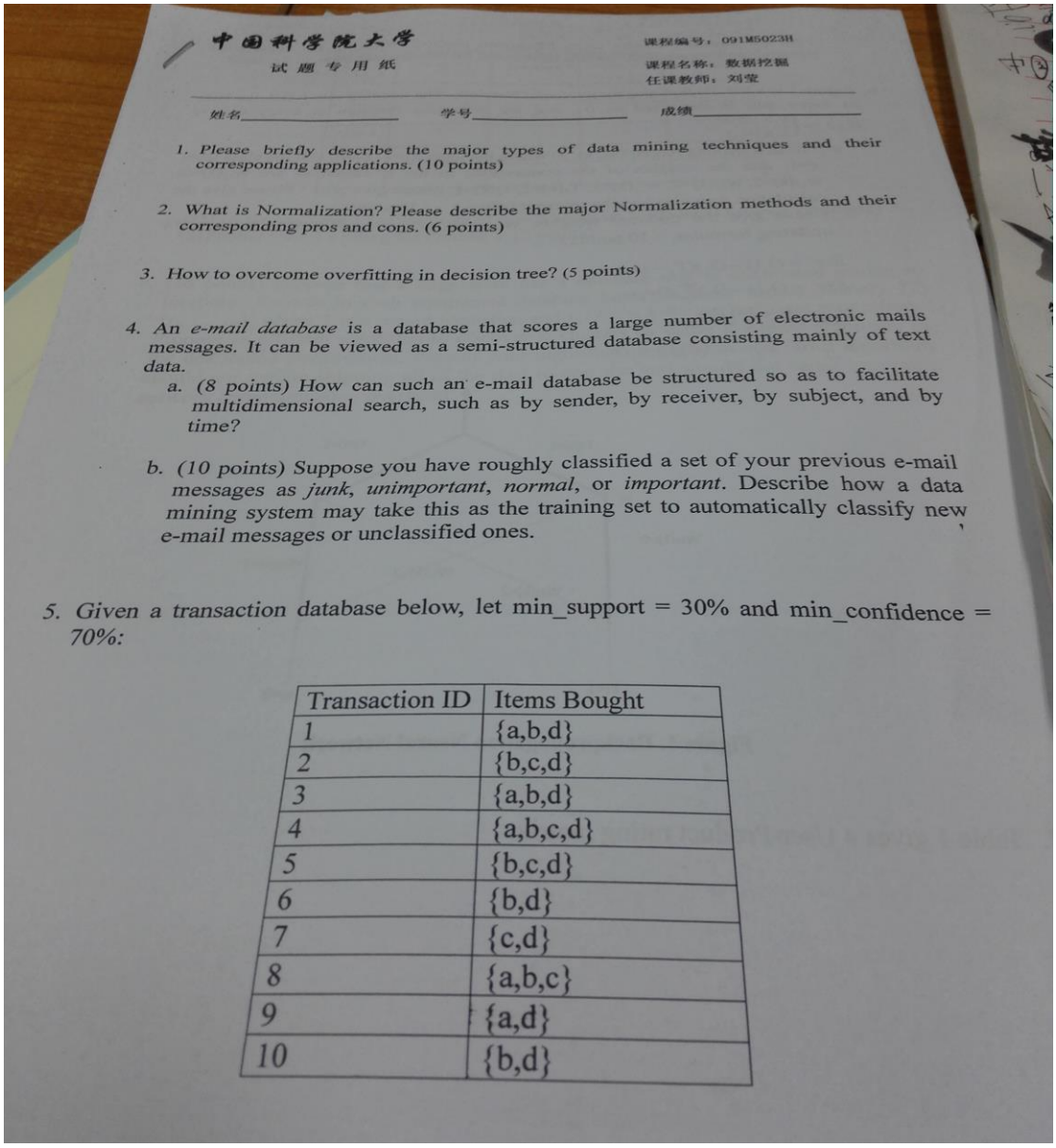
2017 年的题目跟往年的题目相比，题型变化比较大，只有第 5 题和第 8 题是往年的。

计算量不算很大，但是由于平时相关的题目做得不多，所以会手生，算起来比较慢，100 分钟做不完。

想争取高分的可以提前把数据仓库（数据立方体，星状图），分类（Grain 信息增益决策树归纳，朴素贝叶斯，神经网络），聚类（K-means），频繁挖掘（Apriori，FP 树）这几个部分的计算题多做几遍，伪代码建议试着自己多写一些。

《数据挖掘概念技术（第三版）》的课后题可以拿来练手。



中国科学院大学
试 题 专 用 纸

课程编号：091M5023H
课程名称：数据挖掘
任课教师：刘莹

姓名_____ 学号_____ 成绩_____

1. Please briefly describe the major types of data mining techniques and their corresponding applications. (10 points)

2. What is Normalization? Please describe the major Normalization methods and their corresponding pros and cons. (6 points)

3. How to overcome overfitting in decision tree? (5 points)

4. An *e-mail database* is a database that scores a large number of electronic mails messages. It can be viewed as a semi-structured database consisting mainly of text data.
   a. (8 points) How can such an e-mail database be structured so as to facilitate multidimensional search, such as by sender, by receiver, by subject, and by time?
   b. (10 points) Suppose you have roughly classified a set of your previous e-mail messages as *junk*, *unimportant*, *normal*, or *important*. Describe how a data mining system may take this as the training set to automatically classify new e-mail messages or unclassified ones.

5. Given a transaction database below, let min_support = 30% and min_confidence = 70%:

Transaction ID	Items Bought
1	{a,b,d}
2	{b,c,d}
3	{a,b,d}
4	{a,b,c,d}
5	{b,c,d}
6	{b,d}
7	{c,d}
8	{a,b,c}
9	{a,d}
10	{b,d}

Find all frequent itemsets using FP-growth method. Write up the condition pattern base for each item, and the conditional FP-tree for each item. (15 points)

6. Figure 1 is a BP (Backpropagation) Neural Network. The learning rate l=0.9, the Bias at every unit is initialized as 0, and the activation function at every unit is $f(x) = \begin{cases} x, x \geq 1 \\ 1, x < 1 \end{cases}$

a. Given a training record (x_1, x_2, z) where the input x_1=1, x_2=0, and the class label z=1, and the weights of the connections in the k^{th} iteration are $w_{11}(k)$=0, $w_{12}(k)$=2, $w_{21}(k)$=2, $w_{22}(k)$=1, $T_1(k)$=1, $T_2(k)$=1, please give z(k) (Please give the calculation formulas and the relevant values). (10 points)

b. Please give the updated weights, w_{11}, w_{12}, w_{21}, w_{22}, following the weight updating formulas. (10 points)

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\theta_j = \theta_j + (l)Err_j$$

$$w_{i,j} = w_{i,j} + (l)Err_j O_i$$

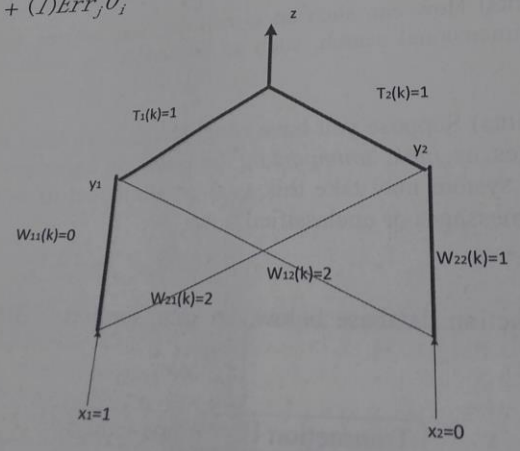


Figure 1. Backpropagation Neural Network

7. Table 1 gives a User-Product rating matrix.

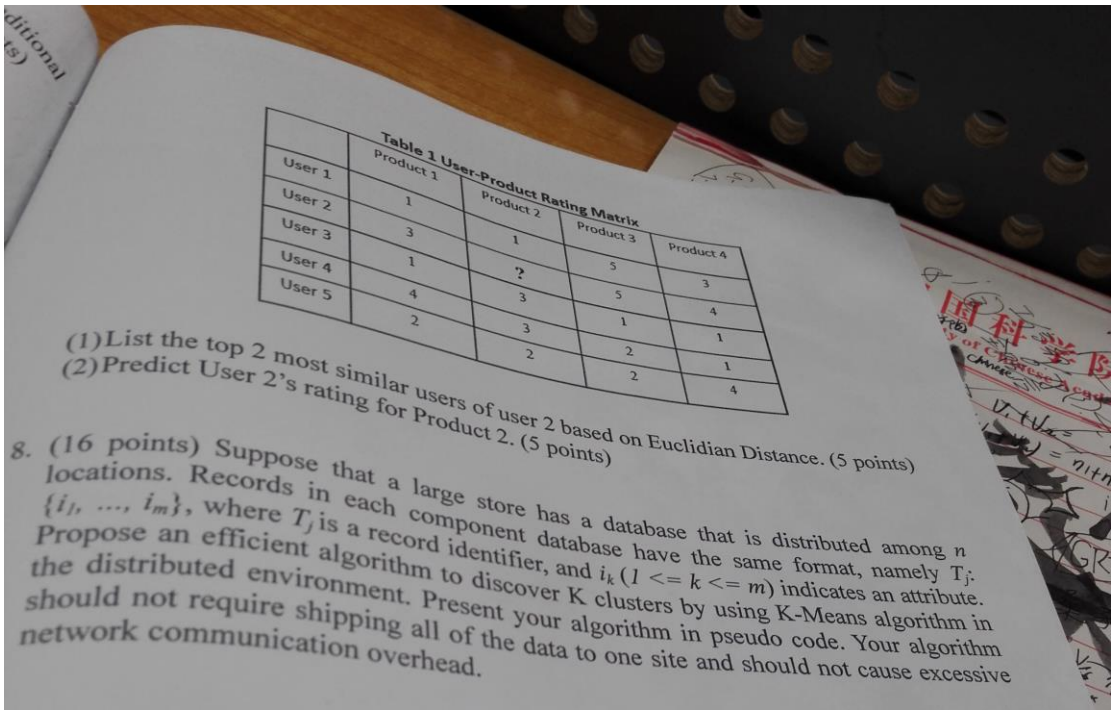


Table 1 User-Product Rating Matrix

	Product 1	Product 2	Product 3	Product 4
User 1	1	1	5	3
User 2	3	?	5	4
User 3	1	3	1	1
User 4	4	3	2	1
User 5	2	2	2	4

(1) List the top 2 most similar users of user 2 based on Euclidian Distance. (5 points)
(2) Predict User 2's rating for Product 2. (5 points)

8. (16 points) Suppose that a large store has a database that is distributed among n locations. Records in each component database have the same format, namely T_j: $\{i_j, \ldots, i_m\}$, where T_j is a record identifier, and i_k ($1 <= k <= m$) indicates an attribute. Propose an efficient algorithm to discover K clusters by using K-Means algorithm in the distributed environment. Present your algorithm in pseudo code. Your algorithm should not require shipping all of the data to one site and should not cause excessive network communication overhead.