University of Roehampton London

CMP020L012A	Data Analytics	2022/23

Week-9

Naïve Bayes

Solution

Answer the following Multiple-Choice Questions

1. Naïve bayes is a _____ algorithm.
   a) Classification
   b) Clustering
   c) Regression
   d) All of the above

2. Which from the following is/are the assumption in Naïve Bayes Classifier?
   a) All the classes are independent of each other
   b) All the predictors of the datapoints in the dataset are independent to each other
   c) Both a) and b)
   d) None

3. Naïve Bayes work exceptionally good on?
   a) Qualitative Values
   b) Quantitative Values
   c) Either a) or b)
   d) None

4. A probabilistic model of data within each class is an example of ____ .
   a) Discriminative Classification
   b) Probabilistic Classification
   c) Both above
   d) None of the above

5. Examples of some applications of the Naïve Bayes Algorithm is/are:
   a) Sentiment Analysis
   b) Classifying Articles
   c) Spam filtering from email
   d) All of the above

6. Which of the following statement(s) about Naïve Bayes is/are true?
   a) It can be used for Binary Classification problem
   b) It can be used for Multiple Classification problem
   c) Both a) and b)
   d) None of the above

7. Which of the following is not correct about Naïve Bayes?
   a) All the attributes (predictors) are equally important.
   b) Attributes are statistically dependent of each other
   c) Attributes are statistically independent of each other
   d) Attributes (predictors) can be numeric or, nominal

8. The prior probability is estimated on the "Label"/ "Class" value of the dataset in Naïve Bayes.
   a) True
   b) False

9. The Class/label value is mutually exclusive and exhaustive, is ____.
   a) True
   b) False

10. The qualitative attribute in the test set, which is not observed in the training set can cause 0 "Zero" probability.
    a) True
    b) False

University of Roehampton London

Answer the following Question

The following training set contains the data about a particular train in London being ontime/ late/ very late/ even getting cancelled due to some weather conditions and different days of the month that are the predictors of this dataset.

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Now, you are to find the most likely classification for the new unseen/new datapoint as given below:

weekday	winter	high	heavy	????

University of Roehampton London

Answer:

To find the posterior probability, we should first determine the Prior and the Conditional probabilities required for the new/unseen datapoint.

	class = on time	class = late	class = very late	class = cancelled
day = weekday	**9/14 = 0.64**	**1/2 = 0.5**	**3/3 = 1**	**0/1 = 0**
day = saturday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	1/1 = 1
day = sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
day = holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
season = spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1/1 = 1
season = summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
season = autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
season = winter	**2/14 = 0.14**	**2/2 = 1**	**2/3 = 0.67**	**0/1 = 0**
wind = none	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
wind = high	**4/14 = 0.29**	**1/2 = 0.5**	**1/3 = 0.33**	**1/1 = 1**
wind = normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
rain = none	5/14 = 0.36	1/2 =0.5	1/3 = 0.33	0/1 = 0
rain = slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
rain = heavy	**1/14 = 0.07**	**1/2 = 0.5**	**2/3 = 0.67**	**1/1 = 1**
Prior Probability	**14/20 = 0.70**	**2/20 = 0.10**	**3/20 = 0.15**	**1/20 = 0.05**

So, the posterior probabilities for the respective class values are:

Ontime: 0.70 X 0.64 X 0.14 X 0.29 X 0.07 = 0.0013

Late: 0.10 X 0.50 X 1.00 X 0.50 X 0.50 = 0.0125

Very Late: 0.15 X 1.00 X 0.67 X 0.33 X 0.67 = **0.0222**

Cancelled: 0.05 X 0.00 X 0.00 X 1.00 X 1.00 = 0.00

Based on the largest probability value for the new datapoint, the class is determined as:

Very Late