Veri Madenciliği

Bölüm 5. Sınıflandırma 1

Doç. Dr. Suat Özdemir http://ceng.gazi.edu.tr/~ozdemir

Gözetimli & Gözetimsiz Öğrenme

- Predictive Data Mining vs. Descriptive Data Mining
- Gözetimli (Supervised) öğrenme= **sınıflandırma (clasification)**
  - Öğrenme kümesindeki sınıfların sayısı ve hangi nesnenin hangi sınıfta olduğu biliniyor.
- Gözetimsiz (Unsupervised) öğrenme = demetleme (clustering)
  - Öğrenme kümesinde hangi nesnenin hangi sınıfta olduğu bilinmiyor. Genelde sınıf sayısı da bilinmiyor.
  - 2 hafta sonra işleyeceğiz

Sınıflandırma ve tahmin

- Sınıflandırma (Classification)
  - Kategorik sınıf etiketlerini öngörme
  - Bir model oluşturur ve veriyi sınıflandırır
    - Ögrenme seti (the training set)
    - Sınıf etiketleri (class label) biliniyor
  - Sınıfı bilinmeyen veriler (sınama seti) oluşturulan modele göre sınıflandırılır
  - Sınıflar arasında ilişki yok
- Sayısal öngörü, tahmin (Numeric Prediction)
  - Sürekli değere sahip fonksiyonları modeller
    - Bu modellere göre bilinmeyen ya da eksik değerleri tahmin eder
  - Öngörülen değerler arasında ilişki var

Sınıflandırma – Problem/Amaç/Yöntem

- Sınıflandırma = **ayrık** değişkenlerin hangi kategoride (sınıfta) olduklarını diğer nitelikleri kullanarak tahmin etme/öngörme
- Girdi:
  - Ayrık nesnelerden oluşan veri kümesi (öğrenme kümesi):
  - Her nesne niteliklerden oluşur, niteliklerden biri sınıf bilgisidir (sınıf etiketi)
- Yöntem:
  - Sınıf niteliğini belirlemek için diğer nitelikleri kullanarak bir model oluşturulur
  - Bulunan modelin başarımı belirlenir (sınama kümesi ile)
  - Model başarımı: doğru sınıflandırılmış sınama kümesi örneklerinin oranı
- Çıktı:
  - Sınıf etiketi belli olmayan nesneler oluşturulan model kullanılarak mümkün olan en iyi şekilde doğru sınıflara atanır

Sınıflandırma Uygulamaları

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

Sınıflandırma için Veri Önişleme

- Veri dönüşümü:
  - Sürekli nitelik değeri ayrık hale getirilir
    - 0-25 yas -> Genç
    - 26-50 yas -> Orta yaş
    - 51 ve ustu -> Yaşlı
  - Normalizasyon ([-1,...,1],[0,...,1])
- Veri temizleme:
  - gürültüyü azaltma (noise reduction)
  - gereksiz nitelikleri silme

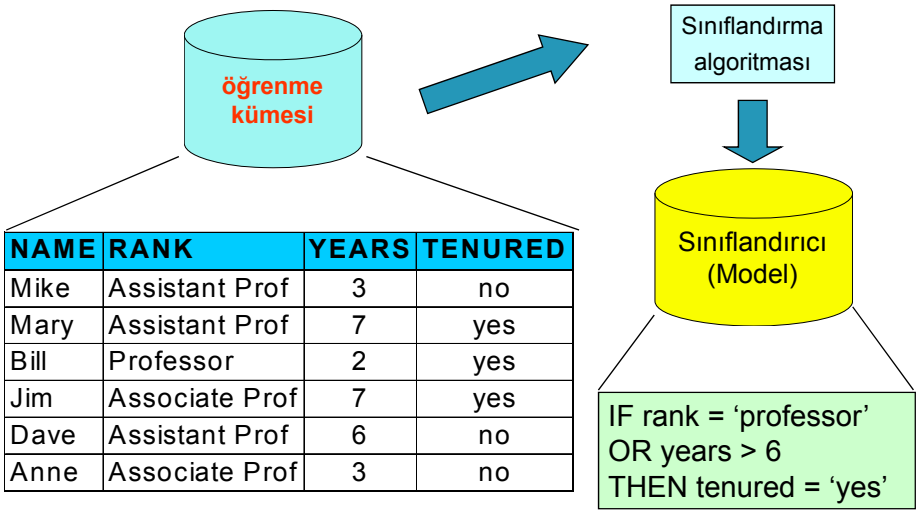
Sınıflandırma İşlemi

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

Sınıflandırma İşlemi: Model Oluşturma

- Model Oluşturma:
  - Her nesnenin sınıf etiketi olarak tanımlanan niteliğinin belirlediği bir sınıfta olduğu varsayılır
  - 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

Model Oluşturma

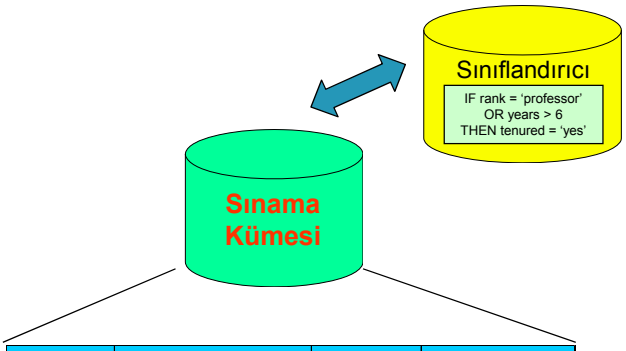


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Sınıflandırma İşlemi: ModelDeğerlendirme

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

Modeli değerlendirme



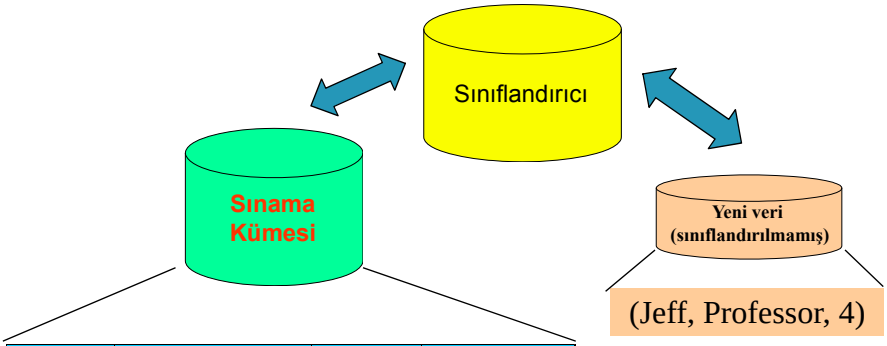
NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Model başarımı %75

Sınıflandırma İşlemi: Modeli Kullanma

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

Modeli kullanma



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4)

Tenured?

Yes

Sınıflandırıcı Başarımını Değerlendirme

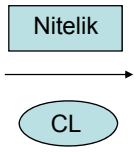
- Doğru sınıflandırma başarısı
  - Hız
    - modeli oluşturmak için gerekli süre
    - sınıflandırma yapmak için gerekli süre
  - Kararlı olması
    - veri kümesinde gürültülü ve eksik nitelik değerleri olduğu durumlarda da iyi sonuç vermesi
  - Ölçeklenebilirlik
    - büyük miktarda veri kümesi ile çalışabilmesi
  - Anlaşılabilir olması
    - kullanıcı tarafından yorumlanabilir olması
- Kuralların yapısı
  - birbiriyle örtüşmeyen kurallar

Sınıflandırma Yöntemleri

- **Karar ağaçları (decision trees)**
- Bayes sınıflandırıcılar (Bayes classifier)
- Yapay sinir ağları (artificial neural networks)
- 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)

Karar ağaçları

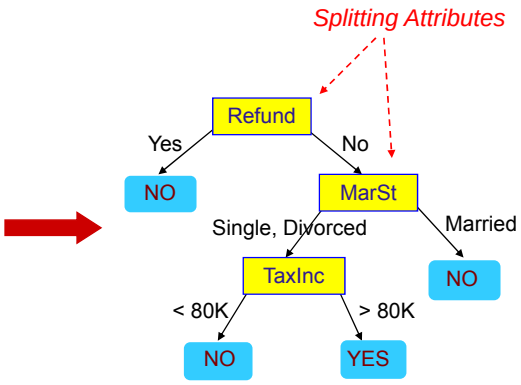
- Akış diyagramı şeklinde ağaç yapısı
  - Her ara düğüm bir nitelik sınaması
  - Dallar sınama sonucu
  - Yapraklar sınıflar



Karar Agacı Örneği - 1

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

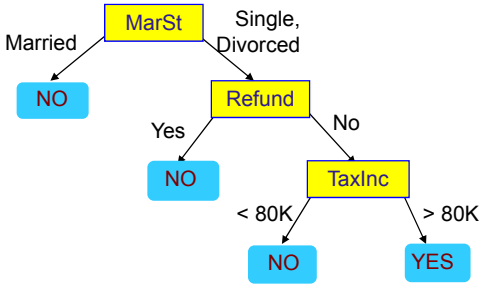
Training Data



Model: Decision Tree

Karar Ağacı Örneği - 2

	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

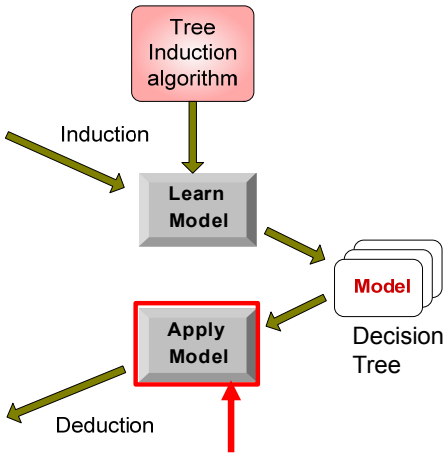
Karar Ağacı Sınıflandırma

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

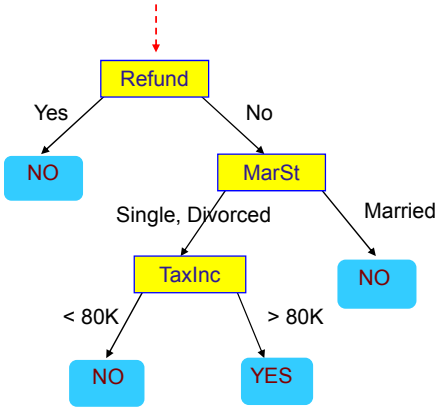
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Modelin Test Verisine Uygulanması

Start from the root of tree.



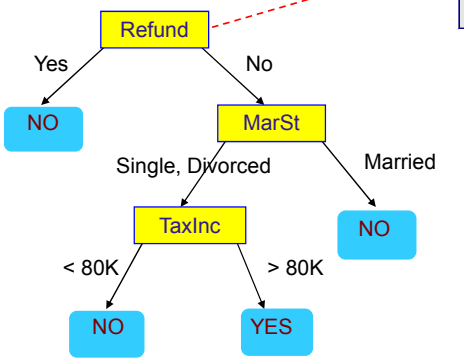
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Modelin Test Verisine Uygulanması

Test Data

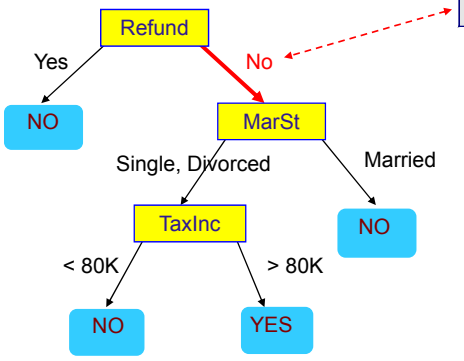
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Modelin Test Verisine Uygulanması

Test Data

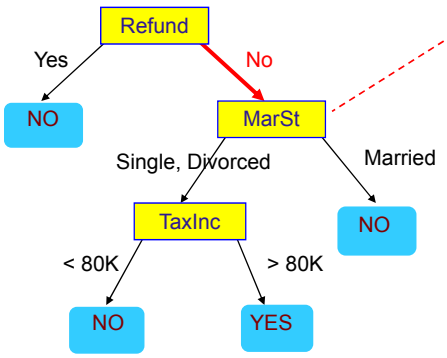
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Modelin Test Verisine Uygulanması

Test Data

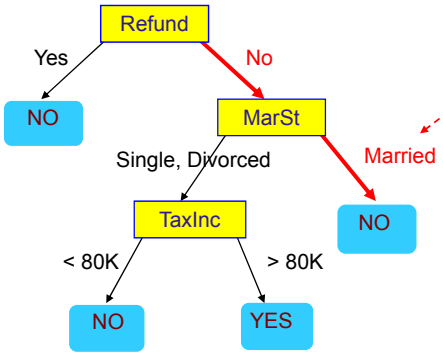
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



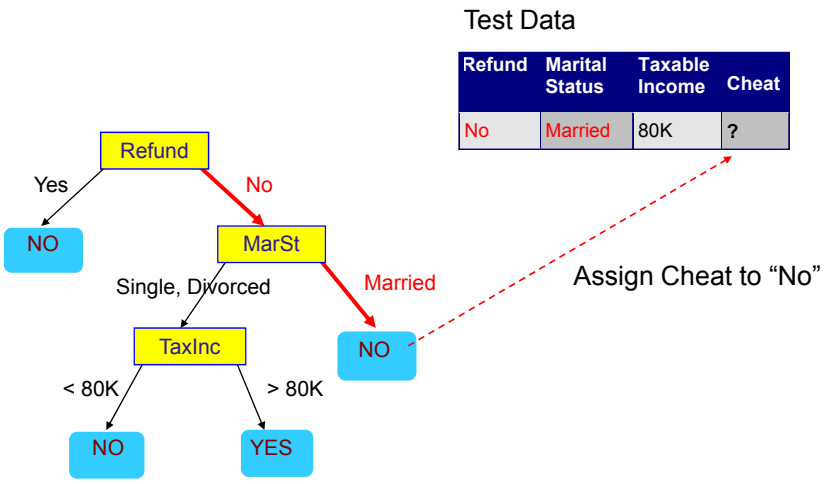
Modelin Test Verisine Uygulanması

Test Data

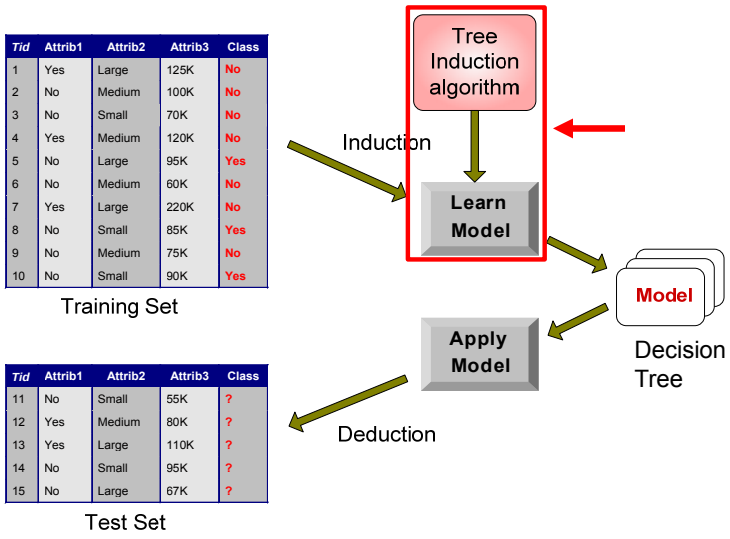
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Modelin Test Verisine Uygulanması



Karar Ağacı Sınıflandırma



Karar Ağacı Oluşturma

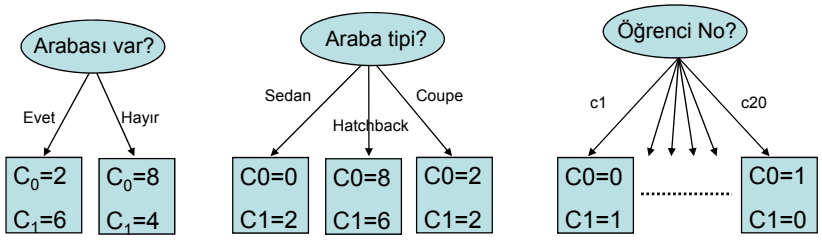
- Ağaç **top-down recursive divide-and-conquer** bir yaklaşım ile oluşturulur.
  - ağaç bütün verinin oluşturduğu tek bir düğümle başlıyor
  - nitelikler kategorik (eğer sürekli nitelikler varsa önceden ayrıştır)
  - eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor
  - **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

Karar ağaçları

- Karar ağacı oluşturma yöntemleri genel olarak iki aşamadan oluşur:
  - 1. ağaç oluşturma
    - en başta bütün öğrenme kümesi örnekleri kökte seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor
  - 2. ağaç budama
    - öğrenme kümesindeki gürültülü verilerden oluşan ve sınama kümesinde hataya neden olan dalları silme (sınıflandırma başarımını artırır)

En İyi Bölen Nitelik Hangisi?

- Bölmeden önce:
  - 10 örnek C_0 sınıfında (Erkek öğrenciler)
  - 10 örnek C_1 sınıfında (Kız öğrenciler)



En iyi bölen nitelik

- "Greedy" yaklaşım
  - çoğunlukla aynı sınıfa ait örneklerin bulunduğu (homojen) düğümler tercih edilir
- Düğümün kalitesini ölçmek için bir yöntem

Homojen değil (Kalitesiz)

Homojen (Kaliteli)

En iyi bölen nitelik seçimi

- İyilik Fonksiyonu (Goodness Function)
  - Farklı algoritmalar farklı iyilik fonksiyonları kullanabilir:
    - **Bilgi kazancı (information gain): ID3**
    - **Kazanç oranı (gain ratio): C4.5**
      - bütün niteliklerin ayrık değerler aldığı varsayılıyor
      - sürekli değişkenlere uygulamak için değişiklik yapılabilir
    - **Gini index: CART, IBM IntelligentMiner**
      - bütün niteliklerin sürekli değerler aldığı varsayılıyor
      - her nitelik için farklı bölme değerleri olduğu varsayılıyor
      - bölme değerlerini belirlemek için başka yöntemlere (demetleme gibi) ihtiyaç var
      - ayrık değişkenlere uygulamak için değişiklik yapılabilir

Bilgi kazancı (Information gain)

- Bir torbadaki topların renkleri farklı ise belirsizlik fazladır
- Topların hepsi aynı renkte ise belirsizlik yoktur
- Information theory 'e (bilgi kuramına) dayanır
- The concept was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication".
- The **Shannon entropy** (entropi) or **information entropy** is a measure of the uncertainty associated with a random variable.
  - Belirsizliğin ölçütü

Entropi

- $p_1, p_2, .., p_m$ toplamları 1 olan olasılıklar.

$$Entropi = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- örnekler aynı sınıfa aitse entropi=0
- örnekler sınıflar arasında eşit dağılmışsa entropi=1
- örnekler sınıflar arasında rastgele dağılmışsa 0<entropi<1

Bilgi kazanımı: Information Gain

- Bilgi kuramı kavramlarını kullanarak karar ağacı oluşturulur.
- Sınıflandırma sonucu için en az sayıda karşılaştırma yapmayı hedefler.
- Ağaç bir niteliğe göre dallandığında entropi ne kadar düşer?

En iyi bölen nitelik seçimi: Information Gain

- Bilgi kazanımı en yüksek olan nitelik seçilir
- p_i D öğrenme kümesindeki bir varlığın C_i sınıfına ait olma olasılığı, $|C_{i,D}|/|D|$ olarak ifade edilir
- D içindeki bir varlığı sınıflandırmak için gerekli bilgi (D nin entropisi):

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- D kümesi A niteliğine göre v parçaya bölündükten sonra D yi sınıflandırmak için gerekli olan bilgi:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- A niteliğine göre bölünmeden dolayı bilgi kazancı

$$Gain(A) = Info(D) - Info_A(D)$$

Müşteri veritabanı

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

ÖRNEK:

buys_computer sınıfı için en iyi bölen niteliği bulunuz?

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

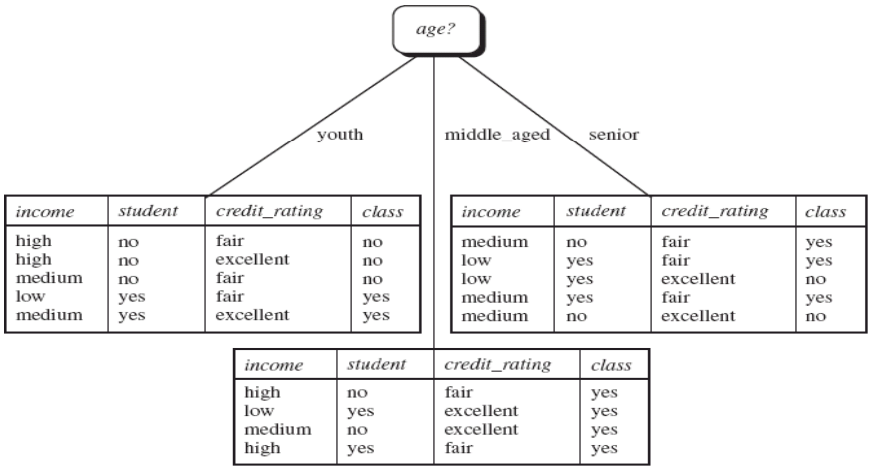
$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

ÖRNEK

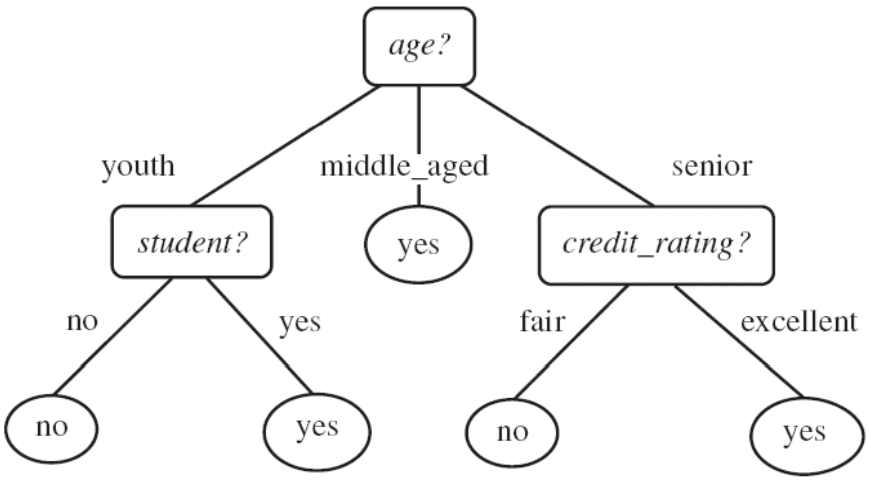


income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

"buys_computer" sınıfı için karar ağacı



En iyi bölen nitelik seçimi: Kazanım Oranı (Gain Ratio)

- Bilgi kazanımı metodu çok çeşitli değerlere sahip nitelikleri seçme eğilimdedir
- Bu problemi çözmek için C4.5 (a successor of ID3) kazanım oranını kullanır (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

- GainRatio(A) = Gain(A)/SplitInfo(A)
- En yüksek kazanım oranına sahip nitelik seçilir

Gini index (CART, IBM IntelligentMiner)

- *D kümesi n* sınıfdan örnekler içeriyorsa, gini index, *gini*(*D*) şu şekilde ifade edilir (p_j *j* sınıfının *D* kümesinde görülme sıklığıdır)

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

- *D* kümesi A niteliğine göre ikiye D_1 ve D_2 olarak bölünürse, *gini* index *gini*(*D*) şu şekilde ifade edilir

$$gini_{\ A}(D) = \frac{|D_1|}{|D|} gini\ (D_1) + \frac{|D_2|}{|D|} gini\ (D_2)$$

- Kirlilikteki azalma (Reduction in Impurity)

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- En küçük $gini_{split}(D)$ ye sahip nitelik (or the largest reduction in impurity) bölme noktası olarak seçilir

Örnek : Gini Index

- Önceki örnekte 9 kişi buys_computer = "yes" ve 5 kişi "no" sınıfında

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

- income niteliğinin D'yi 2'ye böldüğünü kabul edelim: 10 in D_1: {low, medium} and 4 in D_2 :{high}

$$gini_{income \in \{low,medium\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

$Gini_{\{low,high\}}$ is 0.458; $Gini_{\{medium,high\}}$ is 0.450. En düşük Gini index değerini verdiğinden {low,medium} (and {high}) ayrımı seçilir.