UIUC-CS412 "Introduction to Data Mining" (Fall 2014)

Midterm Exam

Friday, Oct. 17, 2014
75 minutes, 75 marks

Name: NetID:

1 [12]	2 [13]	3 [10]	4 [20]	5 [17]	6 [3]	Total [75]

1. [12] Knowing and Preprocessing Data.

   (a) [10] For each of the following scenarios, state which technique is preferred, and briefly explain why.

      i. Classify transactions into two different classes: normal vs. fraudulent. In order to perform sampling in the preprocessing step, which sampling technique is better: stratified sampling or simple random sampling?

      ii. In order to compare the efficiency between Chevrolet and Ford cars, which visualization technique is better: bar chart or boxplot?

      iii. In order to find out advantages and disadvantages of Chevrolet cars compared with Ford cars, which visualization technique is better: parallel coordinates or scatterplot matrix?

iv. Given election survey results, in order to study the correlation between voters' genders and their voting preferences, which measure is better: χ^2 test or correlation coefficient (Pearson's product moment coefficient)?

v. To find the Walgreens stores closest to Illini Union, which similarity/distance measure is better: Minkowsky distance or cosine similarity?

(b) [2] List the value ranges of the following measures.

i. Min-Max normalization

ii. Correlation coefficient (Pearson's product moment coefficient)

2. [13] Principal Component Analysis (PCA).
Consider 10 data points in 2-D space, as specified in Table 1.

X	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
Y	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

Table 1: Data points in 2-D space.

So that you do not need to calculate, we give you the following statistics:

$$\mu_x = \sum_{i=1}^{10} x_i = 0$$

$$\mu_y = \sum_{i=1}^{10} y_i = 0$$

$$\delta_x^2 = \frac{1}{10}\sum_{i=1}^{10} x_i^2 = 0.5549$$

$$\delta_y^2 = \frac{1}{10}\sum_{i=1}^{10} y_i^2 = 0.6449$$

$$\delta_{xy} = \frac{1}{10}\sum_{i=1}^{10} x_i y_i = 0.5539$$

(a) [3] Use the given statistics to give a formula to calculate the correlation coefficient (Pearson's product moment coefficient) of the data points in Table 1. (Numerical results are not required). Based on your observation, are the two dimensions positively or negatively correlated? Briefly explain your answer.

(b) [3] Write down the covariance matrix for the data points in Table 1.

(c) [2] Given the direction of the first principal component, as shown as the line in Figure 1, draw the projection of data point C into the new feature space constructed from the first principal component on Figure 1.

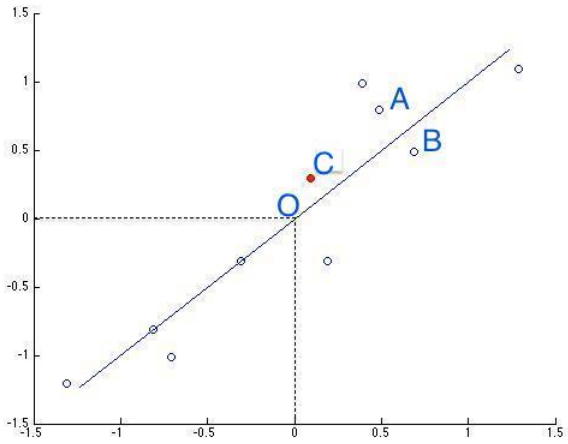

Figure 1: Visualization of data points and principal components

(d) [2] Draw the second principal component on Figure 1, and make sure that it goes through the origin.

(e) [3] Assume that points A and B belong to two different classes. Is it still a good strategy to project the data into the new feature space, constructed from the first principal component? Why or why not? Briefly explain your answer.

3. [20] Data Warehousing and OLAP for Data Mining.
Suppose the base cuboid of a data cube contains four cells

$$(a_1, a_2, a_3, b_4, a_5, \ldots, a_{10})$$
$$(a_1, a_2, b_3, a_4, a_5, \ldots, a_{10})$$
$$(a_1, b_2, a_3, a_4, a_5, \ldots, a_{10})$$
$$(b_1, a_2, a_3, a_4, a_5, \ldots, a_{10})$$

where $a_i \neq b_i$, $\forall i = 1, 2, 3, 4$.

(a) [5] How many cuboids are there in the full data cube?

(b) [5] How many **nonempty aggregate closed** cells are there in the full cube?

(c) [5] How many **nonempty aggregate** cells are there in the full cube?

(d) [5] If we set minimum support = 2, how many **nonempty aggregate** cells are there in the corresponding iceberg cube?

4. [10] Data Cube Implementation.

Suppose we use Bottom-Up Computation to materialize cubes. Consider a 3-D data array containing three dimensions A, B, C. The data contained in the array is as follows:

$(a_0, b_0, c_0) : 1$	$(a_0, b_0, c_1) : 1$	$(a_0, b_0, c_2) : 1$
$(a_1, b_1, c_0) : 3$	$(a_1, b_1, c_1) : 3$	$(a_1, b_1, c_2) : 3$
$(a_0, b_2, c_0) : 1$	$(a_0, b_2, c_1) : 1$	$(a_0, b_2, c_2) : 1$

Suppose we construct an iceberge cube for dimension A, B, C with different orders of exploration.

(a) [4] Draw the trace trees of expansion with regard different exploration orders: A, B, C and C, B, A, respectively.

(b) [6] If we set minimum support = 4 with the exploration order of A, B, C, how many cells would be considered/computed?

5. [17] Frequent Pattern and Association Mining.
   A database with 150 transactions has its FP-tree shown in Figure 2. Let relative $min\_sup = 0.4$ and $min\_conf = 0.7$.

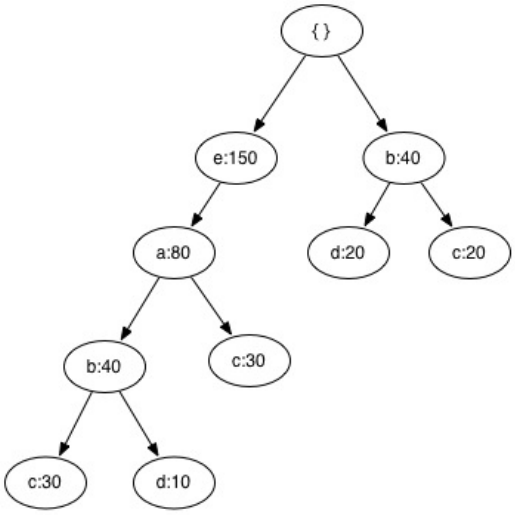

Figure 2: FP tree of a transaction DB

   i. [5] Show c's conditional (i.e., projected) database.

   ii. [6] Present all frequent 3-itemsets and 2-itemsets.

iii. [6] Present two frequent association rules based on the given relative minimum support and confidence.

6. [3] (Opinion).

   (a) I ☑ like ☐ dislike the exams in this style.

   (b) In general, the exam questions are ☐ too hard ☐ too easy ☐ just right.

   (c) I ☐ have plenty of time ☐ have just enough time ☐ do not have enough time to finish the exam questions.