AMA546 Statistical Data Mining

Exercise 1 Classification

1. If a set of **data** has mean > median > mode, then this set of data is ( ).
   A. Right-skewed
   B. Left-skewed
   C. Bell shape
   D. Symmetric

2. **Naive Bayes** is a special type of Bayesian classifier, with X as the feature variable and C as the class label. One of its assumptions is: ( ).
   A. The prior probabilities of each class are equal.
   B. The prior distribution is a normal distribution with mean 0 and standard deviation $\sqrt{2}/2$.
   C. Given the label, the feature variables X are conditionally independent.
   D. X|C follows a Gaussian distribution.

3. In **logistic regression**, what problems can L1 regularization and L2 regularization solve? ( )
   A. Insufficient data volume
   B. Mismatched training data
   C. Overfitting in training
   D. Training speed is too slow

4. Which measure does the ID3 algorithm use for node classification in constructing **classification trees**?( )
   A. Gini index B. Information gain C. Information gain ratio D. Accuracy

5. Which of the following does not prevent **overfitting**? ( )
   A. Adding a regularization term
   B. Increasing the number of samples
   C. Building a more complex model
   D. Bootstrap resampling

6. ( ) is an observation that differs significantly from other observations to the extent that it is suspected to be generated by a different mechanism.
   A. Boundary point B. Outlier C. Core point D. Centroid

7. The measure for **evaluating** the quality of a classification model is ( ).
   A. Accuracy and recall

B. Accuracy and confidence

C. Accuracy and lift

D. Confidence and lift

8. In general, **KNN** method performs better when ( ) is true.

   A. there are many samples, but they are not typical

   B. there are few samples, but they are typical

   C. the samples are clustered

   D. the samples are distributed in a chain-like manner

9. (Multiple choice) In **classification** problems, we often encounter situations where the number of positive and negative samples is unequal, such as having 100,000 positive samples and only 10,000 negative samples. The most appropriate method for handling this situation is to ( )

   A. repeat the negative samples 10 times to generate 100,000 samples, shuffle the order and participate in classification

   B. perform classification directly, utilizing the data to the fullest extent possible

   C. randomly select 10,000 from the 100,000 positive samples to participate in classification

   D. set the weight of each negative sample to 10 and the weight of each positive sample to 1, participating in the training process

10. (Multiple choice) Which statement about **Random Forest** is incorrect?

    A. Each subtree in RF is independently and identically distributed.

    B. The model variance in RF decreases as the number of subtrees increases.

    C. RF primarily reduces model variance by increasing the correlation between subtrees.

    D. The model bias in RF decreases as the number of subtrees increases.

11. Short answer: Analyze the difference between regression and **classification**.

12. Calculation (**decision tree classifier**): The following table of data sets contains two attributes X and Y, and two class labels "+" and "−". Each attribute takes three different values, 0, 1, or 2. The concept for class "+" is Y=1, and the concept for class "−" is X=0 and X=2.

    (a) Build a decision tree for this dataset. Can this decision tree capture the concepts of ""+" and "−"?

    (b) What are the accuracy, precision, recall, and F1 of the decision tree? (Note that precision, recall, and F1 are all defined with respect to the "+" class.)

X	Y	Num of instances	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

Table 1

(c) Build a new decision tree using the cost function below. Can this new decision tree capture the concept of "+"?

$$C(i,j) = \begin{cases} 0 & i = j \\ 1 & i = +, j = - \\ \frac{\text{Number of } - \text{ instances}}{+\text{Number of instances}} & i = -, j = + \end{cases}$$

(Hint: only need to change the nodes of the original decision tree.)

13. Calculation (**dissimilarity**): Given two tuples (22, 1, 42, 10) and (20, 0, 36, 8):
    (a) Calculate the Euclidean distance between these two objects.
    (b) Calculate the Manhattan distance between these two objects.
    (c) Calculate the Minkowski distance between these two objects using q = 3.

14. Calculation (**attribute selection measures**): The table below shows a training set D of labeled tuples randomly selected from a customer database. In this example, each attribute is discrete-valued, and continuous-valued attributes have been generalized. The class label attribute, buys computer, has two different values (i.e., yes, no), so there are two different classes (i.e., m = 2). Let the C1 class correspond to yes and the C2 class correspond to no.
    (a) Calculate the information gain of attribute "income".
    (b) Calculate the information gain ratio of attribute "income".
    (c) Calculate the Gini index of attribute "income".

15. Calculation (**Naive Bayes**): Given the table data in the previous question, we hope to use the Naive Bayes classifier to predict the class label of an unknown principle. The training data is also the table data in Question 5. The data tuple is described by the attributes "age", "income", "student", and "credit rating", and the class label attribute "buys computer" has two different values. Please use the Naive

RID	Age	Income	Student	Credit_rating	Class
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

Table 2

Bayes method to classify X:

$$Y = (age = youth, Income = medium, Student = yes, Credit\_rating = Fair)$$

16. (**ROC curve**) The tuples in the table below have been sorted in descending order of the probability values returned by the classifier. For each tuple:

Tuple	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.6
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.4

Table 3

(a) Calculate true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

(b) Calculate true positive rate (TPR) and false positive rate (FPR).

(c) Plot the ROC curve for the data.

17. Calculation (**decision tree**): There is a dataset shown in the table below. Please write down the formula to calculate the information gain when splitting the data by attributes A and B (no need to calculate the final result), and explain the role of calculating information gain in classification algorithms.

A	B	Class Label
T	F	*
T	T	*
T	T	*
T	F	#
T	T	*
F	F	#
F	F	#
F	F	#
T		#
T	F	#

Table 4