Northeastern University, College of Computer and Information Science

CS 6220 Data Mining Techniques
Midterm Exam

Due Sunday October 20th, 2019 at 11:59pm PST.

Name: _____Yuanjie Yue_____

Note:

1. This is an individual exam. Group work is not permitted.
2. This exam contains 10 pages (feel free to expand the number of pages if needed)
3. This exam is open book and open notes.
4. Write your answers on this document itself for submission.
5. This exam is graded out of 100 points.
6. Do as much as you can; partial credit will be given for partially correct answers.
7. Please notify the instructor if any question is unclear.
8. The Northeastern University Code of Student Conduct applies.

1: (10 points)

Data may be missing from a dataset in three different ways: missing at random (MAR), missing not at random (MNAR), and missing completely at random (MCAR). So far as you can discern, which of these three types of "missingness" does the following dataset exhibit. Support your answer.

Petal Length	Petal Width	Species Type
low	*missing*	Setosa
low	*missing*	Setosa
medium	low	Setosa
medium	medium	Versicolour
medium	high	Versicolour
medium	high	Virginica
high	medium	Versicolour
high	medium	Virginica
high	high	Versicolour
high	high	Virginica

Answer 1:
It is MAR, missing at random.

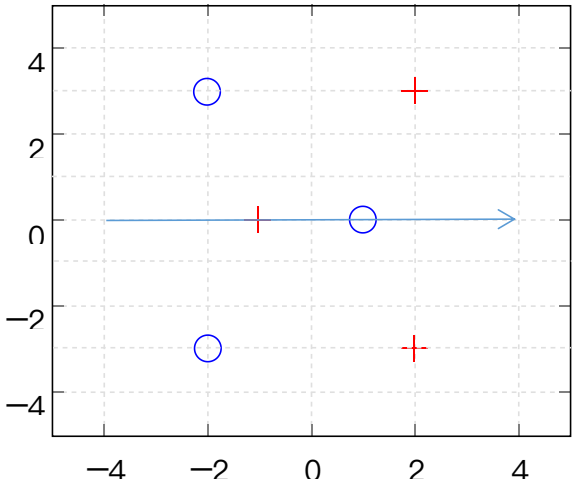
The missing data are all of Setosa species and its relative Petal Length are all 'low', which means they depend on the values of observed variables, so it is not MCAR.

Meanwhile, the Petal Width is observed with all of the three values, 'low', 'medium' and 'high', so the missing value itself does not depend on the values of any of the missing or unobserved variables.

Therefore, this case belongs to MAR.

2: (10 points)

Consider performing principal component analysis (PCA) on the following data. Draw the direction of the second principal component. Use the Insert → Shapes feature to draw an arrow indicating the component's direction.



Data on which to perform PCA (circles and crosses denote different classes of instances).

Answer 2:

PCA is aimed to help us generate fewer features by reducing the data dimension.

In the first projection, we usually maximize the variance of the data. In this case, we could easily tell that the maximum variance lays almost parallel with the Y axis, we will pick the positive Y.

Then the second projection we will pick the direction that is perpendicular to the first projection, which means that its direction is parallel with the X axis.

3: (10 points)

Suppose the FDA wants to investigate a hypothesis that pesticides present in the water supply are responsible for spinach poisoning. The farms that use pesticides are geographically dispersed and investigating every farm across the country is found to be too time consuming. Suppose that you instead recommend using a sampling strategy for this problem. What sampling strategy should you recommend and why?

Answer 3:
I recommend a Cluster Random Sampling strategy.

The main reason is that the farms are geographically dispersed across the country, therefore, these farms might be varied from each other on the usage amount of pesticides due to their different severity of pest. We all know that the level of pest is related to climate and different regions might have completely different climate, such as temperature and precipitation, thus a Simple Random Sampling may be imbalanced.

We could group or segment the farms based on regions and climate, then randomly select from each group. In this way, the samples chosen should be representative.

4: (10 points)

Consider the following database of transactions. Extract and list all association rules with a minimum support of 60% (3/5) and a minimum confidence of 80% (4/5). For each rule, please draw an arrow to indicate the direction of implication between items.

Hint: For the correct answer, you only need to consider 2-itemsets (two-item itemsets).

TID	Transactions
1	bread, milk
2	bread, diapers, beer, eggs
3	milk, diapers, beer, coke
4	bread, milk, diapers, beer
5	bread, milk, diapers, coke

Answer 4:

beer -> diapers, support = 3/5, confidence = 3/3

There are six items, bread, milk, diapers, beer, eggs, coke.
The eggs support = 1/5, coke support = 2/5, which is lower than minimum support, no need to consider.
Therefore, we will check the rules between these four items, bread, milk, diapers and beer.

bread -> milk, s = 3/5, c = 3/4
-> diapers, s = 3/5, c = 3/4
-> beer, s = 2/5
milk -> bread, s = 3/5, c = 3/4
-> diapers, s = 3/5,c = 3/4
-> beer, s = 2/5
diapers -> bread, s = 3/5, c = 3/4
-> milk, s = 3/5, c = 3/4
-> beer, s = 3/5, c = 3/4
Beer -> bread, s = 2/5
-> milk, s = 2/5
-> diapers, s = 3/5, c = 3/3

5: (10 points)

(a) In association rule mining, a strong association rule satisfies both a minimum support (*min_support*) and minimum confidence (*min_confidence*) threshold. Consider the sales data in the table below. Given *min_support* = 10% (1/10) and *min_confidence* = 33% (1/3), is **hotdogs → hamburgers** a strong rule? Show your calculations.

(b) Is the purchase of hotdogs independent of the purchase of hamburgers? If not, what correlation (positive or negative) exists between the two?

Sales Data of Hamburgers and Hotdogs

	Hotdogs	$\overline{\text{Hotdogs}}$	
Hamburgers	500	2000	2500
$\overline{\text{Hamburgers}}$	1500	1000	2500
	2000	3000	5000

Answer 5:

Before we start, let's take a look at this contingency tables:

	B	$\bar{B}$	
A	$f_{11} = 500$	$f_{10} = 2000$	$f_{1+} = 2500$
$\bar{A}$	$f_{01} = 1500$	$f_{00} = 1000$	$f_{0+} = 2500$
	$f_{+1} = 2000$	$f_{+0} = 3000$	$N = 5000$

1. Not a strong rule.

Support (hotdogs → hamburgers) $= f_{11} / N$
$= 500 / 5000$
$= 10\%$

Confidence (hotdogs → hamburgers) $= f_{11} / f_{+1}$
$= 500 / 2000$
$= 25\%$

2. Not independent.

Let's calculate the Interest Factor following the formula:

I (hotdogs, hamburgers) = S (hotdogs, hamburgers) / (S (hotdogs) * S (hamburgers))
$= (N * f_{11}) / (f_{1+} * f_{+1})$
$= (5000 * 500) / (2500 * 2000)$
$= 0.5 < 1$

This means there is a negative correlation exists between the two.

6: (10 points)

Let minimum support = 20% (1/5) and minimum confidence = 60% (3/5). Given the transaction database below, find the frequent itemsets and confident association rules. Please show your intermediate work.

Database of Transactions

TID	Items Bought
1	egg, sausage, spam
2	egg, spam
3	egg
4	egg, bacon
5	sausage
6	sausage, spam
7	egg, sausage, spam
8	egg, sausage, bacon
9	sausage, spam
10	egg, sausage, spam, bacon

Answer 6:

Since minimum support is 20%, so for 10 pieces of data, the $minsup_{count}$=2.

1. We could find the frequent itemsets as follow

C1	
Itemset	support
{egg}	7
{sausage}	7
{spam}	6
{bacon}	3

F1	
Itemset	support
{egg}	7
{sausage}	7
{spam}	6
{bacon}	3

C2	
Itemset	support
{egg, sausage}	4
{egg, spam}	4
{egg, bacon}	3
{sausage, spam}	5
{sausage, bacon}	2
{spam, bacon}	1

F2	
Itemset	support
{egg, sausage}	4
{egg, spam}	4
{egg, bacon}	3
{sausage, spam}	5
{sausage, bacon}	2

C3	
Itemset	support
{egg, sausage, spam}	3
{egg, sausage, bacon}	2
{egg, spam, bacon}	1
{sausage, spam, bacon}	1

F3	
Itemset	support
{egg, sausage, spam}	3
{egg, sausage, bacon}	2

2. For all the association in F2 and F3, we could find out that there are totally 11 confident rules, they are: 4), 6),7),8),10),11),13),15), 18),19), 21)

We are calculating the confidence of every subsets in the frequent itemsets:

1) egg -> sausage, c = 4/7
2) sausage -> egg, c = 4/7
3) egg -> spam, c = 4/7
4) **spam -> egg, c = 4/6 > 60%**
5) egg -> bacon, c = 3/7
6) **bacon -> egg, c = 3/3 > 60%**
7) **sausage -> spam, c = 5/7 > 60%**
8) **spam -> sausage, c = 5/6 > 60%**
9) sausage -> bacon, c = 2/7
10) **bacon -> sausage, c = 2/3 > 60%**
11) **(egg, sausage) -> spam, c = 3/4 > 60%**
12) spam -> (egg, sausage), c = 3/6
13) **(egg, spam) -> sausage, c = 3/3 > 60%**
14) sausage -> (egg, spam), c = 3/7
15) **(sausage, spam) -> egg, c = 3/3 > 60%**
16) egg -> (sausage, spam), c = 3/7
17) (egg, sausage) -> bacon, c = 2/4
18) **bacon -> (egg, sausage), c = 2/3 > 60%**
19) **(egg, bacon) -> sausage, c = 2/3 > 60%**
20) sausage -> (egg, bacon), c = 2/7
21) **(sausage, bacon) -> egg, c = 2/2 > 60%**
22) egg -> (sausage, spam), c = 2/7

7: (10 points)

One of the drawbacks of *k*-means clustering is that one may arrive at different outcomes depending on the initial value choice for *k*. Explain (1) why the value of *k* is often not obvious, and (2) how is it possible that even when using the same value of *k*, one can arrive at slightly different results across multiple runs of *k*-means.

Answer 7:

1. The value of k is not obvious, mostly because it depends on scale and distribution of data, which may vary so much.

2. This could be explained how the k-means runs.
   a. Pick up k initial cluster centers randomly
   b. Assign each instance to the closest center
   c. Move the center to the mean of each cluster
   d. Repeat a) and b) until converge.

From this process, we could easily tell that, the problem lays in the picking of the k centers in the first step. Since they are generated randomly, it will affects the calculation in the following steps, which may lead to slightly different results across multiple runs.

8: (10 points)

Given the following data points $p_1 = 1$; $p_2 = 3$; $p_3 = 8$; $p_4 = 10$; $p_5 = 16$:

(a) Compute final centroids for k-means with $k = 3$ clusters using initial centroids of 4, 5, and 8.

(b) Compute the sum of squared error (SSE) for the clusters generated in (a) with the 3 ground truth clusters $\{p_1, p_2\}$, $\{p_3, p_4\}$, $\{p_5\}$.

SSE is given by

$$SSE = \sum_{i=1}^{k} \left(\sum_{x \in C_i} \|x - c_i\|^2 \right),$$

where $x \in C_i$ is all points in cluster C_i with centroid c_i and $i = 1 \ldots k$.

Answer 8:

1. Final centroids are [2, 9, 16]

We could do the calculation following these 4 steps:

a. Pick up k initial cluster centers randomly
b. Assign each instance to the closest center
c. Move the center to the mean of each cluster
d. Repeat a) and b) until converge.

It will takes us 3 rounds of repeating the a) and b)

After 1st round, the centers will be [2, 5, 11].

After 2nd round, the centers will be [2, 8, 13]

After 3rd round, the centers will be [2, 9, 16]

Then the centers will not change, which means that it converged, we are done, the final clusters are [1, 1, 2, 2, 3].

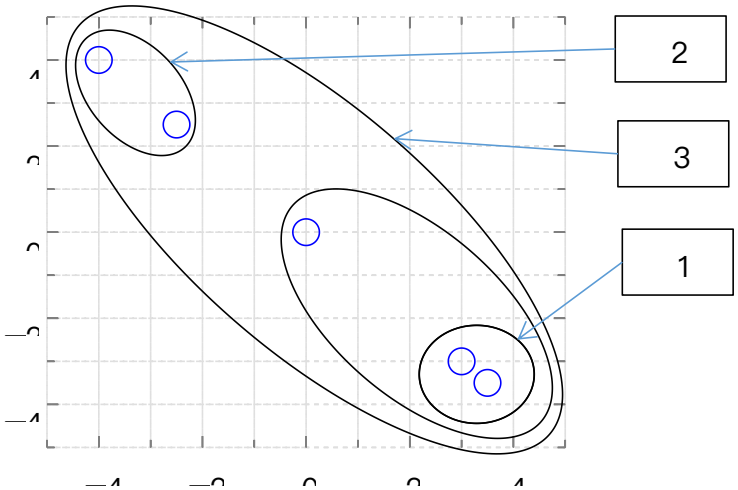
2. Let's compute the SSE as follows:

SSE = [(1 - 2)^2 + (3 - 2)^2] + [(8 - 9)^2 + (10 - 9)^2] + [(16 - 16)^2]

= 4

9: (10 points)

Consider performing complete link (MAX) hierarchical clustering on the following data. Draw the clusters that are found by complete link clustering, labeling them in the order in which they are found. Hint: The correct answer can be obtained visually; you do not need to compute the proximity matrix.



Data to cluster (circles denote instances)

Answer 9:

10: (10 points)

I have a dataset with 1,000,000 instances and 10,000 features. The data is highly non-linear with complex feature interactions, but almost all of the variability of the dataset can be explained by using less than 1% of the features. I need to reduce the set of 10,000 features to a smaller set of exactly 100 features in a computationally feasibly way and I have to maintain the interpretability of the output. What data reduction technique should I use? Why?

Answer 10:
I think we could use the Local Linear Embedding (LLE).

As is mentioned, the data is highly non-linear, we know that PCA can not help us seperate the non-linear data so well. So here, we could use LLE to help us better analyzing overlapping local neighborhoods and determine local structure.