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 [20]	4 [10]	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. **Note: correct answers without correct explanation receive 0 point. 5 unexplained answers get 1 point in total**

      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?
      [ANSWER: stratified sampling because of the unbalance between normal and fradulent. In other words, as normal transactions are much more popular than fraudulent, if we do random sampling, it is likely we will see very few fradulent transactions in the sample. In your anwer, you must explicitly say the 2 types of transactions are unbalanced to get the full score.]

      ii. In order to compare the efficiency between Chevrolet and Ford cars, which visualization technique is better: bar chart or boxplot?
      [ANSWER: boxplot because besides mean or any other types of aggregation, it expresses additional useful statistics such as quantiles. If you use bar chart to express histogram (which actually doesn't make a good sense when comparing with boxplot), and argue it is more informative than boxplot, we give you one point because even histogram is more informative, it is harder to interprete.]

      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?
      [ANSWER: parallel coordinates because it shows comparisons for all properties together in a 2-D graph, while the scatterplot matrix shows multiple graphs for all possible pairs, which is harder to interprete. Scatterplot matrix is more suitable for pairwise correlation analysis. The answer is quite standard, because I did mention it in the first demonstration in class. Many of you think advantage and disadvantage are two single characters, which

doesn't make sense. We must understand them as something like: among MPG, speed, acceleration, which are the advantages and disadvantages of Chevy, compared to Ford cars. If you argue that scatterplot matrix shows more details, so you can learn more from the graph, we will give you 1 point.]

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)?
[ANSWER: χ^2 test because features are categorical/nominal rather than continuous. You must explicitly say the features are nominal to get full score. I saw a lot of answers like writing characteristics of χ^2 test from slides/text book in general case, which will get zero point]

v. To find the Walgreens stores closest to Illini Union, which similarity/distance measure is better: Minkowsky distance or cosine similarity?
[ANSWER: Minkowsky distinace because it measures geometrical distance, while cosine-similarity is not. Answers, such as data points are numeric, are not acceptable because cosine similarity works with numeric data points as well. Or if you argue cosine similarity only works for sparse data or documents, you will get zero points too because they are just a few applications of cosine similarity]

(b) [2] List the value ranges of the following measures. **Note: each following question is worth only 1 point!**

i. Min-Max normalization
[ANSWER: $[new\_min, new\_max]$ or $[0, 1]$ are all acceptable. In this question, you don't have to explain. In previous questions, if you don't explain, you will receive zero point. If all your answers are unexplained and most of them are correct, we give you 1 point for all]

ii. Correlation coefficient (Pearson's product moment coefficient)
[ANSWER: $[-1, 1]$]

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:

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

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

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

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

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

Note: We appologize for using δ^2 instead of σ^2 for variance, but as we have the definition following it, we think it is clear enough.

(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.

[ANSWER: $0.5539/\sqrt{(0.5549 * 0.6449)} \approx 0.93$. Conclusion: positively correlated. We asked you to use the provided statistics, so to get the full score, you must either present final results, numerical formula, or formula containing only provided statistics. Many of you only present the formula in the textbook which even use a,b instead of x,y, you will receive zero point]

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

[ANSWER: Just need to plug the provided statistics into the correct positions of the covariance matrix $\begin{bmatrix} 0.5549 & 0.5539 \\ 0.5539 & 0.6449 \end{bmatrix}$ If you calculate it manually, you will receive full point as long as you give us the final matrix.]

(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.

[ANSWER: Many of you seemed to forget to answer this question. We are sorry but we will have to give you zero point for that]

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

[ANSWER: The second principal component should be orthogonal with the first one. Many of you think the origin is at bottom-left corner, we think it is unimportant, so you still receive full score for that. A few of you did manually calculate the PCA, but I did not see the correct answer by using that method. The figure 1 is the answer for both this and the last sub-question.]

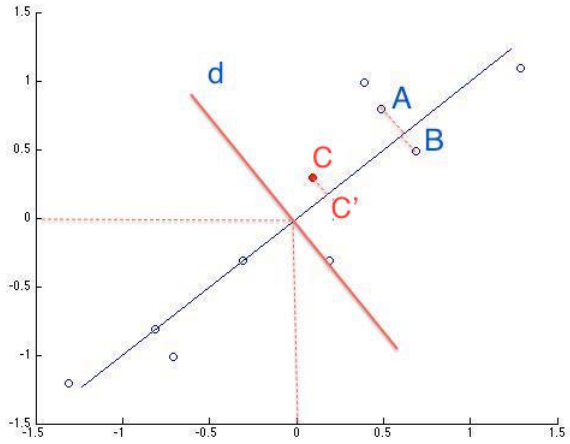

Figure 1: Visualization of data points and principal components

(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.

[ANSWER: No, because the projected points are not discriminative while they should be so that classification algorithms can separate them into 2 different categories. Please note that using PCA is not always a good idea. You must experiment or observe from graph to see if you should use that. Many of you think the labels do not matter, only variance matters. However, the fact is variance minimization is the PCA assumption, which is not neccessary true for all cases. You must carefully examine the results to decide if the assumption holds and if PCA really works]

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?

[ANSWER: As there are 10 attributes, and there is no dimension with concept hierarchy, so there are $(1+1)^{10} = 1024$ cuboids in the data cube.]

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

[ANSWER: There are 11 nonempty aggregate closed cells in the data cube, as follows:

$(*, *, *, *, a_5, \ldots, a_{10})$
$(a_1, *, *, *, a_5, \ldots, a_{10})$
$(*, a_2, *, *, a_5, \ldots, a_{10})$
$(*, *, a_3, *, a_5, \ldots, a_{10})$
$(*, *, *, a_4, a_5, \ldots, a_{10})$
$(a_1, a_2, *, *, a_5, \ldots, a_{10})$
$(a_1, *, a_3, *, a_5, \ldots, a_{10})$
$(a_1, *, *, a_4, a_5, \ldots, a_{10})$
$(*, a_2, a_3, *, a_5, \ldots, a_{10})$
$(*, a_2, *, a_4, a_5, \ldots, a_{10})$
$(*, *, a_3, a_4, a_5, \ldots, a_{10})$

]

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

[ANSWER: The number of (nonempty) aggregate cells in the data cube is 3004.
For these cells, the last 6 dimensions have the same values, we only need to consider the first 4 dimensions.
The total number of cells is 3^4, and there are the following empty cells:

i. All four dimensions are a, 1;
ii. There are two dimensions with value b, $\binom{4}{2} * 2^2 = 24$;
iii. There are three dimensions with value b, $\binom{4}{3} * 2 = 8$;
iv. All four dimensions are b, 1;

The number of base cells is 4. The aggregate cell count is $2^6 * (3^4 - 1 - 24 - 8 - 1) - 4 = 3004$.]

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

[ANSWER: If the minimum support = 2, only a would appear in the aggregate cell, the total count is $11 \times 2^6 = 704$.
If we only consider the first 4 dimensions, the cell would only contain a and *,

(a) Contain 2 a's, there are $\binom{4}{2} = 6$;
(b) Contains 1 a's, there are $\binom{4}{1} = 4$;
(c) Contains 0 a's, there are $\binom{4}{0} = 1$;

In total, there are $(6 + 4 + 1) \times 2^6 = 704$.]

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.

[Answer: The trace tree of expansion of ABC is shown in Figure 2, and CBA is shown in Figure 4.]

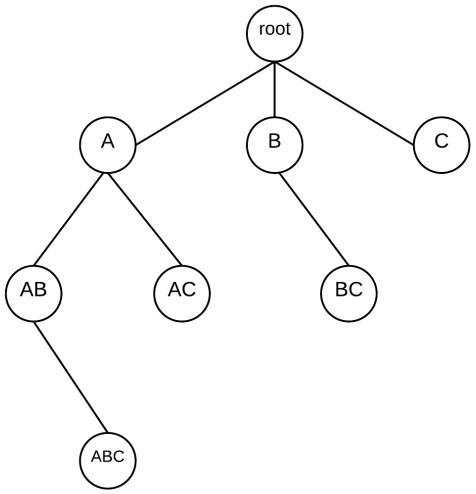

Figure 2: Trace tree of expansion (ABC)

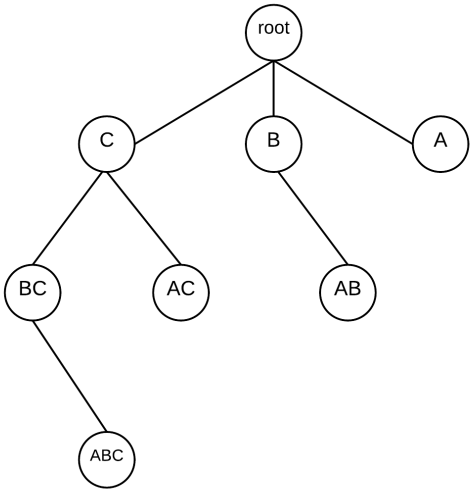

Figure 3: Trace tree of expansion (CBA)

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

[ANSWER: First we list the count of each cells in all the cuboids, as follows:

All $(*,*,*) : 15 - expansion$

A $(a_0,*,*) : 6 - expansion$
A $(a_1,*,*) : 9 - expansion$

AB $(a_0,b_0,*) : 3$
AB $(a_1,b_1,*) : 9 - expansion$
AB $(a_0,b_2,*) : 3$

ABC $(a_1,b_1,c_0) : 3$
ABC $(a_1,b_1,c_1) : 3$
ABC $(a_1,b_1,c_2) : 3$

AC $(a_0,*,c_0) : 2$
AC $(a_0,*,c_1) : 2$
AC $(a_0,*,c_2) : 2$
AC $(a_1,*,c_0) : 3$
AC $(a_1,*,c_1) : 3$
AC $(a_1,*,c_2) : 3$

B $(*,b_0,*) : 3$
B $(*,b_1,*) : 9 - expansion$
B $(*,b_2,*) : 3$

BC $(*,b_1,c_0) : 3$
BC $(*,b_1,c_1) : 3$
BC $(*,b_1,c_2) : 3$

C $(*,*,c_0) : 5$
C $(*,*,c_1) : 5$
C $(*,*,c_2) : 5$

Consider the order of A, B, C, the number of cells to be computed is

$$1(All) + 2(A) + 3(AB) + 3(ABC) + 6(AC) + 3(B) + 3(BC) + 3(C) = 24$$

]

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

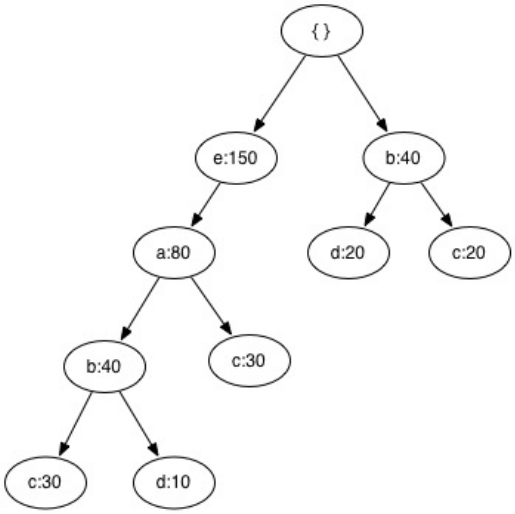

Figure 4: FP tree of a transaction DB

i. [4] Show c's conditional (i.e., projected) database.

ii. [4] Present all frequent 3-itemsets and 2-itemsets.

iii. [4] Present two association rules with support and confidence containing 2 or 3 items.

Answer: [**Note: The answers of i. and ii. are the same with e:110 and e:150]

i. eab:30, ea:30, b:20

ii. ae:80, eac: 60, ce:60, ac:60.

iii. If we use $e : 150$:
$a \rightarrow e : (80/150, 80/80)$,
$ae \rightarrow c : (60/150, 60/80)$
$ac \rightarrow e : (60/150, 60/60)$
$ce \rightarrow a : (60/150, 60/60)$
$a \rightarrow ce : (60/150, 60/80)$
$c \rightarrow ae : (60/150, 60/80)$
$c \rightarrow e : (60/150, 60/80)$
$a \rightarrow c : (60/150, 60/80)$
$c \rightarrow a : (60/150, 60/80)$
If you use $e : 110$, one additional decision rule is $e \rightarrow a : (80/150, 80/110)$

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.