Module No : 03

Classification

Bayes Classification Method : Naive Bayes classification

- Bayesian classifiers are statistical classifiers.
- They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
- A simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree
- Naive Bayesian classifiers assume 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.
- It is made to simplify the computations involved and, in this sense, is considered “naive.”

Q. Predict a class label using naive Bayes classification

The tuple we wish to classify is

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

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

Probabilities of Class Attribute:

P(buys_comp) / yes	
P(buys_comp) / no	

Probabilities of Age Attribute:

P(youth/yes)=	P(youth/no)=
P(m_aged/yes)=	P(m_aged/no)=
P(senior/yes)=	P(senior/no)=

Probabilities of income attribute

P(high/yes)=	P(high/no)=
P(medium/yes)=	P(medium/no)=
P(low/yes)=	P(low/no)=

Probabilities of attribute student

P(yes/yes)=	P(yes/no)=
P(no/yes)=	P(no/no)=

RID	age	income	student	credit_rating	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

Probabilities of attribute credit rating

P(fair/yes)=	P(fair/no)=
P(excellent/yes)=	P(excellent/no)=

The tuple we wish to classify is

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

buys_comp — age — income — student — credit_rating

P(X/yes) = P(youth/yes) * P(medium/yes) * P(yes/yes) * P(fair/yes)

**= * * * **

=

buys_comp — age — income — student — credit_rating

P(X/no) = P(age/no) * P(medium/no) * P(yes/no) * P(fair/no)

**= * * * **

=

The tuple we wish to classify is

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

buys_comp: age, income, student, credit_rating

$$P(X/yes) = P(youth/yes) * P(medium/yes) * P(yes/yes) * P(fair/yes)$$

$$= 0.222 * 0.444 * 0.667 * 0.667$$

$$= 0.044$$

buys_comp: age, income, student, credit_rating

$$P(X/no) = P(age/no) * P(medium/no) * P(yes/no) * P(fair/no)$$

$$= 0.600 * 0.400 * 0.200 * 0.400$$

$$= 0.019$$

P(X/yes) = P(youth/yes) * P(medium/yes) * P(yes/yes) * P(fair/yes) = 0.044

P(X/no) = P(age/no) * P(medium/no) * P(yes/no) * P(fair/no) = 0.019

and

Probabilities of Class Attribute:

P(buys_comp) / yes	**0.643**
P(buys_comp) / no	**0.357**

to find the final result, we compute

P(x/yes) * P(buys_comp/yes)=0.044*0.643=0.028

P(x/no) * P(buys_comp/no) =0.019*0.357=0.007

Therefore, the naive Bayesian classifier predicts buys computer = yes for tuple X

Predict the class label using naive bayes classifier

X=(outlook=*sunny*,temp=*mild*,humidity=*high*,windy=*true*)

No.	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P(PlayTennis = yes) = 9/14 = .64$$

$$P(PlayTennis = no) = 5/14 = .36$$

NAIVE BAYES CLASSIFIER

Example - 1

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5
Windy		
Strong	3/9	3/5
Weak	6/9	2/5

NAIVE BAYES CLASSIFIER – Example -1

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

$$v_{NB}(yes) = P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$v_{NB}(no) = P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = 0.205 \qquad v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = 0.795$$

Model evaluation and selection

- Now we know what is classification, how classifiers works so we may built a classification model
- For example, suppose you used previous sales data to build a classifier to predict customer purchasing behaviour
- In this example we would like to analyse how our model can predict the purchasing behaviour of future customers.(data on which classifier has not been trained)
- We may built different classifiers and we can compare their accuracy/performance by applying various evaluation matrics
- Before we discuss the various evaluation matrics, we need to understand some basic terminologies

Model evaluation and selection

- MODEL : a model is created by applying an algorithms(or statistical calculations) to data to generate predictions/classifications of new data.
- Given data set is partitioned into subsets
  - Training data set
  - Testing data set
- Training data set: training data set is used to derive the model or train the model
- Testing data set: the models accuracy is estimated by using testing data set

Model evaluation and selection

- Positive tuples : positive tuples of the class attribute (in our last example positive tuples are *buys_computer= yes)*
- Negative tuples : negative tuples of the class attribute (in our last example negative tuples are *buys_computer= no)*
- Suppose we use our classifier on a test set of labeled tuples.
- P is the number of positive tuples and N is the number of negative tuples.
- For each tuple, we compare the classifier's class attribute prediction with the tuple's known class attribute value.

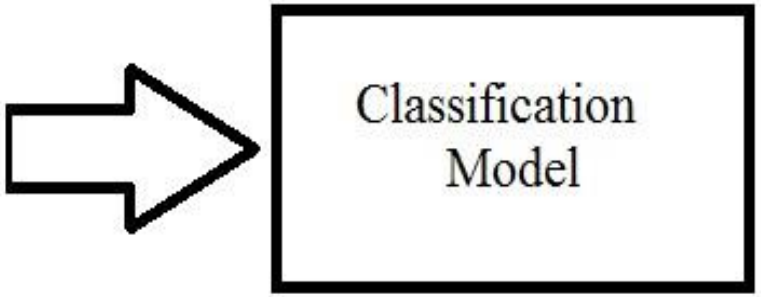
Model evaluation and selection

There are four additional terms we need to know that are

- True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- True negatives (TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives.
- False positives (FP) Type I Error: These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class *buys_computer=no* for which the classifier predicted *buys_computer=yes*). Let FP be the number of false positives.
- False negatives (FN) Type II Error: These are the positive tuples that were mislabeled as negative (e.g., tuples of class *buys_computer=yes* for which the classifier predicted *buys_computer=no*). Let FN be the number of false negatives.

Labeled Samples:

RID	age	income	student	credit_rating	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



RID	age	income	student	credit_rating	buys_computer
1	youth	high	no	fair	no yes
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes no
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no yes
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes no
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 no
14	senior	medium	no	excellent	no

TP= TN= FP= FN=