AMA546 Statistical Data Mining

Assignment 1

1. Which of the following statements about overfitting and underfitting is correct? ( ).

   A. Overfitting is generally characterized by high bias

   B. Underfitting is generally characterized by high variance

   C. Overfitting can be mitigated by reducing the number of variables

   D. Underfitting can be solved by regularization

> **Solution:** C
>
> A. Overfitting is generally characterized by low bias and high variance.
> B. Underfitting is generally characterized by high bias and low variance.
> D. Underfitting can be detected but not solved by regularization.

2. The minimum and the maximum values of attribute income are 12000 yuan and 98000 yuan, respectively. Using the method of maximum/minimum normalization, mapping the values of the attribute to the range of $[0, 1]$. For the income attribute, 73600 yuan will be transformed into ( ).

   A. 0.821

   B. 1.224

   C. 1.458

   D. 0.716

> **Solution:** D
>
> $$\frac{73600 - 12000}{98000 - 12000} = 0.716$$

3. Which distance below focuses on the direction of the vector?( )

   A. Euclidean distance

   B. Hamming distance

   C. Jaccard distance

   D. Cosine distance

Solution: D

See page 157. The cosine similarity does not take the length of the two data objects into account when computing similarity.

4. In the ID3 algorithm, information gain refers to ( ).
   A. The degree of information overflow
   B. The degree of information increase
   C. The degree of entropy increase
   D. The degree of entropy decrease

Solution: D

The information gain is defined as the entropy of the parent node minus the weighted average of the entropy of children nodes, which is the degree of entropy decrease.

5. Which of the following statements about SVM is incorrect?( )
   A. The process of using kernel functions in SVM is essentially a process of feature transformation (feature engineering).
   B. SVM has good classification performance for linearly non-separable data.
   C. Because SVM uses kernel functions, there is no risk of overfitting.
   D. The support vectors in SVM are a few data points.

Solution: C

The SVM uses kernel functions, but it can still overfitting.

6. What is the effect if both L1 and L2 norms are introduced to punish large parameters in the losisic regression? ( )
   A. It can performs variable selection and prevent overfitting to a certain extent
   B. It can solve the problem of dimensionality curse
   C. It can speed up the calculation
   D. It can obtain more accurate results

Solution: A

L1 norm, also refer to the lasso penalty, can performs variable selection. Both L1 and L2 norm can punish large parameters and prevent overfitting to a certain extent.

7. Which two evaluation criteria for classification algorithms do the following two descriptions correspond to, respectively? ( )

   (1) When a police officer catches a thief, it measures how many of the people caught by the police are thieves.

   (2) It measures what proportion of thieves have been caught by the police in total.

   A. Precision, Recall

   B. Recall, Precision

   C. Precision, ROC

   D. Recall, ROC

Solution: A

$Precision = \frac{TP}{TP+FP}$ measures the fraction of the people caught by the police are thieves.
$Recall = \frac{TP}{TP+FN}$ measures proportion of thieves have been caught by the police in total.

8. (Multiple Choice) Suppose a student accidentally duplicated a feature in the training data while using the Naive Bayesian model. Which of the following statements about NB is correct? ( )

   A. The assumption of the Naive Bayesian model has not been violated

   B. The accuracy of the model will decrease compared to the case without duplicate features

   C. The Naive Bayesian model can be used for least squares regression

   D. In this case, the conclusion obtained by the student may be incorrect

Solution: B, D.

A: The assumption of the Naive Bayesian model refers to the conditional independent of the attributes. The duplicated attribute violates the assumption.

C: The Naive Bayesian model has nothing to do with the least squares regression.

9. Analysis Question: Tom is using SVM to build a spam email classifier. If an email is a spam, its label is y=1, otherwise y=0.

(a) List at least three characteristics that Tom can extract from the Email for classification.

(b) In Tom's training set, 99% of the emails are legitimate, and 1% are spam. Suppose this label imbalance causes the trained model to classify all emails as legitimate. What is the accuracy and recall in this case?

(c) If Tom wants to avoid the problem in (2) and train a model that can identify as many spam emails as possible, what should Tom do?

Solution:

(a) Like the email address, the length of the email and the title of the email.

(b) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0.99}{1} = 0.99$
$Recall = \frac{TP}{TP+FN} = \frac{0}{0.01} = 0$

(c) Performing over sampling before building the model.

10. Calculation Question (Naive Bayes Classification): Consider the following training set, where the last column, "Purchase," is the label. Please use the Naive Bayes method to determine whether a **young, low-income, non-student, and medium-credit** customer has a tendency to purchase a computer.

ID	Age	Income	Student	Credit	Purchase
1	young	high	no	middle	no
2	young	high	no	good	no
3	mid-age	high	no	middle	yes
4	old	middle	no	middle	yes
5	old	low	yes	middle	yes
6	old	low	yes	good	no
7	mid-age	low	yes	good	yes
8	young	middle	no	middle	no
9	young	low	yes	middle	yes
10	old	middle	yes	middle	yes
11	young	middle	yes	good	yes
12	mid-age	middle	no	good	yes
13	mid-age	high	yes	middle	yes
14	old	middle	no	good	no

Solution:

(a) $P(young, low \subseteq income, non \subseteq student, medium \subseteq credit|no)P(no)$
$= P(young|no)P(low \subseteq income|no)P(non \subseteq student|no)P(medium \subseteq credit|no)P(no)$
$= \frac{4}{8}\frac{2}{8}\frac{5}{7}\frac{3}{7}\frac{6}{16} = 0.014$

$P(young, low \subseteq income, non \subseteq student, medium \subseteq credit|yes)P(yes)$
$= P(young|yes)P(low \subseteq income|yes)P(non \subseteq student|yes)P(medium \subseteq credit|yes)P(yes)$
$= \frac{3}{12}\frac{2}{6}\frac{4}{11}\frac{7}{11}\frac{10}{16} = 0.012$

Since $0.014 > 0.012$, the predictive label will be "no".

11. Calculation (Classification, AUC): Bob just labeled a set of 14 emails as legitimate or spam. Alice uses this set of emails to test her score-based classifier f. Alice sets a threshold θ, and for any email x, if $f(x) > \theta$, x will be marked as spam, and if $f(x) \leq \theta$, x will be marked as legitimate.

Let 1 refer to the a spam email, 0 refer to the a legitimate email.

(a) Set $\theta = 20$, calculate the true positive rate and false positive rate of f based on Bob's labels.

(b) Calculate the sample AUC of f based on Bob's labels.

Email ID	Bob's label	f-score of the email
1	spam	77.2
2	spam	69
3	spam	65
4	legitimate	30
5	spam	22
6	legitimate	21.11
7	legitimate	10
8	legitimate	7
9	legitimate	3
10	spam	0.33
11	legitimate	-3
12	legitimate	-6
13	legitimate	-15
14	legitimate	-77

Solution:

(a) Let 1 refer to the a spam Email, 0 refer to the a legitimate Email.

		Actual		
		1	0	Total
Spam	1	4	2	6
	0	1	7	8
	Total	5	9	

Thus:

$$\begin{aligned} TPR &= \frac{TP}{AP} \\ &= \frac{4}{5} \\ &= \boxed{.8000} \end{aligned}$$

$$\begin{aligned} FPR &= \frac{FP}{AN} \\ &= \frac{2}{9} \\ &= \boxed{.2222} \end{aligned}$$

(b)

	<	AP				
		77.2	69	65	22	0.33
AN	30	1	1	1	0	0
	21.11	1	1	1	1	0
	10	1	1	1	1	0
	7	1	1	1	1	0
	3	1	1	1	1	0
	-3	1	1	1	1	1
	-6	1	1	1	1	1
	-15	1	1	1	1	1
	-77	1	1	1	1	1

$$
\begin{aligned}
AUC &= \frac{\sum_{y \in AP} \sum_{x \in AN} \mathbb{I}_{f(y)<f(x)}}{\#ofAP \cdot \#ofAN} \\
&= \frac{5 * 9 - 6}{5 * 9} \\
&= .8667
\end{aligned}
$$