

Problem 1(20) Do you agree with the following statements? Give a brief explanation to support your conclusion.

- 1.1 Data mining is a combination of statistics, database system management and artificial intelligence;
- 1.2 Ten-fold Cross-Validation is the best method to break any data set into training and testing sets;
- 1.3 Supervised learning is to study under supervision and unsupervised learning is self-study;
- 1.4 Missing value is a kind of noisy data;
- 1.5 Based on the MDL principle, instance-based learning is the best learning algorithm because it is very simple and makes no mistakes in the training data set.

Problem 2 Consider a classification problem with the following two point sets

$$\begin{aligned} S_1 &= \{(1, 2), (4, 2), (4, 3)\} \\ S_2 &= \{(0, 2), (-1, 1)\}. \end{aligned}$$

- 2.1 Use linear regression to find linear models for the points in S_1 and S_2 , respectively.
- 2.2 Use those two linear models to construct a linear classifier for S_1 and S_2 .

Problem 3 The following database has five transactions. Use Apriori algorithm to find all the frequent item sets with minimum support 3.

TID	data	items(bought)
T1	10/15/00	{K, A, D, B}
T2	10/15/00	{D, A, C, E, B}
T3	10/15/00	{C, A, D, B}
T4	10/20/00	{A, B, D}
T5	10/18/00	{C, D, E}

Problem 4 The following table gives the marks of two students ‘A’ and ‘B’ for courses C_1, \dots, C_{10} . The symbol ? in the grade column denotes that the student has not taken the corresponding course.

Course	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
A	75	80	84	92	85	76	87	68	80	77
B	83	93	80	75	85	95	87	77	90	?

4.1 Based on the table, predict what are the grades that ‘A’ and ‘B’ might get for a new course with 90% confidence.

4.2 Compare the performance of these two students and state your conclusion.

age	income	married	credit_rating	Loan
21	low	no	excellent	no
25	low	no	fair	no
27	medium	no	excellent	no
28	medium	no	fair	no
29	medium	yes	fair	yes
31	medium	no	excellent	yes
32	high	?	fair	no
36	high	yes	fair	yes
41	medium	yes	fair	yes
45	low	yes	fair	no
?	low	yes	excellent	no
47	medium	yes	fair	yes

Table 5: Training set from a bank database

5.4 Use Bayes’ model to predict the new instance $(35, \text{medium}, \text{yes}, \text{fair})$?

- 5.1** Propose an algorithm to deal with the missing values in the above table and discuss the potential bias if you simply cast the missing value as a global value and use one R algorithm in this application;
- 5.2** One criteria in binary splitting for numerical attributes is to use the information gain obtained by the splitting. For example, if we select 30 as a breaking point for the attribute ‘age’, then we can split the data set into two subsets. One subset contains all the instances for which the attribute ‘age’ has a value less than or equal to 30, and another subset contains all the instances with ‘age’ larger than 30. Find the breaking point with maximal information gain for the attribute ‘age’ in the above table(you can only consider the breaking points where the class values change such as 28, 31).
- 5.3** Use the breaking point you selected from question 5.2 and one R algorithm to construct 1-level decision tree for the above data set.