BİM448-VERİ MADENCİLİĞİ VE BİLGİ KEŞFİ
YILİÇİ SINAVI

Yrd.Doç.Dr. Songül ALBAYRAK — 5 Mayıs 2009

Sınav Süresi: 90 dakika

Öğrencinin Adı ve Soyadı: Öğrenci No:

SORULAR

1-[35puan] Aşağıda verilen 8 örnek ve 3 özellikten oluşan küçük veriseti için pozitif ve negatif sınıfları belirlemek için bir karar ağacı oluşturunuz. Karar ağacını oluşturmak için entropy ve bilgi kazancını hesaplayınız.

T veriseti

Hair	Weight	Lotion	Result
blonde	light	no	sunburned (positive)
blonde	average	yes	none (negative)
brown	average	yes	none
blonde	average	no	sunburned
red	heavy	no	sunburned
brown	heavy	no	none
brown	heavy	no	none
blonde	light	yes	none

$$Entropy(T) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0,95 \quad \{3P, 5N\}$$

Blonde = 4 {2+, 2-}
Brown = 3 {3-}
Red = 1 {1+}

$$Entropy_{hair}(T) = \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{3}{8}\left(-\frac{3}{3}\log_2\frac{3}{3} - 0\right) + \frac{1}{8}\left(-1\log_2 1\right) = 0,5$$

Light = 2 {1+, 1-}
Average = 3 {2+, 1-}
Heavy = 3 {1+, 2-}

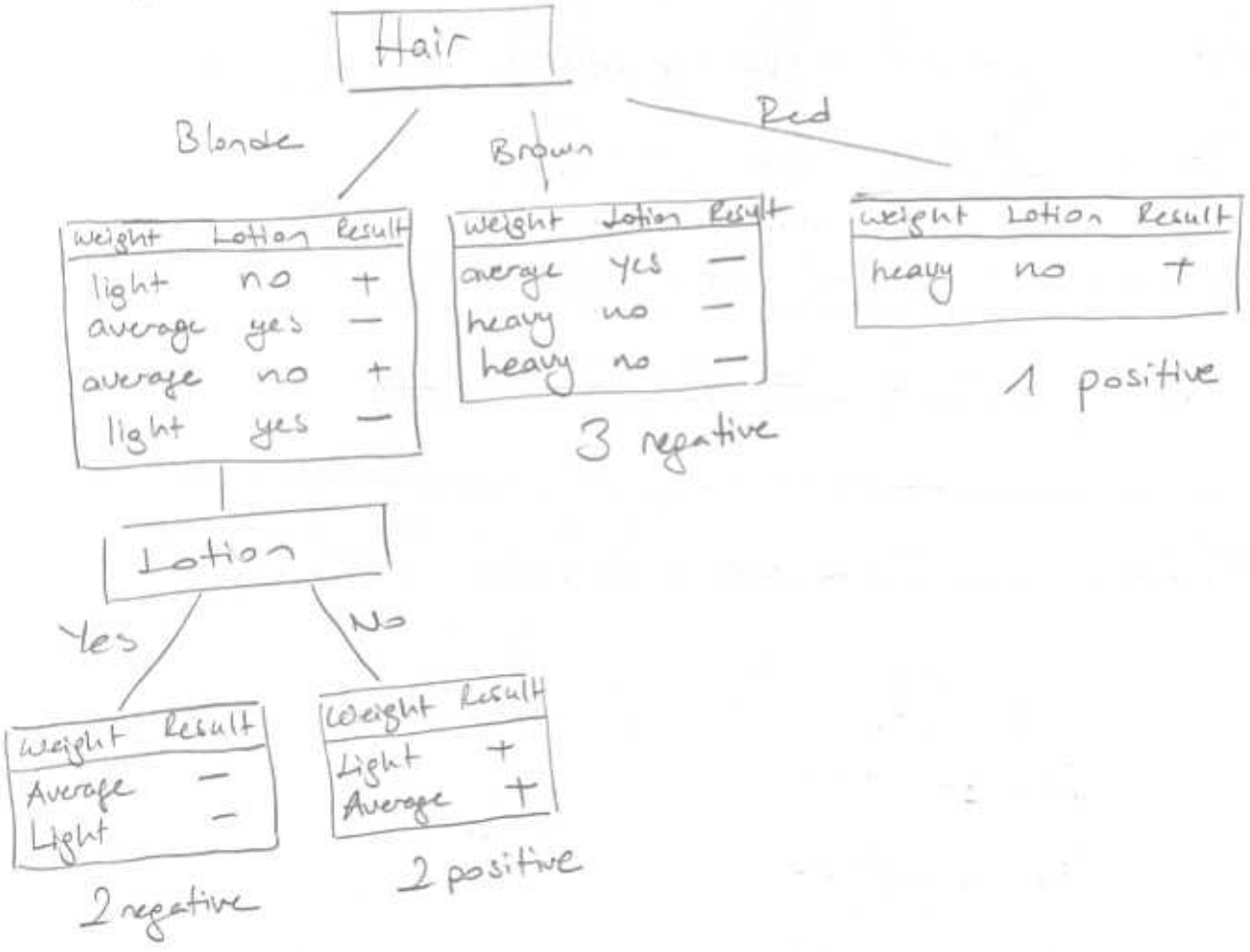
$$Entropy_{weight}(T) = \frac{2}{8}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{8}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) + \frac{3}{8}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0,94$$

No = 5 {3+, 2-}
Yes = 3 {3-}

$$Entropy_{lotion}(T) = \frac{5}{8}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \frac{3}{8}\left(-\frac{3}{3}\log_2\frac{3}{3} - 0\right) = 0,61$$

Özellik	kazanç
Hair	$Gain_{hair} = Entropy(\tau) - Entropy_{hair}(\tau) = 0,95 - 0.5 = 0,45$
Weight	$Gain_{weight} = 0,95 - 0,94 = 0,01$
Lotion	$Gain_{lotion} = 0,95 - 0,61 = 0,34$

Saç rengi maksimum kazanç verdiği için karar ağacının köküne yerleştirilir.



weight	lotion	result
light	no	+
average	yes	−
average	no	+
light	yes	−

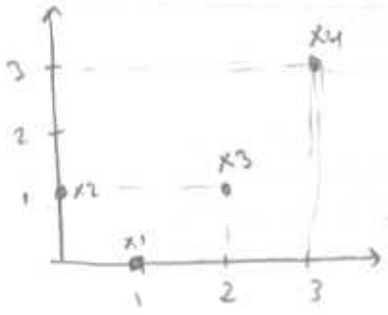
weight	lotion	result
average	yes	−
heavy	no	−
heavy	no	−

weight	lotion	result
heavy	no	+

weight	result
Average	−
Light	−

weight	result
Light	+
Average	+

2-[30 puan] Verisetinde iki boyutlu uzayda verilmiş 4 adet örnek bulunmaktadır. X1= {1,0}, X2={0,1}, X3={2,1}, X4={3,3}
K-ortalamalı kümeleme yöntemine göre verisetini iki kümeye bölmek için başlangıç küme merkezleri X1 ve X4 olarak seçilmiş olsun. Algoritmanın ilk iki adımı için küme elemanlarını ve yeni küme merkezlerini hesaplayınız. Hesaplamada Eucledean mesafesi kullanınız.



$C_1 = (1,0)$ ve $C_2 = (3,3)$ ise

X_1 için $d(X_1, C_1) = 0$ $\quad d(X_1, C_2) = \sqrt{1}$

X_2 için $d(X_2, C_1) = \sqrt{2}$ $\quad d(X_2, C_2) = \sqrt{13}$

X_3 için $d(X_3, C_1) = \sqrt{2}$ $\quad d(X_3, C_2) = \sqrt{5}$

X_4 için $d(X_4, C_1) = \sqrt{13}$ $\quad d(X_4, C_2) = 0$

C_1 kümesinin üyeleri $\{X_1, X_2, X_3\} \Rightarrow$ yeni küme merkezi $C_1 = (1, \frac{2}{3})$

C_2 kümesinin üyeleri $\{X_4\} \Rightarrow$ yeni küme merkezi $C_2 = (3,3)$

$C_1 = (1, \frac{2}{3})$ $\quad C_2 = (3,3)$

X_1 için $d(X_1, C_1) = 2/3$ $\quad d(X_1, C_2) = \sqrt{13}$

X_2 için $d(X_2, C_1) = \sqrt{10/9}$ $\quad d(X_2, C_2) = \sqrt{13}$

X_3 için $d(X_3, C_1) = \sqrt{10/9}$ $\quad d(X_3, C_2) = \sqrt{5}$

X_4 için $d(X_4, C_1) = \sqrt{85/9}$ $\quad d(X_4, C_2) = 0$

C_1 kümesinin üyeleri $\{X_1, X_2, X_3\} \Rightarrow C_1 = (1\ 2/3)$

C_2 kümesinin üyeleri $\{X_4\} \Rightarrow C_2 = (3,3)$

– İki iterasyonda küme merkezleri değişmediği için iterasyon sonlanır.

3-[35 puan] Aşağıdaki tabloda verilen 10 örnek ve 3 özellikten oluşan verisetini eğitim veriseti olarak kullanarak Naive-Bayes sınıflayıcı modeli geliştirilecektir. Bu verisetinde 10 otomobilin rengi, modeli ve üretilidiği yer(yerli veya ithal) özelliklerine göre çalınma riski(var veya yok şeklinde) belirtilmiştir. Hangi sınıfa ait olduğu bilinmeyen (Red Domestic SUV) otomobilin çalınma riskini Naive Bayes sınıflayıcı kullanarak belirleyiniz.

Example No:	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

5 YES
5 NO

$$P(R,D,SUV|Yes) = P(R|Yes)\,P(D|Yes)\,P(SUV|Yes) \times P(Yes)$$

$$= \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{5}{10} = \frac{6}{250}$$

$$P(R,D,SUV|No) = P(R|No)\,P(D|No)\,P(SUV|No) \times P(No)$$

$$= \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{5}{10} = \frac{18}{250}$$

NO > YES

STOLEN = NO