Classification

Classification

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

- Classification models predict categorical class labels.

Naive Bayes Classification

- Let D be a training set of tuples and their associated class labels.

- Tuple is represented by an n-dimensional attribute vector, X = (x1, x2,..., xn),

- Measurements made on the tuple from n attributes, respectively, A1, A2,..., An

Naive Bayes Classification

- Naïve Bayesian classifier predicts that tuple X belongs to the class Ci, if and only if

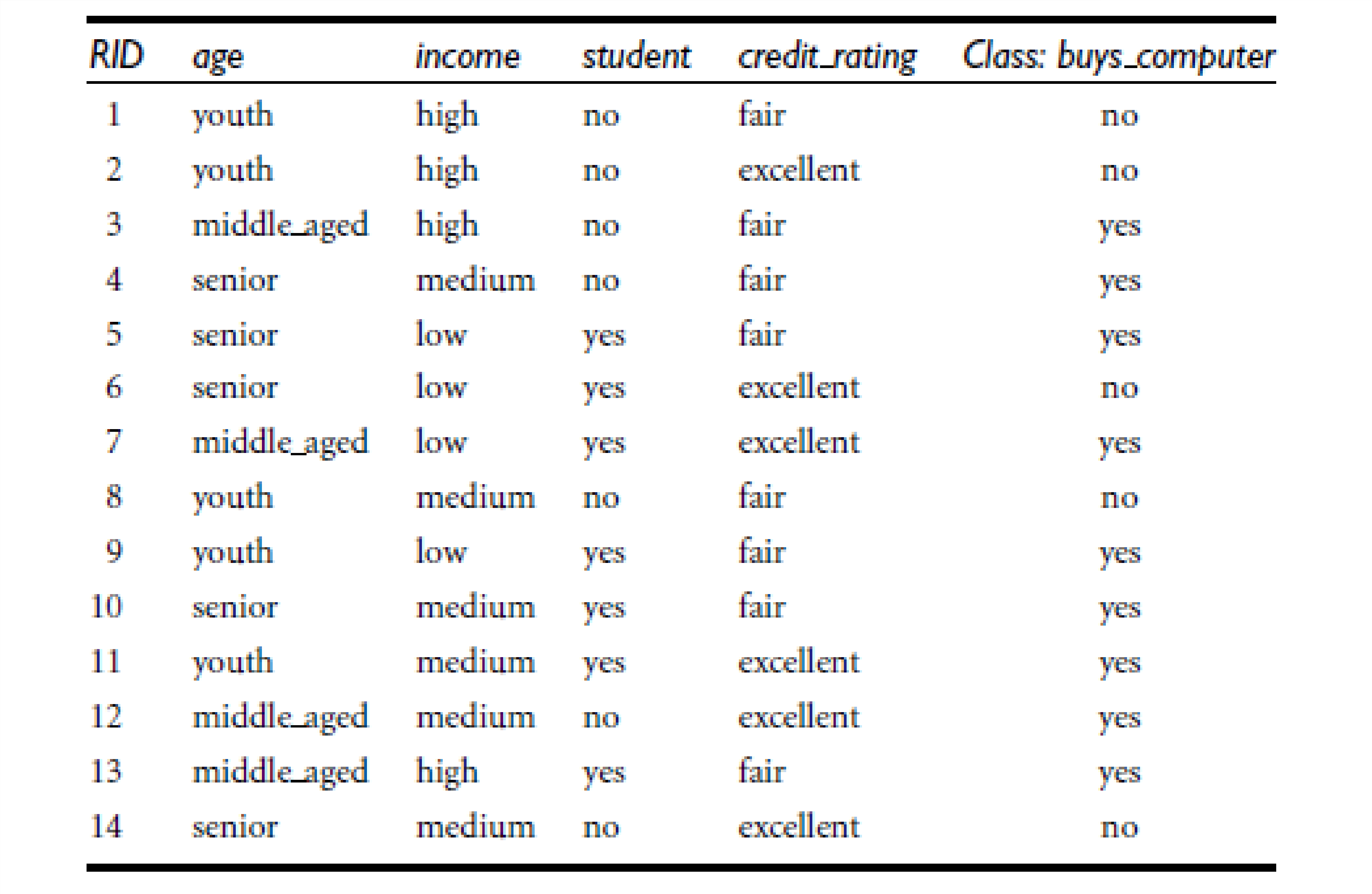
$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \quad \text{for } 1 \le j \le m, j \ne i.$$

Bayes’ theorem

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

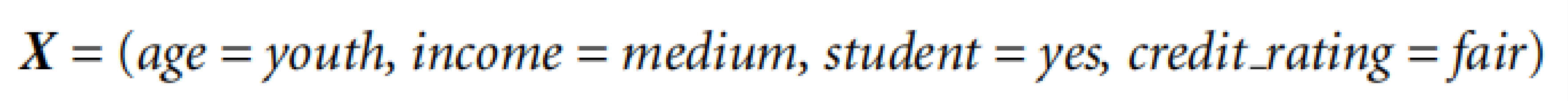
$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i). \end{aligned}$$

Naive Bayes - Example



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

- Classify the following x,



$$X = (age = youth,\ income = medium,\ student = yes,\ credit\_rating = fair)$$

- Prior probablity

P(buys computer = yes) = 9/14 = 0.643

P(buys computer = no) = 5/14 = 0.357

Conditional Probablity

- P(age = youth | buys computer = yes) = 2/9 = 0.222
- P(age = youth | buys computer = no) = 3/5 = 0.600
- P(income = medium | buys computer = yes) = 4/9 = 0.444
- P(income = medium | buys computer = no) = 2/5 = 0.400
- P(student = yes | buys computer = yes) = 6/9 = 0.667
- P(student = yes | buys computer = no) = 1/5 = 0.200
- P(credit rating = fair | buys computer = yes) = 6/9 = 0.667
- P(credit rating = fair | buys computer = no) = 2/5 = 0.400

- P(X|buys computer = yes) =

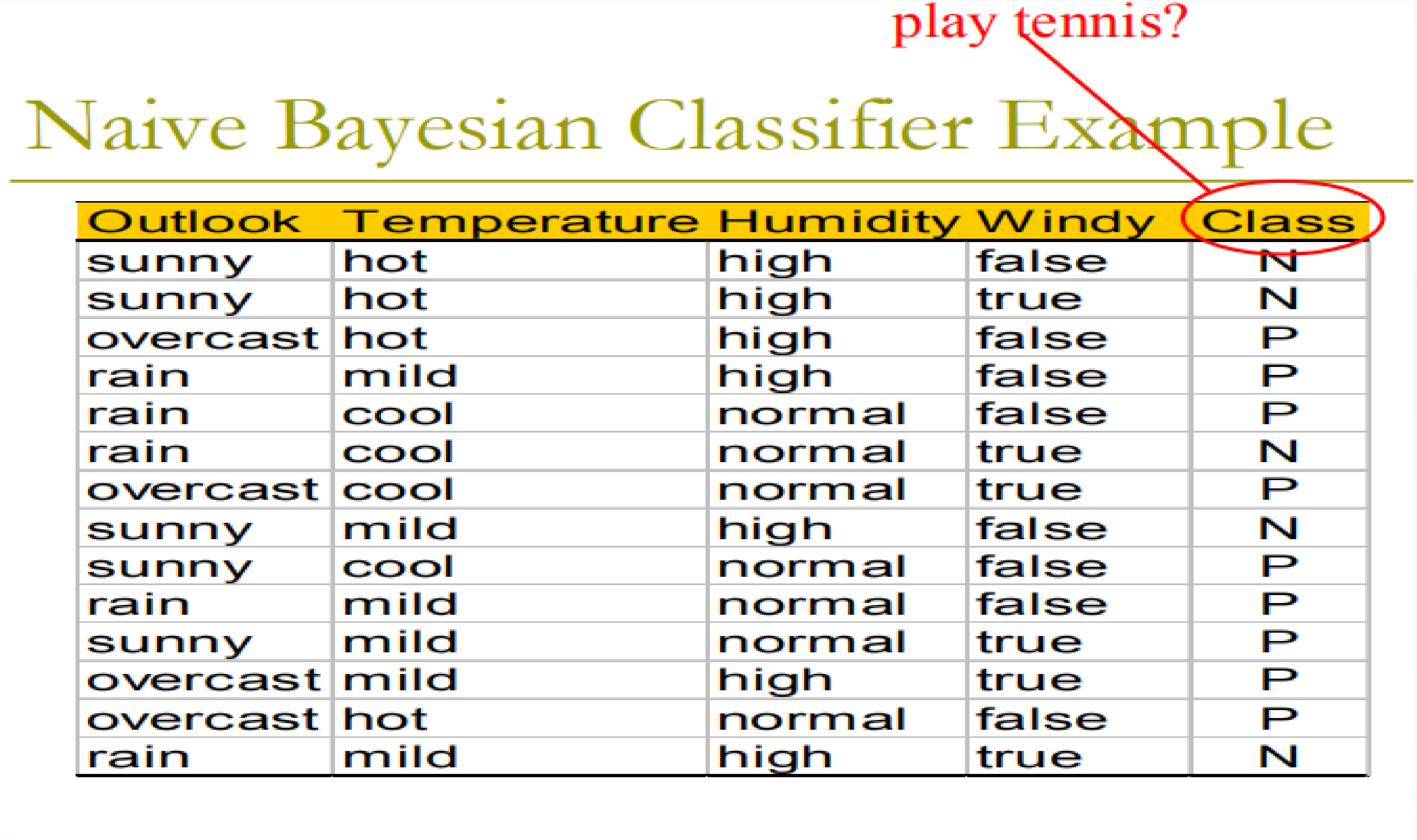
P(age = youth | buys computer = yes) ×
P(income = medium | buys computer = yes) ×
P(student = yes | buys computer = yes) ×
P(credit rating = fair | buys computer = yes)

$= 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.$

- P(X|buys computer = no) = 0.600 × 0.400 × 0.200 × 0.400 = 0.019.
- To find the class, Ci, that maximizes P(X|Ci)P(Ci), Compute
- P(X|buys computer = yes)P(buys computer = yes) = 0.044 × 0.643
  = 0.028
- P(X|buys computer = no)P(buys computer = no) = 0.019 × 0.357
  = 0.007

Naïve Bayesian classifier predicts buys computer = yes for tuple X.

Practice Question???



play tennis?

Naive Bayesian Classifier Example

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Classify a new sample X

X =

- „outlook = sunny
- „temperature = cool
- „humidity = high
- „windy = false

- Play Tennis = ?

Reference:

Data Mining: Concepts and Techniques , Jiawei Han and Micheline Kamber