Bayes Classification

Shubham kumar
Dept. of CS&IT
MGCUB

- Bayesian classifier can predict the probability that a given tuple belongs to a particular class.
- It is based on Bayes' theorem.
- Bayesian classifier gives more speed and accuracy as compared to Decision tree classifier.
- First we recall the Bayes' theorem in probability then we will understand the working of Bayesian classification method.

- Bayes' theorem:

Let X be a data tuple which is described by measurements made on a set of n attributes.

[Note : In Bayesian term X is also called evidence]

Let H be some hypothesis.

We assume the hypothesis H: the data tuple X belongs to a specified class C.

Then,

for classification problems, we have to find P(H/X) i.e. the probability of hypothesis H when X is given.

Example:

Suppose we have employee database consists of several attributes like department, age and salary, etc. And status is the class label attribute which is either senior or junior.

Let X is the detail of an employee.

i.e. X=(department= sales, age= 35, salary=40K)

Let the Hypothesis H: employee belongs to senior class.

Then,

[1] Posterior probability:

P(H/X) is the probability that the employee X will belongs to senior class given that we know the employee's department, age and salary.

P(X/H) is the probability that an employee, X, is of sales department, with age 35 and have salary 40K, given that, it belongs to senior class.

[2] Prior probability:

P(H) is the probability that any given employee will belong to the senior class regardless of any information like department , age etc.

P(X) is the probability of an employee whose age is 35 years, department is sales and the salary is 40K.

The posterior probability can be calculated using Bayes' theorem as:

$$P(H/X) = \frac{P(X/H).\ P(H)}{P(X)}$$

Bayesian classification method:

Let D is the training data set.

Each tuple in D is represented by an n- dimensional attribute vector, X =(x1, x2, x3, ...,xn) where x1, x2, x3….., xn are the values of attributes A1, A2, A3, ..., An respectively.

Let there are m classes, C1, C2, C3, ..., Cm.

Then,

The Bayesian classifier predicts that a given tuple X belongs to the class C_i if and only if the posterior probability $P(C_i / X)$ is highest.

i.e.

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

By Bayes' theorem-

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

If the class prior probabilities $P(C_i)$ are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1)= P(C2) = P(C3)= .... = P(Cm).

Our goal is to maximize $P(C_i / X)$.

- If the data sets have many attributes then it would be expensive to compute $P(X/C_i)$. To reduce this computation naïve assumption is made.
- Naïve Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes.
- This assumption is called **class- conditional independence.**

- Now, since there are no dependence relationships among the attributes.

therefore,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i).$$

It is easy to calculate $P(x1/C_i)$, $P(x2/C_i)$, . ., $P(xn/C_i)$ from the training tuples.

Hence, to predict the class label of X, $P(X/C_i)$. $P(C_i)$ is evaluated for each class C_i and the maximum one will be the predicted class label.

Example: Let us take the training data set D as:

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

The data tuples are described by the attributes age, income, student, and credit_rating. The class label attribute buys_computer, has two distinct values [yes, no].

Let C1 if for class yes and C2 is for class no.

The tuple w wish to classify is;

X= (age= youth, income= medium, student= yes, credit_rating= fair)

We have to maximize $P(X/C_i). P(C_i)$, for i= 1, 2.

The prior probability of each class can be computed as:

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

To compute $P(X|C_i)$, for $i = 1, 2$, we compute the following conditional probabilities:

$$P(age = youth \mid buys\_computer = yes) = 2/9 = 0.222$$
$$P(age = youth \mid buys\_computer = no) = 3/5 = 0.600$$
$$P(income = medium \mid buys\_computer = yes) = 4/9 = 0.444$$
$$P(income = medium \mid buys\_computer = no) = 2/5 = 0.400$$
$$P(student = yes \mid buys\_computer = yes) = 6/9 = 0.667$$

$$P(student = yes \mid buys\_computer = no) = 1/5 = 0.200$$
$$P(credit\_rating = fair \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(credit\_rating = fair \mid buys\_computer = no) = 2/5 = 0.400$$

Using these probabilities, we obtain

$$\begin{aligned} P(\boldsymbol{X}|buys\_computer = yes) = {} & P(age = youth \mid buys\_computer = yes) \\ & \times P(income = medium \mid buys\_computer = yes) \\ & \times P(student = yes \mid buys\_computer = yes) \\ & \times P(credit\_rating = fair \mid buys\_computer = yes) \\ = {} & 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044. \end{aligned}$$

Similarly,

$$P(\boldsymbol{X}|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019.$$

To find the class, C_i, that maximizes $P(\boldsymbol{X}|C_i)P(C_i)$, we compute

$$P(\boldsymbol{X}|buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(\boldsymbol{X}|buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

Therefore, the naïve Bayesian classifier predicts buys_computer= yes for given tuple X.

Reference

Jiawei Han, Micheline kamber and Jian pei. “DATA MINING concepts and Techniques” 3/e, Elsevier, 2012