , X örneğinin C sınıfına ait olduğu
- h hipotezinin sonrasal olasılığı (*posteriori probability*)

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$
- MAP (maximum posteriori) hipotez

$$h_{MAP} = \arg \max_{h \in H} P(h|D) = \arg \max_{h \in H} P(D|h)P(h)$$
- Çok sayıda olasılığı önceden kestirmek gerekiyor

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

68

Örnek

- $H + \bigcirc \xrightarrow{H} C$
 - $P(H) = P(\text{apple}) \quad P(X) = P(H + \bigcirc)$
 - $P(X|H) = P(\text{apple}) \text{ ise } H + \bigcirc$
- $$P(h|X) = \frac{P(X|h)P(h)}{P(X)}$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

69

Naïve Bayes Sınıflandırıcı

- $X = (X_1, X_2, \dots, X_n)$ örneğinin C sınıfında olma olasılığı ($P(C|X)$) nedir?
- $\frac{P(X|C_i)P(C_i)}{P(X)}$ değerini enbüyükte
 $\rightarrow P(X|C_i)P(C_i)$ değerini enbüyükte
- $P(C_i) = |S_i| / |S|$, S_i : C_i sınıfına ait örneklerin sayısı
- $P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$, $P(x_k|C_i) = s_{ik} / s_i$
- Hesaplama maliyetini azaltıyor, sadece sınıf dağılımları hesaplanıyor
- Naïve: nitelikler bağımsız

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

70

Hava Durumu Verisi için Olasılıklar

Outlook	Temperature		Humidity		Windy		Play	
	Yes	No	Yes	No	Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4
Overcast	4	0	Mild	4	2	Normal	6	1
Rainy	3	2	Cool	3	1			
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5
Rainy	3/9	2/5	Cool	3/9	1/5			

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

1

Hava Durumu Verisi için Olasılıklar

Outlook	Temperature		Humidity		Windy		Play	
	Yes	No	Yes	No	Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4
Overcast	4	0	Mild	4	2	Normal	6	1
Rainy	3	2	Cool	3	1			
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5
Rainy	3/9	2/5	Cool	3/9	1/5			

Yeni veri

$$P(C_i | X) = P(X | C_i) \times P(C_i) = \prod_{k=1}^n P(x_k | C_i) \times P(C_i)$$

$$\begin{aligned} P(\text{"yes"}|X) &= 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053 \\ P(\text{"no"}|X) &= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206 \end{aligned}$$

Normalize edilmiş olasılıklar:

$$P(\text{"yes"}) = 0.0053 / (0.0053 + 0.0206) = 0.205$$

$$P(\text{"no"}) = 0.0206 / (0.0053 + 0.0206) = 0.795$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

72

Sürekli Veriler için Olasılık

- Verinin normal dağılımdan geldiği varsayılmıyor.
- Her sınıf-nitelik çifti için bir olasılık hesaplanıyor.

$$P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_j)^2}{2\sigma_{ij}^2}}$$

- Gelir için sınıf=-1
 - ortalama=110
 - varyans=2975

$$P(Gelir = 120 | -1) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Örnek

X=(Geri Ödeme="Hayır", Medeni Durum="Evli", Gelir=120k)

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolan dinci
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Bosanmis	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Bosanmis	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

P(Geri Ödeme="Evet"|-1)=3/7
 P(Geri Ödeme="Hayır"|-1)=4/7
 P(Geri Ödeme="Evet")=0
 P(Geri Ödeme="Hayır")=1
 P(Medeni Durum="Evli"|-1)=4/7
 P(Medeni Durum="Bekar"|-1)=2/7
 P(Medeni Durum="Boşanmış"|-1)=1/7
 P(Medeni Durum="Evli")=0
 P(Medeni Durum="Bekar")=2/3
 P(Medeni Durum="Boşanmış")=1/3

Gelir:
 Sınıf=-1
 ortalama=110
 varyans=2975
 Sınıf=1
 ortalama=90
 varyans=25

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Örnek

X=(Geri Ödeme="Hayır", Medeni Durum="Evli", Gelir=120k)

P(Geri Ödeme="Evet"|-1)=3/7
 P(Geri Ödeme="Hayır"|-1)=4/7
 P(Geri Ödeme="Evet")=0
 P(Geri Ödeme="Hayır")=1
 P(Medeni Durum="Evli"|-1)=4/7
 P(Medeni Durum="Bekar"|-1)=2/7
 P(Medeni Durum="Boşanmış"|-1)=1/7
 P(Medeni Durum="Evli")=0
 P(Medeni Durum="Bekar")=2/3
 P(Medeni Durum="Boşanmış")=1/3

Gelir:
 Sınıf=-1
 ortalama=110
 varyans=2975
 Sınıf=1
 ortalama=90
 varyans=25

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Olasılığın Sıfır Olması

- Her sınıfta bir niteliğin her değeri olmazsa
 - koşullu olasılıklardan biri 0
 - o sınıf'a ait olma olasılığı 0
- Olasılıklar

Original : $P(A_i | C) = \frac{N_{ic}}{N_c}$ k: nitelik sayısı
 Toplamları 1 olmak zorunda

Laplace : $P(A_i | C) = \frac{N_{ic} + 1}{N_c + k}$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Bayes Sınıflandırıcılar

- Avantajları:
 - gerçeklemesi kolay
 - çoğu durumda iyi sonuçlar
- Dezavantajları
 - varsayımlı: sınıf bilgisi verildiğinde nitelikler bağımsız
 - gerçek hayatı değişkenler birbirine bağımlı
 - değişkenler arası ilişki modellenemiyor
- Çözüm:
 - Bayes ağıları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

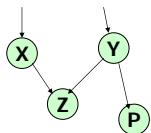
Konular

- Sınıflandırma işlemi
 - Sınıflandırma tanımı
 - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
 - Karar ağaçları
 - Yapay sinir ağları
 - Bayes sınıflandırıcılar
 - Bayes ağıları

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

Bayes Ağları

- Netliklerin altkümesinin birbiri ile bağımsız olduğunu varsayıyor
- Yönlü çevrimsiz çizge (directed acyclic graph) ve koşullu olasılık tablolarından oluşur
- Her değişken A için bir tablo var
 - netliğin ebeveynlerine olan koşullu olasılıkları

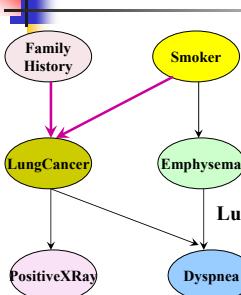


- düğümler: rasgele değişkenler
- ayritlar: olasılıklı bağılilik
- X ve Y, Z değişkeninin ebeveyini
- Y, P değişkeninin ebeveyini
- Z ve P arasında bağ yok

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

79

Örnek



(FH, S) (FH, ~S)(~FH, S) (~FH, ~S)

LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

LungCancer için koşullu olasılık tablosu

Bayes Ağları

$$P(LC="yes" | FH="yes", S="yes") = 0.8$$
$$P(LC="no" | FH="no", S="no") = 0.9$$

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

80

Bayes Ağlarının Eğitilmesi

- Ağ yapısı ve tüm değişkenler biliniyorsa koşullu olasılıklar hesaplanır
- Ağ yapısı belli ancak bazı değişkenler eksik ise yinelemeli öğrenme uygulanır
 - gradient descent algoritması
- D. Heckerman. [A Tutorial on Learning with Bayesian Networks](#). In *Learning in Graphical Models*, M. Jordan, ed.. MIT Press, Cambridge, MA, 1999. Also appears as Technical Report MSR-TR-95-06, Microsoft Research, March, 1995.

<http://www.ninova.itu.edu.tr/EgitimDetay.aspx?eId=195>

81