BAYESIAN LEARNING

Bayesian Classifiers

Bayesian classifier is statistical classifier, and are based on Bayes theorem

They can calculate the probability that a given sample belongs to a particular class

Its results are comparable to the performance of other classifiers, such as decision tree and neural networks in many cases

BAYESIAN LEARNING

Bayes Theorem

Let *X* be a data sample, e.g. red and round fruit

Let *H* be some hypothesis, such as that *X* belongs to a specified class *C* (e.g. X is an apple)

For classification problems, we want to determine *P(H|X)*, the probability that the hypothesis *H* holds given the observed data sample *X*

BAYESIAN LEARNING

Prior & Posterior Probability

The probability *P(H)* is called the prior probability of *H*, i.e the probability that any given data sample is an apple, regardless of how the data sample looks

The probability *P(H|X)* is called posterior probability. It is based on more information, then the prior probability *P(H)* which is independent of *X*

BAYESIAN LEARNING

Bayes Theorem

It provides a way of calculating the posterior probability

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

$P(X|H)$ is the posterior probability of X given H (it is the probability that X is red and round given that X is an apple)

$P(X)$ is the prior probability of X (probability that a data sample is red and round)

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

It works as follows:

1. **Each data sample is represented by an n-dimensional feature vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the sample from n attributes, respectively $A_1, A_2, \ldots A_n$**

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

2. Suppose that there are m classes C_1, C_2, … C_m. Given an unknown data sample, X (i.e. having no class label), the classifier will predict that X belongs to the class having the highest posterior probability given X

Thus if $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$
then X is assigned to C_i

This is called Bayes decision rule

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

3. We have $P(C_i|X) = P(X|C_i)\ P(C_i)\ /\ P(X)$

As $P(X)$ is constant for all classes, only $P(X|C_i)\ P(C_i)$ needs to be calculated

The class prior probabilities may be estimated by

$$P(C_i) = s_i / s$$

where s_i is the number of training samples of class C_i

& s is the total number of training samples

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$

For example, assuming the attributes of colour and shape to be Boolean, we need to store 4 probabilities for the category apple

$$P(\neg red \wedge \neg round \mid apple)$$
$$P(\neg red \wedge round \mid apple)$$
$$P(red \wedge \neg round \mid apple)$$
$$P(red \wedge round \mid apple)$$

If there are 6 attributes and they are Boolean, then we need to store 2^6 probabilities

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

In order to reduce computation, the naïve assumption of *class conditional independence* is made

This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample (we assume that there are no dependence relationships among the attributes)

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

Thus we assume that $P(X|C_i) = \prod^n_{k=1} P(x_k|C_i)$

Example
$P(\text{colour} \wedge \text{shape} \mid \text{apple}) = P(\text{colour} \mid \text{apple})\, P(\text{shape} \mid \text{apple})$

For 6 Boolean attributes, we would have only 12 probabilities to store instead of $2^6 = 64$
Similarly for 6, three valued attributes, we would have 18 probabilities to store instead of 3^6

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$, …, $P(x_n|C_i)$ can be estimated from the training samples, where

For an attribute A_k, which can take on the values x_{1k}, x_{2k}, …
e.g. colour = red, green, …

$$P(x_k|C_i) = s_{ik}/s_i$$

where s_{ik} is the number of training samples of class C_i having the value x_k for A_k
and s_i is the number of training samples belonging to C_i

e.g. P(red|apple) = 7/10 if 7 out of 10 apples are red

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

Example:

rid	age	income	student	credit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	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	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

Example:

Let C_1 = class buy computer and C_2 = class not buy computer

The unknown sample:
X = {age = "< 30", income = medium, student = yes, credit-rating = fair}

The prior probability of each class can be computed as

P(buy computer = yes) = 9/14 = 0.643
P(buy_computer = no) = 5/14 = 0.357

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

Example:

To compute P(X|Ci) we compute the following conditional probabilities

$$
\begin{aligned}
&P(age = \text{"}<30\text{"} \mid buys\_computer = yes) &= 2/9 = 0.222\\
&P(age = \text{"}<30\text{"} \mid buys\_computer = no) &= 3/5 = 0.600\\
&P(income = medium \mid buys\_computer = yes) &= 4/9 = 0.444\\
&P(income = medium \mid buys\_computer = no) &= 2/5 = 0.400\\
&P(student = yes \mid buys\_computer = yes) &= 6/9 = 0.667\\
&P(student = yes \mid buys\_computer = no) &= 1/5 = 0.200\\
&P(credit\_rating = fair \mid buys\_computer = yes) &= 6/9 = 0.667\\
&P(credit\_rating = fair \mid buys\_computer = no) &= 2/5 = 0.400
\end{aligned}
$$

BAYESIAN LEARNING

Naïve (Simple) Bayesian Classification

Example:

Using the above probabilities, we obtain

$$P(X|buys\_computer = yes) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X|buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$$

And hence the naïve Bayesian classifier predicts that the student will buy computer, because

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

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

BAYESIAN LEARNING

Gaussian Naïve Bayes Classifier

Numerical variables need to be transformed to their categorical counterparts (binning) before constructing their frequency tables. The other option we have is using the distribution of the numerical variable to have a good guess of the frequency. For example, one common practice is to assume normal distributions for numerical variables.

The probability density function for the normal distribution is defined by two parameters (mean and standard deviation).

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$ Mean

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$ Standard deviation

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ Normal distribution

Example:

		Humidity	Mean	StDev
Play Golf	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
	no	85 90 70 95 91	86.2	9.7

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

BAYESIAN LEARNING

The Zero-Frequency Problem

- One of the disadvantages of Naïve Bayes is that if you have no occurrences of a class label and a certain attribute value together then the frequency-based probability estimate will be zero. And this will get a zero when all the probabilities are multiplied.

BAYESIAN LEARNING

Laplace smoothing or correction for handling zero frequency problem

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	3/5

Temp	Yes	No
Hot	2/9	3/5
Mild	4/9	2/5
Cool	3/9	2/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Classify new example

(outlook = Overcast, temp = Cool,

Humidity = High, Wind =Strong)

$p(PlayTennis = yes) = \frac{9}{14}$ $p(PlayTennis = no) = \frac{5}{14}$

$p(yes|new\ Instance)$

$= p(yes) * p(Outlook = Overcast|yes) * p(Temp = cool|yes) * p(Hum = high|yes) * p(Wind = strong|yes)$

$p(no|new\ Instance)$

$= p(no) * p(Outlook = Overcast|no) * p(Temp = cool|no) * p(Hum = high|no) * p(Wind = strong|no)$

BAYESIAN LEARNING

Laplace smoothing or correction for handling zero frequency problem

- Laplace smoothing is a smoothing technique that handles the problem of zero probability in Naïve Bayes.

$$P_{LAP,k}(x|y) = \frac{c(x,y)+k}{c(y)+k|X|}$$

- K represents the smoothing parameter (greater than zero)
- X represent number of different values x can have

$$p(outlook = overcast|no) = \frac{0+1}{5+1*3} = \frac{1}{8}$$