Bayesian Classifier

Bayesian Classifier

- Statistical Classifier
- Predict Class Membership Probabilities
- Based on Bayes Theorem
- High Accuracy and Speed in large Databases
- Prior Probability: *Probability of X, P(X)*
- Posterior Probability: *Probability of X when consider some condition Y, P(X|Y)*
- A Simple Bayesian Classifier Naive Bayesian Classifier
    - Class Conditional Independence

MLE vs MAP

Maximum Likelihood Estimation

$$\begin{aligned}\theta_{MLE} &= \arg\max_{\theta} f(X_1, X_2, \ldots, X_n \,|\, \theta) \\ &= \arg\max_{\theta} \sum_i \log f(X_i \,|\, \theta)\end{aligned}$$

Maximum A Priori

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} f(\theta \,|\, X_1, X_2, \ldots, X_n) \\ &= \arg\max_{\theta} \left(\log g(\theta) + \sum_i \log f(X_i \,|\, \theta) \right)\end{aligned}$$

Conditional Independence

A and B are independent if,

$$P(A \cap B) = P(A) \times P(B)$$
$$\forall_{a,b} : P(A = a \cap B = b) = P(A = b) \times P(B = b)$$

A and B are conditionally independent given C if,

$$P(A, B \,|\, C) = P(A \,|\, C) \times P(B \,|\, C)$$
$$\forall_{a,b,c} : P(A = a \cap B = b \,|\, C = c) = P(A = b \,|\, C = c) \times P(B = b \,|\, C = c)$$

Bayes Theorem

Likelihood

Prior Probability

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

Normalization Constant

Naive Bayes Classifier

- Attributes are conditionally independent
- Consider *n*-dimensional attribute vector, $X = (X_1, X_2, \ldots, X_n)$
- Consider m classes, $C = (C_1, C_2, \ldots, C_m)$
- Naive Bayes predicts that a tuple belongs to some class C_i

  for given condition ***X*** if and only if,

$$P(C_i \mid X) > P(C_j \mid X)\, for\, 1 \leq j \leq m,\ j \neq i$$

- Goal: maximize $P(C_i \mid X)$

Naive Bayes Equations

Likelihood: $$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$
$$= P(X_1 \mid C_i) \times P(X_2 \mid C_i) \times \ldots \times P(X_n \mid C_i)$$

Categorical Attribute:

$$P(X_k \mid C_i) = \frac{|D_{X_k, C_i}|}{D_{C_i}}$$

Continuous-valued Attribute:
Gaussian Distribution

$$P(X_k \mid C_i) = g(X_k, \mu_{C_i}, \sigma_{C_i})$$
$$g(X, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Sample Data: D

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes

ID	age	income	student	creditRating	buyComputer
1	youth	low	no	fair	???

Sample Data: Problem Definition

ID	age	income	student	creditRating	buyComputer
1	youth	low	no	fair	???

Here,

$$X = (age = youth, income = low, student = no, creditRating = fair)$$

Find, $$= \max(P(C_{yes} \mid X), P(C_{no} \mid X))$$

$$= \max\left(\frac{P(X \mid C_{yes})P(C_{yes})}{P(X)}, \frac{P(X \mid C_{no})P(C_{no})}{P(X)}\right)$$

$$= \max(P(X \mid C_{yes})P(C_{yes}), P(X \mid C_{no})P(C_{no}))$$

Sample Data: P(buyComputer=yes|X)

$$= \max(P(X \mid C_{yes})P(C_{yes}), P(X \mid C_{no})P(C_{no}))$$

ID	age	income	student	creditRating	buyComputer
1	youth	low	no	fair	???

Necessary Statistics,

$$P(buyComputer = yes)$$

$$P(age = youth \mid buyComputer = yes)$$

$$P(income = low \mid buyComputer = yes)$$

$$P(student = no \mid buyComputer = yes)$$

$$P(creditRating = fair \mid buyComputer = yes)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=yes|X)

Necessary Statistics,

$$P(buyComputer = yes) \quad \boxed{= \frac{9}{14}}$$

$$P(age = youth \mid buyComputer = yes)$$

$$P(income = low \mid buyComputer = yes$$

$$P(student = no \mid buyComputer = yes$$

$$P(creditRating = fair \mid buyCompute$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=yes|X)

Necessary Statistics,

$$P(buyComputer = yes) \boxed{= \frac{9}{14}}$$

$$P(age = youth \mid buyComputer = yes) \boxed{= \frac{2}{9}}$$

$$P(income = low \mid buyComputer = yes$$

$$P(student = no \mid buyComputer = yes$$

$$P(creditRating = fair \mid buyCompute$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=yes|X)

Necessary Statistics,

$$P(buyComputer = yes) \boxed{= \frac{9}{14}}$$

$$P(age = youth \,|\, buyComputer = yes) \boxed{= \frac{2}{9}}$$

$$P(income = low \,|\, buyComputer = yes \boxed{= \frac{2}{9}}$$

$$P(student = no \,|\, buyComputer = yes$$

$$P(creditRating = fair \,|\, buyCompute$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=yes|X)

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Necessary Statistics,

$$P(buyComputer = yes) \quad \boxed{= \frac{9}{14}}$$

$$P(age = youth \mid buyComputer = yes) \quad \boxed{= \frac{2}{9}}$$

$$P(income = low \mid buyComputer = yes \quad \boxed{= \frac{2}{9}}$$

$$P(student = no \mid buyComputer = yes \quad \boxed{= \frac{3}{9}}$$

$$P(creditRating = fair \mid buyCompute$$

Sample Data: P(buyComputer=yes|X)

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Necessary Statistics,

$$P(buyComputer = yes) \boxed{= \frac{9}{14}}$$

$$P(age = youth \,|\, buyComputer = yes) \boxed{= \frac{2}{9}}$$

$$P(income = low \,|\, buyComputer = yes \boxed{= \frac{2}{9}}$$

$$P(student = no \,|\, buyComputer = yes \boxed{= \frac{3}{9}}$$

$$P(creditRating = fair \,|\, buyCompute \boxed{= \frac{4}{9}}$$

Sample Data: P(buyComputer=yes|X)

Necessary Statistics,

$$P(buyComputer = yes) \quad \boxed{= \frac{9}{14}}$$

$$P(age = youth \,|\, buyComputer = yes) \quad \boxed{= \frac{2}{9}}$$

$$P(income = low \,|\, buyComputer = yes) \quad \boxed{= \frac{2}{9}}$$

$$P(student = no \,|\, buyComputer = yes) \quad \boxed{= \frac{3}{9}}$$

$$P(creditRating = fair \,|\, buyComputer = yes) \quad \boxed{= \frac{4}{9}}$$

$$\begin{aligned} P(X \,|\, buyComputer = yes) \times P(buyComputer = yes) &= \frac{2}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{4}{9} \times \frac{9}{14} \\ &= \frac{8}{1701} \end{aligned}$$

Sample Data: P(buyComputer=no|X)

ID	age	income	student	creditRating	buyComputer
1	youth	low	no	fair	???

Necessary Statistics,

$$P(buyComputer = no)$$

$$P(age = youth \mid buyComputer = no)$$

$$P(income = low \mid buyComputer = no$$

$$P(student = no \mid buyComputer = no$$

$$P(creditRating = fair \mid buyComputer = no)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=no|X)

Necessary Statistics,

$$P(buyComputer = no) \quad \boxed{= \frac{5}{14}}$$

$$P(age = youth \mid buyComputer = no)$$

$$P(income = low \mid buyComputer = no)$$

$$P(student = no \mid buyComputer = no)$$

$$P(creditRating = fair \mid buyComputer = no)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=no|X)

Necessary Statistics,

$$P(buyComputer = no) \quad \boxed{= \frac{5}{14}}$$

$$P(age = youth \,|\, buyComputer = no) \quad \boxed{= \frac{3}{5}}$$

$$P(income = low \,|\, buyComputer = no)$$

$$P(student = no \,|\, buyComputer = no)$$

$$P(creditRating = fair \,|\, buyComputer = no)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=no|X)

Necessary Statistics,

$$P(buyComputer = no) \quad \boxed{= \frac{5}{14}}$$

$$P(age = youth \mid buyComputer = no) \quad \boxed{= \frac{3}{5}}$$

$$P(income = low \mid buyComputer = no) \quad \boxed{= \frac{2}{5}}$$

$$P(student = no \mid buyComputer = no)$$

$$P(creditRating = fair \mid buyComputer = no)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Sample Data: P(buyComputer=no|X)

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Necessary Statistics,

$$P(buyComputer = no) \quad \boxed{= \frac{5}{14}}$$

$$P(age = youth \mid buyComputer = no) \quad \boxed{= \frac{3}{5}}$$

$$P(income = low \mid buyComputer = no) \quad \boxed{= \frac{2}{5}}$$

$$P(student = no \mid buyComputer = no) \quad \boxed{= \frac{3}{5}}$$

$$P(creditRating = fair \mid buyComputer = no) \quad \boxed{= \frac{2}{5}}$$

Sample Data: P(buyComputer=no|X)

Necessary Statistics,

$$P(buyComputer = no) \boxed{= \frac{5}{14}}$$

$$P(age = youth \mid buyComputer = no) \boxed{= \frac{3}{5}}$$

$$P(income = low \mid buyComputer = no) \boxed{= \frac{2}{5}}$$

$$P(student = no \mid buyComputer = no) \boxed{= \frac{3}{5}}$$

$$P(creditRating = fair \mid buyComputer = no) \boxed{= \frac{2}{5}}$$

$$\begin{aligned} P(X \mid buyComputer = no) \times P(buyComputer = no) &= \frac{3}{5} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{5}{14} \\ &= \frac{18}{875} \end{aligned}$$

Sample Data: Prediction

$$P(X \mid buyComputer = yes) \times P(buyComputer = yes) < P(X \mid buyComputer = no) \times P(buyComputer = no)$$

$$buyComputer = no$$

$$\begin{aligned} P(X \mid buyComputer = no) \times P(buyComputer = no) &= \frac{3}{5} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{5}{14} \\ &= \frac{18}{875} \end{aligned}$$

$$\begin{aligned} P(X \mid buyComputer = yes) \times P(buyComputer = yes) &= \frac{2}{9} \times \frac{2}{9} \times \frac{3}{9} \times \frac{4}{9} \times \frac{9}{14} \\ &= \frac{8}{1701} \end{aligned}$$

Zero Probability

$$= P(age = middleaged \mid buyComputer = no)$$
$$= \frac{0}{5}$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
1[illegible]	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

$$X = (age = middleaged, income = low, student = no, creditRating = fair)$$

Laplacian Correction

Add an extra count for each different value.

For attribute age,

- ➔ *count(youth | no) = 4 (Actually 3)*
- ➔ *count(middleaged | no) = 1 (Actually 0)*
- ➔ *count(senior | no) = 3 (Actually 2)*
- ➔ *count(no) = 8 (Actually 5)*

$$= P(age = middleaged \mid buyComputer = no)$$
$$= \frac{1}{8}$$

$$X = (age = middleaged, income = low, student = no, creditRating = fair)$$

ID	age	income	student	creditRating	buyComputer
1	youth	high	yes	fair	no
2	youth	low	no	excellent	no
3	middleaged	medium	no	fair	yes
4	middleaged	medium	no	fair	yes
5	senior	medium	yes	fair	yes
6	senior	low	no	excellent	no
7	midleaged	high	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	high	yes	fair	yes
11	youth	high	yes	excellent	yes
12	middleaged	medium	no	excellent	yes
13	middleaged	low	yes	fair	yes
14	senior	high	yes	excellent	no

Thank You