

Assignment 1 for Course CS4TF3

The due date for handing in the answer of this assignment is January 25, 2006. The delay assignments will be penalized by 20% for every day.

Question 1: Discuss the common and difference between data management and data mining.

Question 2: Describe the major tasks in data mining and use some specific examples for illustration.

Question 3: Give several major representation structures in data mining and discuss their merits and weak points.

Question 4: Suppose that the following data represents the marks of the students in a class, listed in an increasing order:

30, 50, 56, 56, 57, 59, 61, 63, 63, 65, 67, 68, 72, 73, 75, 75, 79, 79, 80, 82, 84, 84, 85, 89, 92, 93, 101.

4.1 Smoothing the data by bin means, using a bin depth of 5.

4.2 Use the bin method to smooth the data such that the gap between the highest and lowest marks in each bin is 10.

4.3 Use an algorithm to detect the outlier in the data.

Question 5: Suppose that the following matrix is the data in some data mining application:

$$A = \begin{pmatrix} 3 & 3 & 14 \\ 3 & 3 & \times \\ 2 & 1 & 10 \\ 4 & 5 & 18 \end{pmatrix}. \quad (1)$$

5.1 Fix the missing value in the matrix indicated by \times .

5.2 Use an algorithm to reduce the matrix to one column by removing the redundant columns in the fixed matrix.