分類方法與軟體操作

國立宜蘭大學資訊工程系
吳政瑋助理教授
wucw@niu.edu.tw

教學目標

- 了解機率式分類技術之基礎概念
- 了解樸素貝式分類演算法之運算過程
- 了解樸素貝式分類之限制及缺點
- 以Orange軟體建構樸素貝式分類器

Naive Bayesian Classification Basic Notations (Cont. 1/2)

- Each data sample is a n-dim feature vector
  - $X = (x_1, x_2, \ldots, x_n)$ for attributes $A_1, A_2, \ldots, A_n$
- Suppose there are m classes
  - $C = \{C_1, C_2, .., C_m\}$
- The classifier will predict X to the class C_i that has the highest posterior probability, conditioned on X
  - X belongs to C_i iff $P(C_i|X) > P(C_j|X)$ for all $1 \leq j \leq m, j \neq I$
  - e.g., $C=\{C_1=Yes, C_2=No\}$, X belongs to C_1 iff $P(C_1=Yes|X) > P(C_2=No|X)$

Technical Terms in Naïve Bayesian Theorem
貝氏定理中的專有名詞介紹

給定 C_i 的條件下 X 會發生的事後機率

Likelihood; The Probability of Predictor X Given a Class C_i

C_i 的事前機率

Prior Probability of Class C_i

$$P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$$

Posterior Probability of Class Ci Given a Predictor X

給定 X 的條件下 C_i 會發生的事後機率

Prior Probability of Predictor X

X 的事前機率

Naive Bayesian Classification
Bayesian Theorem

- $P(C_i|X) = P(X|C_i) \times P(C_i) / P(X)$
- *Rationale*
  - $P(C_i|X) = P(C_i \cup X) / P(X)$
  - $P(X|C_i) = P(C_i \cup X) / P(C_i)$
  - $P(C_i \cup X) = P(X|C_i) \times P(C_i)$

- $P(C_i) = s_i / |D|$
  - s_i is the number of training sample of class C_i
  - $|D|$ is the total number of training samples
- Assumption: Independent between Attributes
  - $P(X|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times P(x_3|C_i) \times \ldots \times P(x_n|C_i)$
- $P(X)$ can be ignored

獨立事件
Independent Events

- 假設有一個賣場，年齡分佈與性別分佈分別如下表:

年紀	老年	中年	少年
人數	100	400	500
百分比	10%	40%	50%
機率	0.1	0.4	0.5

性別	男性	女性
人數	400	600
百分比	40%	60%
機率	0.4	0.6

- 若這兩種事件(年紀與性別)為獨立事件，則又是少年同時又是女性的機率為 $(0.5 \times 0.6) = 0.3$

Naive Bayesian Classification
Why $P(X)$ can be ignore?

- X belongs to C_i iff $P(C_i|X) > P(C_j|X)$ for all $1 \le j \le m, j \ne I$

- 若要比較 $P(C_1=Yes|X)$ 與 $P(C_2=No|X)$ 之大小
  - $P(C_1|X) = \boldsymbol{P(X|C_1) \times P(C_1)} / P(X)$
  - $P(C_2|X) = \boldsymbol{P(X|C_2) \times P(C_2)} / P(X)$
  - 只要比較 $\boldsymbol{P(X|C_1) \times P(C_1)}$ 與 $\boldsymbol{P(X|C_2) \times P(C_2)}$

Naive Bayesian Classification
A Running Example

Classify X=(age="<=30", income="medium", student="yes", credit-rating="fair")

P(buys_computer=yes) = 9/14

P(buys_computer=no)=5/14

P(age=<30|buys_computer=yes)=2/9

P(age=<30|buys_computer=no)=3/5

P(income=medium|buys_computer=yes)=4/9

P(income=medium|buys_computer=no)=2/5

P(student=yes|buys_computer=yes)=6/9

P(student=yes|buys_computer=no)=1/5

P(credit-rating=fair|buys_computer=yes)=6/9

P(credit-rating =fair|buys_computer=no)=2/5

P(X|buys_computer=yes)=0.043

P(X|buys_computer=no)=0.019

Training data tuples from the *AllElectronics* customer database.

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

P(buys_computer=yes|*X*) = *P*(X|buys_computer=yes) *P*(buys_computer=yes) = 0.028

P(buys_computer=no|*X*) = *P*(X|buys_computer=no) *P*(buys_computer=no) = 0.0068

Naive Bayesian Classification
A Running Example

Classify X=(age="<=30", income="medium", student="yes", credit-rating="fair")

X= (x_1, x_2, x_3, x_4) = (x_1="<=30", x_2 ="medium", x_3 ="yes", x_4 ="fair")

C_1 = (buys_computer=yes), $P(C_1)$ = 9/14 = 0.64

C_2 = (buys_computer=No), $P(C_2)$= 5/14 = 0.36

$P(x_1|C_1)$ = 2/9 = 0.22

$P(x_1|C_2)$ = 3/5 = 0.6

$P(x_2|C_1)$ = 4/9 = 0.44

$P(x_2|C_2)$ = 2/5 = 0.4

$P(x_3|C_1)$ = 6/9 = 0.67

$P(x_3|C_2)$ = 1/5 = 0.2

$P(x_4|C_1)$ = 6/9 = 0.67

$P(x_4|C_2)$ = 2/5 = 0.4

Training data tuples from the *AllElectronics* customer database.

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31 . . . 40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31 . . . 40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31 . . . 40	medium	no	excellent	yes
13	31 . . . 40	high	yes	fair	yes
14	>40	medium	no	excellent	no

$P(X|C_1) = P(x_1|C_1) \times P(x_2|C_1) \times P(x_3|C_1) \times P(x_4|C_1) = 0.043$ (因為獨立事件)

$P(X|C_2) = P(x_1|C_2) \times P(x_2|C_2) \times P(x_3|C_2) \times P(x_4|C_2) = 0.019$

$P(C_1|X) = P(X|C_1) \times P(C_1) = 0.043 \times 0.64 = 0.028$

$P(C_2|X) = P(X|C_2) \times P(C_2) = 0.019 \times 0.36 = 0.0068$

Exercise 1

department	*status*	*age*	*salary*	*count*
sales	senior	31 . . . 35	46K . . . 50K	30
sales	junior	26 . . . 30	26K . . . 30K	40
sales	junior	31 . . . 35	31K . . . 35K	40
systems	junior	21 . . . 25	46K . . . 50K	20
systems	senior	31 . . . 35	66K . . . 70K	5
systems	junior	26 . . . 30	46K . . . 50K	3
systems	senior	41 . . . 45	66K . . . 70K	3
marketing	senior	36 . . . 40	46K . . . 50K	10
marketing	junior	31 . . . 35	41K . . . 45K	4
secretary	senior	46 . . . 50	36K . . . 40K	4
secretary	junior	26 . . . 30	26K . . . 30K	6

Let *salary* be the class label attribute.

Given a data sample with the values "*systems*", "*junior*", and "*26 . . . 30*" for the attributes *department*, *status*, and *age*, respectively, what would a naive Bayesian classification of the *salary* for the sample be?

Exercise 2

Location = Urban,

Age = Below 21,

Marriage = Married,

Gender = Female,

Loyalty =?

No.	Attributes				Class
	Location	Age	Marriage status	Gender	(Loyalty)
1	Urban	Below 21	Married	Female	Low
2	Urban	Below 21	Married	Male	Low
3	Suburban	Below 21	Married	Female	High
4	Rural	Between 21 and 30	Married	Female	High
5	Rural	Above 30	Single	Female	High
6	Rural	Above 30	Single	Male	Low
7	Suburban	Above 30	Single	Male	High
8	Urban	Between 21 and 30	Married	Female	Low
9	Urban	Above 30	Single	Female	High
10	Rural	Between 21 and 30	Single	Female	High
11	Urban	Between 21 and 30	Single	Male	High
12	Suburban	Between 21 and 30	Married	Male	High
13	Suburban	Below 21	Single	Female	High
14	Rural	Between 21 and 30	Married	Male	Low