DATA MINING I

Decision Tree

Tim Dosen Data Mining I

Diterjemahkan dari buku Data Mining Concepts and Techniques Edisi ketiga dengan penulis Jiawei Han, Micheline Kamber, dan Jian Pei.

Teknologi Sains Data
Fakultas Teknologi Maju dan Multidisiplin
Universitas Airlangga

1

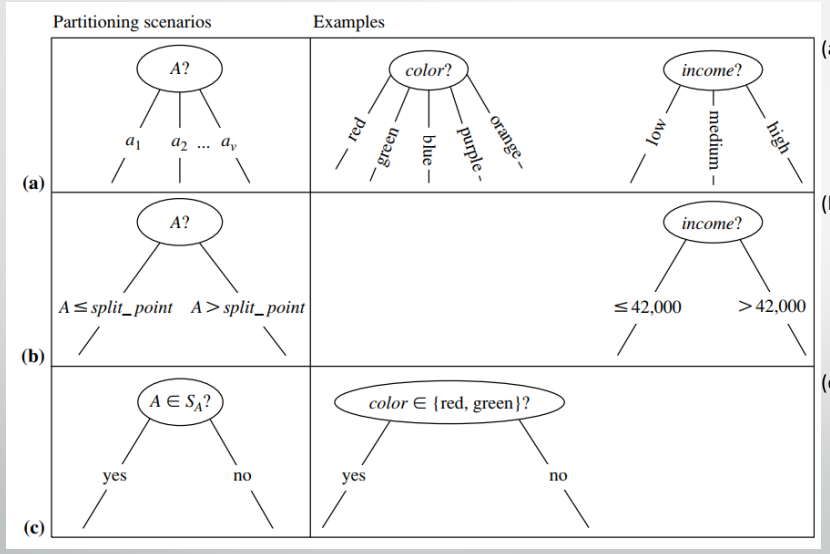
1

Decision Tree

- Pada akhir 1970-an dan awal 1980-an, J. Ross Quinlan, mengembangkan algoritma pohon keputusan yang dikenal sebagai ID3 (Iterative Dichotomiser). Quinlan kemudian mempresentasikan C4.5 (penerus ID3). Pada tahun 1984, sekelompok ahli statistik (L. Breiman, J. Friedman, R. Olshen, dan C. Stone) menerbitkan buku Classification and Regression Trees (CART), yang menggambarkan generasi pohon keputusan biner. ID3 dan CART ditemukan secara independen satu sama lain pada waktu yang hampir bersamaan, tetapi mengikuti pendekatan serupa untuk mempelajari pohon keputusan dari tupel pelatihan. Kedua algoritma landasan ini melahirkan banyak pekerjaan pada induksi pohon keputusan.
- ID3, C4.5, dan CART mengadopsi pendekatan nonbacktracking di mana pohon keputusan dibangun dengan cara membagi-dan-menaklukkan rekursif top-down. Sebagian besar algoritme untuk induksi pohon keputusan juga mengikuti pendekatan top-down, yang dimulai dengan set tupel pelatihan dan label kelas yang terkait. Set pelatihan secara rekursif dipartisi menjadi subset yang lebih kecil saat pohon sedang dibangun.

(1)

2



(a) If *A* is discrete-valued, then one branch is grown for each known value of *A*.

(b) If *A* is continuous-valued, then two branches are grown, corresponding to $A \leq$ *split point* and $A >$ *split point*.

(c) If *A* is discrete-valued and a binary tree must be produced, then the test is of the form $A \in S_A$, where S_A is the splitting subset for *A*.

3

Ukuran Pemilihan Atribut

- Ukuran pemilihan atribut memberikan peringkat untuk setiap atribut yang menjelaskan tupel pelatihan yang diberikan.
- Tiga ukuran pemilihan atribut, diantaranya:
  - Information Gain
  - Gain Ratio
  - Gini Index

4

Information Gain

Atribut dengan perolehan informasi tertinggi dipilih sebagai atribut pemisahan untuk simpul N. Atribut ini meminimalkan informasi yang diperlukan untuk mengklasifikasikan tupel dalam partisi yang dihasilkan dan mencerminkan keacakan terkecil atau "*impurity*" di partisi ini.

Atribut A dengan information gain tertinggi, Gain(A), dipilih sebagai atribut splitting pada node N.

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

Info(D) juga dikenal sebagai entropi D.

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

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

5

Contoh 1:

Untuk menemukan kriteria pemisahan untuk tupel ini, kita harus menghitung perolehan informasi dari setiap atribut.

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

$$Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) = 0.694 \text{ bits.}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

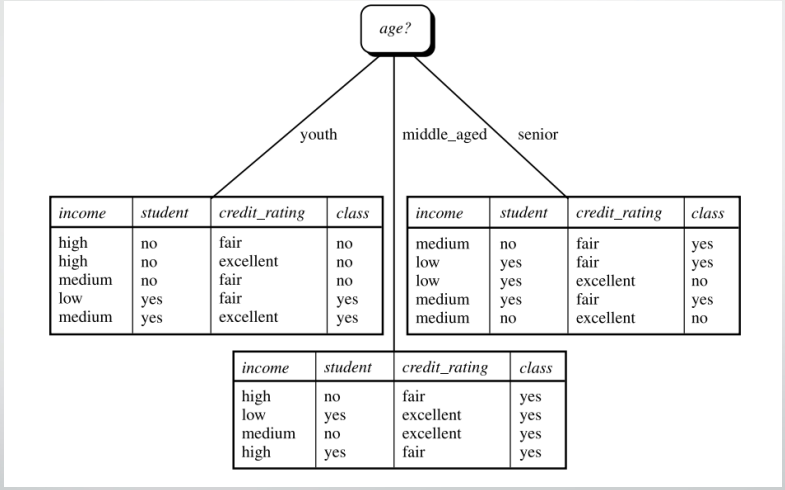
Class-Labeled Training Tuples from the *AllElectronics* Customer Database

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

Gain(income)=0.029 bits
Gain(student)=0.151 bits,
Gain(credit rating)=0.048 bits

Karena usia memiliki perolehan informasi tertinggi di antara atribut, itu dipilih sebagai atribut pemisah.

6



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

7

Gain Ratio

C4.5, penerus ID3, menggunakan ekstensi untuk mendapatkan informasi yang dikenal sebagai gain ratio, yang mencoba untuk mengatasi bias ini. Ini menerapkan semacam normalisasi untuk perolehan informasi menggunakan nilai "informasi terpisah" yang didefinisikan secara analog dengan Info(D) seperti

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

Nilai ini mewakili informasi potensial yang dihasilkan dengan memisahkan set data pelatihan, D, ke dalam partisi v, sesuai dengan hasil v dari pengujian pada atribut A.

Contoh 2: Perhitungan gain ratio untuk atribut income

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right)$$

$$= 1.557.$$

Dari contoh 1, Gain(*income*) D 0.029. Jadi, *GainRatio*(*income*)=0.029/1.557 D 0.019.

8

Gini Index

- Indeks Gini digunakan dalam CART. Menggunakan notasi yang dijelaskan sebelumnya, indeks Gini mengukur impurity dari D, partisi data atau set tupel pelatihan, sebagai

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

$$\Delta Gini(A) = Gini(D) - Gini_A(D).$$

- Atribut yang memaksimalkan pengurangan pengotor (atau, setara, memiliki indeks Gini minimum) dipilih sebagai atribut pemisahan. Atribut ini dan subset pemisahnya (untuk atribut pemisahan bernilai diskrit) atau titik pisah (untuk atribut pemisahan bernilai kontinu) bersama-sama membentuk kriteria pemisahan.

9

THANK YOU

10