C4.5 Algoritması

ID3 algoritmasının nümerik özellik içeren veriye uygulanabilen şeklidir. ID3'ten tek farkı nümerik özelliklerin kategorik hale getirilebilmesini sağlayan bir eşikleme yöntemini içermesidir. Temel mantık nümerik özellik vektöründeki tüm değerler ikili olarak ele alınarak ortalamaları eşik olarak denenir. Hangi eşik değeriyle bilgi kazanımı en iyi ise o değer seçilir. Seçilen eşiğe göre özellik vektörü kategorize edilir ve ID3 uygulanır.

C4.5 Algoritması - Örnek

- Bir otomobil firmasının ürettiği iki farklı model için müşterilerine yapılan memnuniyet araştırması sonuçları aşağıdaki gibidir.

Model	Cinsiyet	Yaş	Memnun
X5	ERKEK	21	HAYIR
X3	KADIN	19	EVET
X5	ERKEK	22	HAYIR
X3	ERKEK	21	EVET
X3	ERKEK	30	EVET
X3	KADIN	60	HAYIR
X3	KADIN	45	HAYIR
X3	ERKEK	55	HAYIR

C4.5 Algoritması - Örnek

S=[3+,5-]

$$\text{Entropi} = -\text{memnun}_{evet}\log_2\text{memnun}_{evet} - \text{memnun}_{hayır}\log_2 \text{memnun}_{hayır}$$

$$= -\ 3/8 \log_2 3/8 - 5/8 \log_2 5/8$$

$$= 0.954$$

C4.5 Algoritması - Örnek

Model	Memnun
X5	HAYIR
X3	EVET
X5	HAYIR
X3	EVET
X3	EVET
X3	HAYIR
X3	HAYIR
X3	HAYIR

E = 0.954

$E(Model_{X5}) = -\ 2/2 \log_2 2/2$

$= 0$

$E(Model_{X3}) = -\ 3/6 \log_2 3/6 - 3/6 \log_2 3/6$

$= 1$

$Gain(S,Model) = 0.954 - [2/8*E(Model_{X5}) + 6/8*E(Model_{X3})]$

$= 0.954 - 0.75$

$= 0.204$

C4.5 Algoritması - Örnek

Cinsiyet	Memnun
ERKEK	HAYIR
KADIN	EVET
ERKEK	HAYIR
ERKEK	EVET
ERKEK	EVET
KADIN	HAYIR
KADIN	HAYIR
ERKEK	HAYIR

E = 0.954

$E(Cinsiyet_{Erkek}) = -\ 2/5 \log_2 2/5 - 3/5 \log_2 3/5$

$= 0.971$

$E(Cinsiyet_{Kadın}) = -\ 1/3 \log_2 1/3 - 2/3 \log_2 2/3$

$= 0.918$

$Gain(S,Cinsiyet) = 0.954 - [5/8*E(Cinsiyet_{Erkek}) + 3/8*E(Cinsiyet_{Kadın})]$

$= 0.954 - 0.951$

$= 0.003$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

$a_1, a_2, \ldots, a_n$

19 21 22 30 45 55 60

20 21 26 37 50 57

$$\frac{a_i + a_{i+1}}{2}$$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 20

E = 0.954

$E(Yaş_{<=20}) = -\ 1/1 \log_2 1/1$

$= 0$

$E(Yaş_{>20}) = -\ 2/7 \log_2 2/7 - 5/7 \log_2 5/7$

$= 0.863$

$Gain(S,Yaş) = 0.954 - [1/8*E(Yaş_{<=20}) + 7/8*E(Yaş_{>20})]$

$= 0.954 - 0.754$

$= 0.2$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 21

E = 0.954

$E(Yaş_{<=21}) = -\ 2/3 \log_2 2/3 - 1/3 \log_2 1/3$

$= 0.918$

$E(Yaş_{>21}) = -\ 1/5 \log_2 1/5 - 4/5 \log_2 4/5$

$= 0.722$

$Gain(S,Yaş) = 0.954 - [3/8*E(Yaş_{<=21}) + 5/8*E(Yaş_{>21})]$

$= 0.954 - 0.796$

$= 0.158$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 26

E = 0.954

$E(Yaş_{<=26}) = -\ 2/4 \log_2 2/4 - 2/4 \log_2 2/4$

$= 1$

$E(Yaş_{>26}) = -\ 1/4 \log_2 1/4 - 3/4 \log_2 3/4$

$= 0.811$

$Gain(S,Yaş) = 0.954 - [4/8*E(Yaş_{<=26}) + 4/8*E(Yaş_{>26})]$

$= 0.954 - 0.905$

$= 0.049$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 37

E = 0.954

$E(Yaş_{<=37}) = -\ 3/5 \log_2 3/5 - 2/5 \log_2 2/5$

$= 0.971$

$E(Yaş_{>37}) = -\ 3/3 \log_2 3/3$

$= 0$

$Gain(S,Yaş) = 0.954 - [5/8*E(Yaş_{<=37}) + 3/8*E(Yaş_{>37})]$

$= 0.954 - 0.607$

$= 0.347$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 50

E = 0.954

$E(Yaş_{<=50}) = -\ 3/6 \log_2 3/6 - 3/6 \log_2 3/6$

$= 1$

$E(Yaş_{>50}) = -\ 2/2 \log_2 2/2$

$= 0$

$Gain(S,Yaş) = 0.954 - [6/8*E(Yaş_{<=50}) + 2/8*E(Yaş_{>50})]$

$= 0.954 - 0.75$

$= 0.204$

C4.5 Algoritması - Örnek

Yaş	Memnun
21	HAYIR
19	EVET
22	HAYIR
21	EVET
30	EVET
60	HAYIR
45	HAYIR
55	HAYIR

Eşik değeri = 57

E = 0.954

$E(Yaş_{<=57}) = -\ 3/7 \log_2 3/7 - 4/7 \log_2 4/7$

$= 0.985$

$E(Yaş_{>57}) = -\ 1/1 \log_2 1/1$

$= 0$

$Gain(S,Yaş) = 0.954 - [7/8*E(Yaş_{<=57}) + 1/8*E(Yaş_{>57})]$

$= 0.954 - 0.862$

$= 0.092$

C4.5 Algoritması - Örnek

Eşik Değeri	Bilgi Kazancı
20	0.2
21	0.158
26	0.049
37	**0.347**
50	0.204
57	0.092

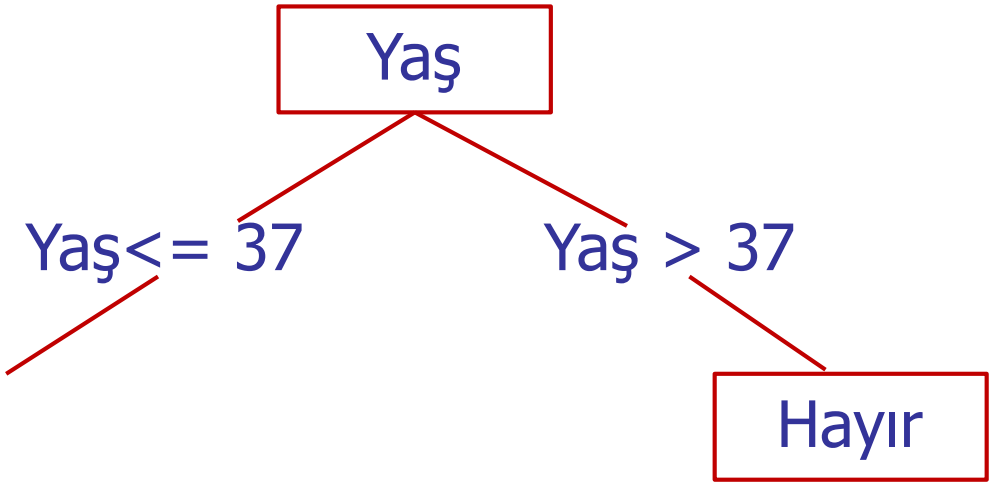
C4.5 Algoritması - Örnek

- Gain(S,Model) = 0.204
- Gain(S,Cinsiyet) = 0.003
- Gain(S,Yaş) = 0.347

Maximum information gain

Yaş

C4.5 Algoritması - Örnek



C4.5 Algoritması - Örnek

Model	Cinsiyet	Yaş	Memnun
X5	ERKEK	Yaş<=37	HAYIR
X3	KADIN	Yaş<=37	EVET
X5	ERKEK	Yaş<=37	HAYIR
X3	ERKEK	Yaş<=37	EVET
X3	ERKEK	Yaş<=37	EVET

$$\text{Entropi} = -\ 2/5 \log_2 2/5 - 3/5 \log_2 3/5$$

$$= 0.971$$

C4.5 Algoritması - Örnek

Model	Memnun
X5	HAYIR
X3	EVET
X5	HAYIR
X3	EVET
X3	EVET

E = 0.971

$E(Model_{X5}) = -\ 2/2 \log_2 2/2$

$= 0$

$E(Model_{X3}) = -\ 3/3 \log_2 3/3$

$= 0$

$Gain(S_{Yaş}, Model) = 0.971 - [2/5 * E(Model_{X5}) + 3/5 * E(Model_{X3})]$

$= 0.971 - 0$

$= 0.971$

C4.5 Algoritması - Örnek

Cinsiyet	Memnun
ERKEK	HAYIR
KADIN	EVET
ERKEK	HAYIR
ERKEK	EVET
ERKEK	EVET

E = 0.971

$E(Cinsiyet_{Erkek}) = -\ 2/4 \log_2 2/4 - 2/4 \log_2 2/4$

$= 1$

$E(Cinsiyet_{Kadın}) = -\ 1/1 \log_2 1/1$

$= 0$

$Gain(S_{Yaş}, Cinsiyet) = 0.971 - [4/5 * E(Cinsiyet_{Erkek}) + 1/5 * E(Cinsiyet_{Kadın})]$

$= 0.971 - 0.8$

$= 0.171$

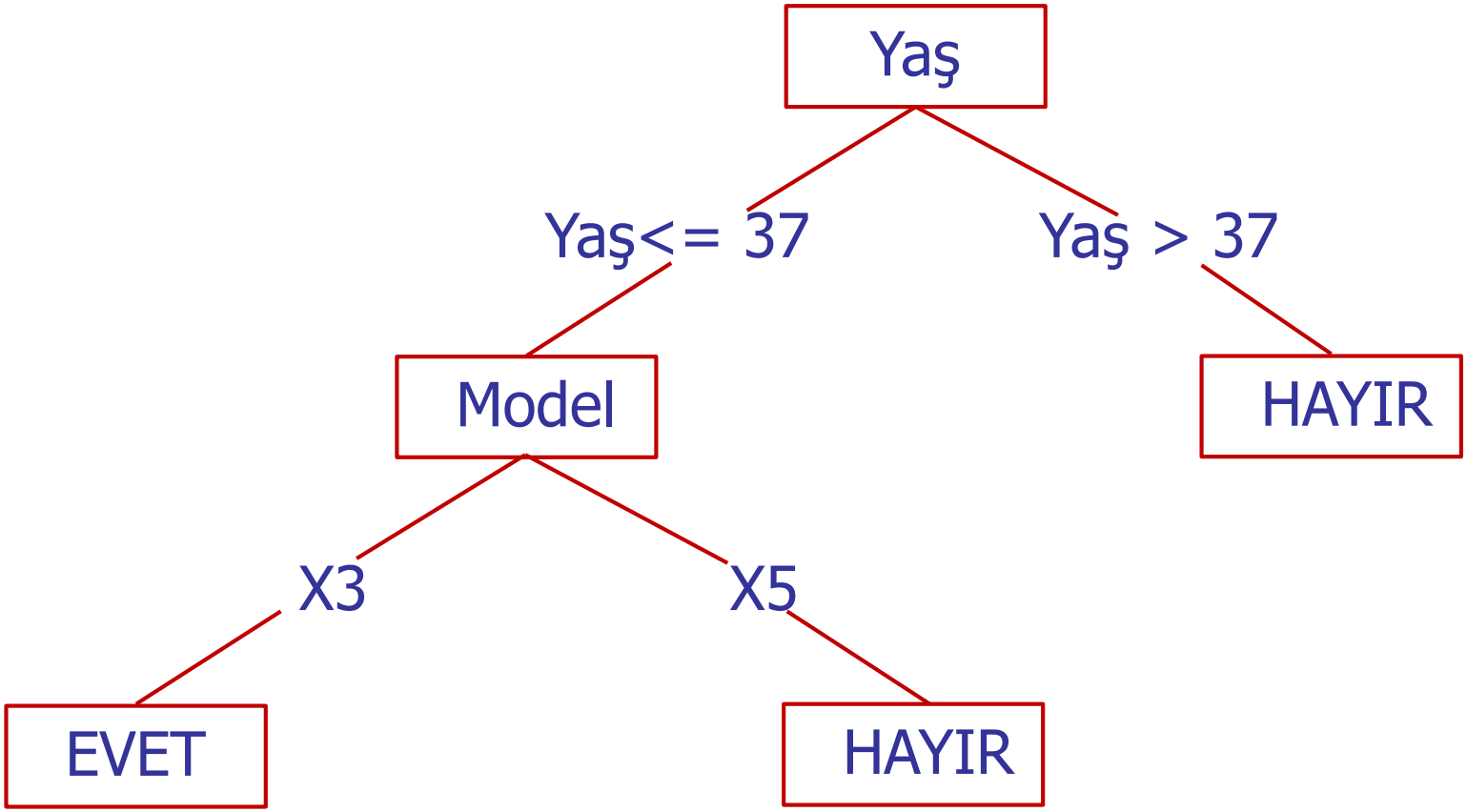
C4.5 Algoritması - Örnek

- $Gain(S_{Yaş}, Model) = 0.971$
- $Gain(S_{Yaş}, Cinsiyet) = 0.171$

Maximum information gain

Model

C4.5 Algoritması - Örnek



C4.5 Algoritması - Örnek

- Veri tablosu incelendiğinde Model=X5 için Memnun=HAYIR ve Model=X3 için Memnun=EVET değerlerinin karşılık geldiği görülmektedir. Cinsiyet niteliğinin herhangi bir etkisi olmadığı anlaşılmaktadır.

C4.5 Algoritması - Örnek

Karar kuralları:

- If (Yaş<=37) $\wedge$ (Model=X3) Then Memnun= EVET
- If (Yaş<=37) $\wedge$ (Model=X5) Then Memnun= HAYIR
- If (Yaş>37) Then Memnun= HAYIR