Büyük Veri İçin İstatistiksel Öğrenme (Statistical Learning for Big Data)

M. Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Bu dersin sunumları, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2017." ve "Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, Stanford University, 2011." kitapları kullanılarak hazırlanmıştır.

Konular

- Denetimli Öğrenmenin Temelleri
- Entropi
- Karar Ağaçları
- ID3 Algoritması
- C4.5 Algoritması

Denetimli Öğrenmenin Temelleri

- **Denetimli (gözetimli) öğrenme**, makine öğrenmesinde **sınıflandırma** veya **tümevarımlı (inductive) öğrenme** şeklinde ifade edilir.
- Denetimli öğrenmede hedef değerler (targets) ile giriş değerleri (inputs) birlikte eğitim kümesi **(training set)** olarak sağlanır.
- Öğrenme işleminde bir kayıt kümesi kullanılır ve özellikler kümesi olarak gösterilir.

  $A = \{A_1, A_2, \ldots, A_n\}$

- Burada, n kümedeki eleman sayısını gösterir.

3

Denetimli Öğrenmenin Temelleri

- Bir veri kümesi aynı zamanda hedef C özelliğine de (sınıf) sahip olabilir.
- $C \cap A = \varnothing$ dir ve aşağıdaki gibi ifade edilir:

  $C = \{c_1, c_2, \ldots, c_m\}, \quad m \geq 2$

- Verilen bir D veri kümesi için öğrenmedeki amaç, **A'daki özellikler ile C'deki sınıflar arasındaki ilişkiyi gösteren** bir **sınıflandırma/tahmin fonksiyonu** oluşturmaktır.
- Elde edilen bu fonksiyon, **sınıflandırma modeli, tahmin modeli** veya **sınıflandırıcı** olarak adlandırılır.

4

Denetimli Öğrenmenin Temelleri

Örnek

- Bir kredi uygulamasına yönelik veri kümesi

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

5

Denetimli Öğrenmenin Temelleri

- Bir veri kümesi ile öğrenen bir model geliştirerek gelecekteki yeni müşterilere ait verilerde kullanılabilir.
- Bu şekilde sınıf etiketlerinin de verildiği öğrenmeye **denetimli (supervised) öğrenme** denilir.
- Öğrenme sürecinde kullanılan veri kümesine **eğitim verisi (training data)** denir.
- Öğrenmeden sonraki değerlendirme sürecinde kullanılan veri kümesine ise **test verisi (test data)** denilmektedir.
- Eğitim verisinin de test verisinin de tüm sistemi temsil etme kapasitesine sahip olması gerekir.
- Test verisi eğitim sürecinde **görülmemiş veri (unseen data)** olarak oluşturulmalıdır.

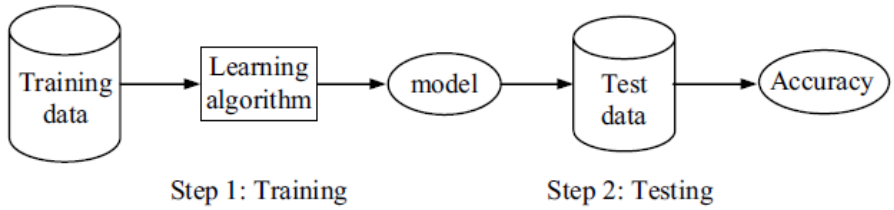
6

Denetimli Öğrenmenin Temelleri

- Geliştirilen modelin **doğruluk değeri (accuracy)**, test verisinde doğru sınıflandırma sayısıyla belirlenir.

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

- Öğrenme süreci training ve test aşamalarından oluşur.



Konular

- Denetimli Öğrenmenin Temelleri
- Entropi
- Karar Ağaçları
- ID3 Algoritması
- C4.5 Algoritması

Entropi

- **Entropi**, rastgele değere sahip bir değişken veya bir sistem için **belirsizlik ölçütüdür.**
- **Enformasyon,** rastsal bir olayın gerçekleşmesi halinde ortaya çıkan bilgi ölçütüdür.
- Bir süreç için entropi, tüm örnekler (durumlar) tarafından içerilen **enformasyonun değeridir.**
- **Eşit olasıklı durumlara sahip sistemler yüksek belirsizliğe sahiptirler.**
- Shannon, bir sistemdeki durum değişikliğinde, **entropideki değişimin enformasyon boyutunu tanımladığını öne sürmüştür.**
- Buna göre **bir sistemdeki belirsizlik arttıkça,** bir durum gerçekleştiğinde **elde edilecek enformasyon boyutu da artacaktır.**

9

Entropi

- Shannon bilgiyi bitlerle ifade ettiği için, logaritmayı 2 tabanında kullanmıştır ve enformasyon formülünü aşağıdaki gibi vermiştir.

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

- $P(x)$, x olayının gerçekleşme olasılığını gösterir.
- Shannon'a göre entropi, **iletilen bir mesajın taşıdığı enformasyonun değeridir.**
- Shannon entropisi H, aşağıdaki gibi ifade edilir:

$$H(X) = E(I(X)) = \sum_{1 \le i \le n} P(x_i) I(x_i)$$

$$= \sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P_{(x_i)} \log_2 P_{(x_i)}$$

10

Entropi

Örnek

- Bir paranın havaya atılması olayı rastsal X sürecini göstersin. Yazı ve tura gelme olasılıkları eşit olduğundan elde edilecek enformasyon,

$$I(X) = \log \frac{1}{P(X)} = \log \frac{1}{0{,}5} = \log 2 = 1$$

olur. **Bu olayın sonucunda 1 bitlik bilgi kazanılmıştır.**

- Entropi değeri ise 1 olarak bulunur.

$$H(X) = -\sum_{i=1}^{2} p_i \log_2 p_i$$
$$= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1$$

11

Entropi

Örnek

- Aşağıdaki 8 elemanlı S kümesi verilsin.

$S = \{evet, hayır, evet, hayır, hayır, hayır, hayır, hayır\}$

- ***"evet"* ve *"hayır"* için olasılık,**

$$p\,(evet) = \frac{2}{8} = 0{,}25 \qquad p\,(hayir) = \frac{6}{8} = 0{,}75$$

- **Entropi değeri,**

$$H(S) = p(evet) \log_2 \frac{1}{p(evet)} + p(hayir) \log_2 \frac{1}{p(hayir)}$$
$$= 0{,}25.\log_2 \frac{1}{0{,}25} + 0{,}75.\log_2 \frac{1}{0{,}75}$$
$$= 0{,}81$$

12

Konular

- Denetimli Öğrenmenin Temelleri
- Entropi
- Karar Ağaçları
- ID3 Algoritması
- C4.5 Algoritması

13

Karar Ağaçları

- Sınıflandırma problemleri için yaygın kullanılan yöntemdir.
- Sınıflandırma doğruluğu diğer öğrenme metotlarına göre çok etkindir.
- Öğrenmiş sınıflandırma modeli ağaç şeklinde gösterilir ve **karar ağacı (decision tree)** olarak adlandırılır.
- **ID3 ve C4.5, entropiye dayalı** sınıflandırma algoritmalarıdır.
- **Twoing ve Gini, CART (Classification And Regression Trees) sınıflandırma ve regresyon ağaçlarına dayalı** sınıflandırma algoritmalarıdır.
- CART algoritmalarında her düğümde bir kritere göre ikili bölünme yapılır.

14

Konular

- Denetimli Öğrenmenin Temelleri
- Entropi
- Karar Ağaçları
- ID3 Algoritması
- C4.5 Algoritması

15

ID3 Algoritması

- **ID3** (Iterative Dichotomiser 3) **algoritması sadece kategorik verilerle çalışmaktadır.**
- Karar ağaçları çok boyutlu veriyi belirlenmiş bir niteliğe göre parçalara böler.
- Her adımda verinin hangi özelliğine göre ne tür işlem yapılacağına karar verilir.
- **Oluşturulabilecek tüm ağaçların kombinasyonu çok fazladır.**
- Karar ağaçlarının en az düğüm ve yaprak ile oluşturulması için farklı algoritmalar kullanılarak bölme işlemi yapılır.

16

ID3 Algoritması

Karar ağacında entropi

- Bir eğitim kümesindeki sınıf niteliğinin alacağı değerler kümesi T, her bir sınıf değeri C_i olsun.
- T sınıf değerini içeren küme için P_t sınıfların olasılık dağılımı

$$P_t = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, ..., \frac{|C_k|}{|T|} \right)$$

şeklinde ifade edilir.

- T sınıf kümesi için ortalama entropi değeri ise,

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

şeklinde ifade edilir. p_i bir sınıf etiketinin olasılığını belirtir.

17

ID3 Algoritması

- **Karar ağaçlarında bölümlemeye hangi düğümden başlanacağı çok önemlidir.**
- Uygun düğümden başlanmazsa ağacın içerisindeki düğümlerin ve yaprakların sayısı çok fazla olacaktır.
- Bir risk kümesi aşağıdaki gibi tanımlansın. C_1=*"var"*, C_2=*"yok"*

RISK = {var, var, var, yok, var, yok, yok, var, var, yok}

$|C_1| = 6$ $\quad |C_2| = 4$ $\quad p_1 = 6/10 = 0{,}6$ $\quad p_2 = 4/10 = 0{,}4$

$$P_{RISK} = \left(\frac{6}{10}, \frac{4}{10} \right)$$

$$H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{6}{10} \log_2 \frac{6}{10} + \frac{4}{10} \log_2 \frac{4}{10} \right) = 0{,}97$$

18

ID3 Algoritması

Dallanma için niteliklerin seçimi

- Öncelikle **sınıf niteliğinin entropisi** hesaplanır.

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Çıkış sınıf vektörü değeri

Giriş özellik vektörü değeri

- Sonra **özellik vektörlerinin sınıfa bağımlı entropileri** hesaplanır.

$$H(X_k) = -\sum_{i=1}^{n} \frac{|T_i|}{|X_k|} \log \frac{|T_i|}{|X_k|} \qquad H(X,T) = \sum_{k=1}^{n} \frac{|X_k|}{|X|} H(X_k)$$

- Son olarak sınıf niteliğinin entropisinden tüm özellik vektörlerinin entropisi çıkartılarak her özellik için **kazanç ölçütü hesaplanır.**

$$Kazanç(X,T) = H(T) - H(X,T)$$

Sınıfa bağımlı entropi

- **En büyük kazanca sahip özellik vektörü** o iterasyon için dallanma düğümü olarak seçilir.

19

ID3 Algoritması

Örnek

- Aşağıdaki tablo için karar ağacı oluşturulsun.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(T) = H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$

20

ID3 Algoritması

Örnek – devam

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(BORÇ_{YÜKSEK}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$H(BORÇ_{DUSUK}) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0{,}863$$

$$H(BORÇ, RISK) = \frac{3}{10}H(BORÇ_{YÜKSEK}) + \frac{7}{10}H(BORÇ_{DUSUK})$$

$$= \frac{3}{10}(0) + \frac{7}{10}(0{,}863) = 0{,}64$$

$$Kazanç(BORÇ, RISK) = 1 - 0{,}64 = 0{,}36$$

21

ID3 Algoritması

Örnek – devam

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(GELIR_{YÜKSEK}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(GELIR_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(GELIR, RISK) = \frac{5}{10}H(GELIR_{YÜKSEK}) + \frac{5}{10}H(GELIR_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(GELIR, RISK) = 1 - 0{,}971 = 0{,}029$$

22

ID3 Algoritması

Örnek – devam

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(STATU_{ISVEREN}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(STATU_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(STATU, RISK) = \frac{5}{10}H(STATU_{YUKSEK}) + \frac{5}{10}H(STATU_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(STATU, RISK) = 1 - 0{,}971 = 0{,}029$$

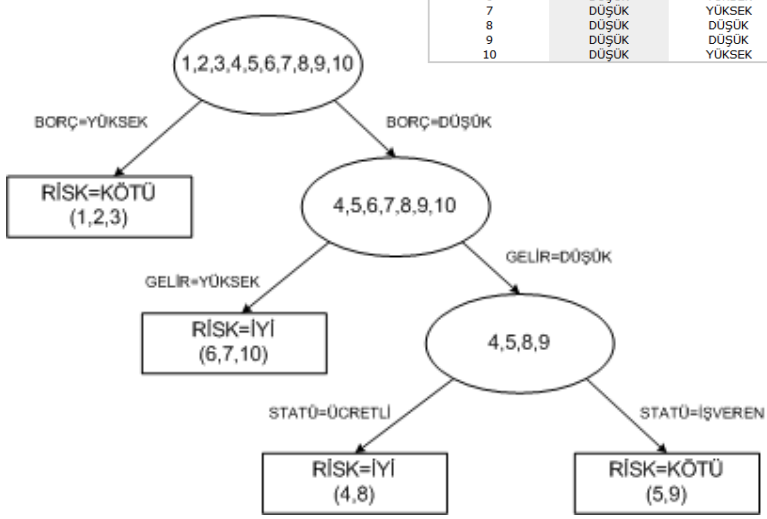
İlk dallanma için uygun seçim BORÇ niteliğidir.

23

ID3 Algoritması

Örnek – devam

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ



24

ID3 Algoritması

Örnek – devam

Karar ağacından elde edilen kurallar

1. **EĞER** (BORÇ = YÜKSEK) **İSE** (RİSK = KÖTÜ)
2. **EĞER** (BORÇ = DÜŞÜK) **VE** (GELİR = YÜKSEK) **İSE** (RİSK = İYİ)
3. **EĞER** (BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = ÜCRETLİ) **İSE** (RİSK = İYİ)
4. **EĞER** (BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = İŞVEREN) **İSE** (RİSK = KÖTÜ)

25

Konular

- Denetimli Öğrenmenin Temelleri
- Entropi
- Karar Ağaçları
- ID3 Algoritması
- C4.5 Algoritması

26

C4.5 Algoritması

- **C4.5 ile sayısal değerlere sahip nitelikler için karar ağacı oluşturmak için Quinlan tarafından geliştirilmiştir.**
- ID3 algoritmasından tek farkı nümerik değerlerin kategorik değerler haline dönüştürülmesidir.
- En büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir.
- Eşik değeri belirlemek için tüm değerler sıralanır ve ikiye bölünür.
- Eşik değer için $[v_i, v_{i+1}]$ aralığının orta noktası alınabilir.

$$t_i = \frac{v_i + v_{i+1}}{2}$$

- Nitelikteki değerler eşik değere göre iki kategoriye ayrılmış olur.

27

C4.5 Algoritması

Örnek

- NİTELİK2 = {65, 70, 75, **80, 85**, 90, 95, 96} için eşik değer **(80+85)/2 = 83** alınmıştır.

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	doğru	sınıf1
a	90	doğru	sınıf2
a	85	yanlış	sınıf2
a	95	yanlış	sınıf2
a	70	yanlış	sınıf1
b	90	doğru	sınıf1
b	78	yanlış	sınıf1
b	65	doğru	sınıf1
b	75	yanlış	sınıf1
c	80	doğru	sınıf2
c	70	doğru	sınıf2
c	80	yanlış	sınıf1
c	70	yanlış	sınıf1
c	96	yanlış	sınıf1

28

C4.5 Algoritması

Örnek – devam

NİTELİK 1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

29

C4.5 Algoritması

Örnek – devam

NİTELİK 1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

$$H(\text{SINIF}) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0{,}940$$

$$H(\text{NITELIK1}_a) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(\text{NITELIK1}_b) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0$$

$$H(\text{NITELIK1}_c) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(\text{NITELIK1, SINIF}) = \frac{5}{14}H(\text{NITELIK1}_a) + \frac{4}{14}H(\text{NITELIK1}_b) + \frac{5}{14}H(\textit{NITELIK1}_c)$$

$$= \frac{5}{14}0{,}971 + \frac{4}{14}0 + \frac{5}{14}0{,}971 = 0{,}694$$

$$\textit{Kazanç}(\text{NITELIK 1, SINIF}) = 0{,}940 - 0{,}694 = 0{,}246$$

30

C4.5 Algoritması

Örnek – devam

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

$$H(\text{NITELIK2}_{ek}) = -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0{,}765$$

$$H(\text{NITELIK2}_{b}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(\text{NITELIK2}, \text{SINIF}) = \frac{9}{14}H(\text{NITELIK2}_{ek}) + \frac{5}{14}H(\text{NITELIK1}_{b})$$

$$= \frac{9}{14}0{,}765 + \frac{5}{14}0{,}971 = 0{,}836$$

$$Kazanç(\text{NITELIK 2}, \text{SINIF}) = 0{,}940 - 0{,}836 = 0{,}104$$

31

C4.5 Algoritması

Örnek – devam

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

$$H(\text{NITELIK3}_{d}) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$$

$$H(\text{NITELIK3}_{y}) = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = 0{,}811$$

$$H(\text{NITELIK3}, \text{SINIF}) = \frac{6}{14}H(\text{NITELIK3}_{d}) + \frac{8}{14}H(\text{NITELIK3}_{y})$$

$$= \frac{6}{14}1 + \frac{8}{14}0{,}811 = 0{,}892$$

$$Kazanç(\text{NITELIK 3}, \text{SINIF}) = 0{,}940 - 0{,}892 = 0{,}048$$

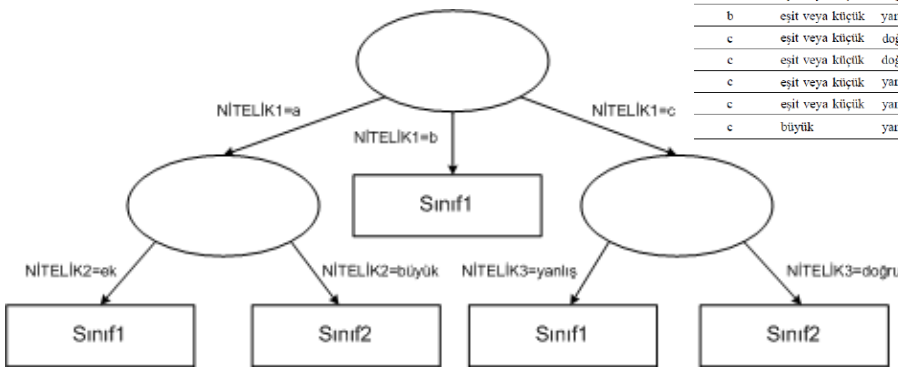
$$Kazanç(\text{NITELIK 3}, \text{SINIF}) < Kazanç(\text{NITELIK 2}, \text{SINIF}) < Kazanç(\text{NITELIK 1}, \text{SINIF})$$

32

C4.5 Algoritması

Örnek – devam

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1



33

C4.5 Algoritması

Örnek – devam

Karar ağacından elde edilen kurallar

1. **EĞER** (NİTELİK1 = a) **VE** (NİTELİK2 = Eşit veya Küçük) **İSE** (SINIF = Sınıf1)
2. **EĞER** (NİTELİK1 = a) **VE** (NİTELİK2 = Büyük) **İSE** (SINIF = Sınıf2)
3. **EĞER** (NİTELİK1 = b) **İSE** (SINIF = Sınıf1)
4. **EĞER** (NİTELİK1 = c) **VE** (NİTELİK3 = yanlış) **İSE** (SINIF = Sınıf1)
5. **EĞER** (NİTELİK1 = c) **VE** (NİTELİK3 = doğru) **İSE** (SINIF = Sınıf2)

34