Northeastern University, College of Computer and Information Science

CS 6220 Data Mining Techniques
Midterm Exam

Due Thursday March 11th, 2021 at 8:59pm PST.

Name (print): Yiman Liu Signature:

Note:

1. This is an individual exam. Group work is not permitted.
2. This exam contains 9 pages (feel free to expand the number of pages if needed)
3. This exam is open book and open notes but you may **not** use on-line.
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.

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

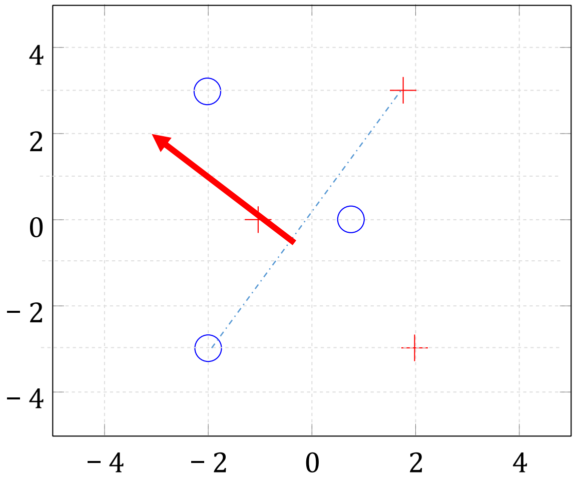
Answer 1:

I think the missingness in this dataset belongs to **missing at random (MAR)**.

1. It is not "missing not at random (MNAR)" because the missing values do not depend on the values of the missing. As we can see, the data for Petal Width includes "low", "medium" and "high", so there is no pattern we can observe from the missing values.

2. Now, we have to decide if it's missing at random (MAR) or missing completely at random (MCAR). As we can see, the data of Petal With is only missing from **low** Petal Length, so we can say the missing data depends on the values of the observed variable. In this case, the observed variable is Petal Length. Thus, it is missing at random.

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:

The first principal component is the direction which has the largest variance. The second principal component is the direction which maximizes variance among all directions orthogonal to the first.

Reference: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf

3: (10 points)
Consider the following database of transactions. Extract and list all association rules with a minimum support of 0.6 (3/5) and a minimum confidence of 0.8 (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 3:

1. Calculate support value for each itemset and select itemset that >= min_support:

Bread, Milk: 3/5 ✓
Bread, Diapers: 3/5✓
Bread, Beer: 2/5
Bread, Eggs: 1/5
Bread, Coke: 1/5

Milk, Diapers: 3/5✓
Milk, Beer: 2/5
Milk, Eggs: 0
Milk, Coke: 2/5

Diapers, Beer: 3/5 ✓
Diapers, Eggs: 1/5
Diapers, Coke: 2/5

Beer, Eggs: 1/5
Beer, Coke: 1/5

Eggs, Coke: 0

2. Calculate confidence value for each valid itemset (use itemset with support >= 3/5):

Bread => Milk: ¾ = 0.75
Milk => Bread: ¾ = 0.75

Bread => Diapers: ¾ = 0.75
Diapers => Bread: ¾ = 0.75

Milk => Diapers: ¾ = 0.75
Diapers => Milk: ¾ = 0.75

Diapers => Beer: ¾ = 0.75
Beer => Diapers: 3/3 = 1 ✓

Thus, association rule with a minimum support of 0.6 (3/5) and a minimum confidence of 0.8 (4/5) is: **Beer ➔ Diapers**

4: (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 4:

I would recommend using **Cluster Sampling**. As the dataset is large and dispersed, it's difficult to collect all the data. Instead, we can group the data based on similarities, and then randomly select from each group.

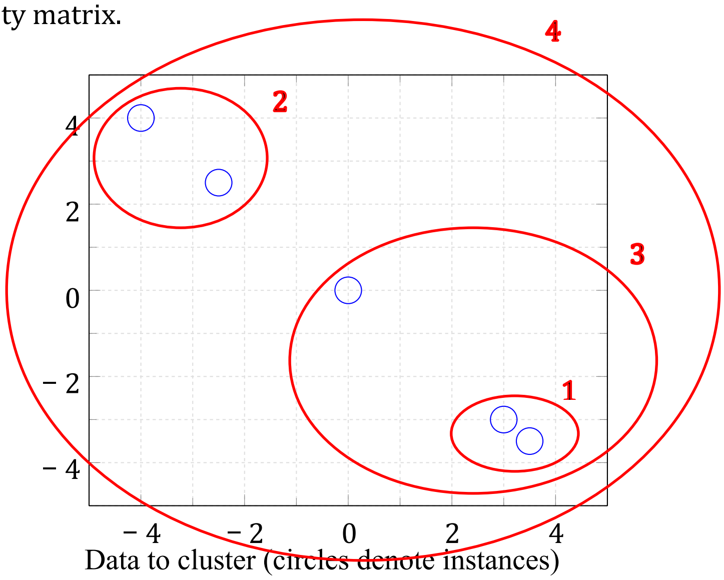
In this case, we can group farms based on their locations first, such as states. For each state, we can further group them based on what pesticides they use as even for one state, the data could be also large.

Then, we could have two ways to select our samples:

> One way is to randomly select entire groups. The advantage is that we don't need to travel to every state to investigate, which can save time and cost. However, the disadvantage is that we may miss some important data points if different groups have significant difference.
>
> Another way is to randomly select an individual from every group. The benefit is that we include almost every case as individuals in the same group share the same locations and use the same pesticides, so one individual could represent the entire group. However, the disadvantage is that we may need to take more cost and time even though the data has already been reduced.

5: (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 5:

6: (10 points)
Two measures of cluster performance are cluster cohesion (compactness, tightness) and cluster separation (isolation). Given two clusters of objects, describe how we would decide whether they are good or poor clusters, based on these two measures.

Answer 6:

In general, **the greater the cohesion within a group and the greater the separation among groups, the better clusters we get**.

We can use the sum of the proximities from each instance within a group to its centroid to represent cohesion. The smaller the sum is, the tighter the cluster is. Separation can be represented as the proximity of the two cluster prototypes. The bigger the proximity is, the more separate the clusters are. Thus, tighter within a group and more separate among groups means good clusters.

The Silhouette Coefficient combines cohesion and separation to evaluate cluster performance. It expresses as:

$S_i = (b_i - a_i) / \max(a_i, b_i)$, where b_i is the minimum average distance from one object to all other objects in other clusters. a_i is the average distance from one object to other object within the same cluster.

In this case, if we are given two clusters of objects, we want to have a greater b_i and smaller a_i. If a_i is smaller enough, like close to 0, then we would get $S_i \sim 1$, which would be good clusters. If we get negative value, which means a_i is greater than b_i, then we get poor clusters.

7: (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 7:

Calculate support:
{hotdogs, hamburgers} => 500 / 5000 = 0.1 = 10% ✓

Calculate confidence:
hotdogs → hamburgers: 500 / 2000 = 0.25 = 25% < 33% ✗

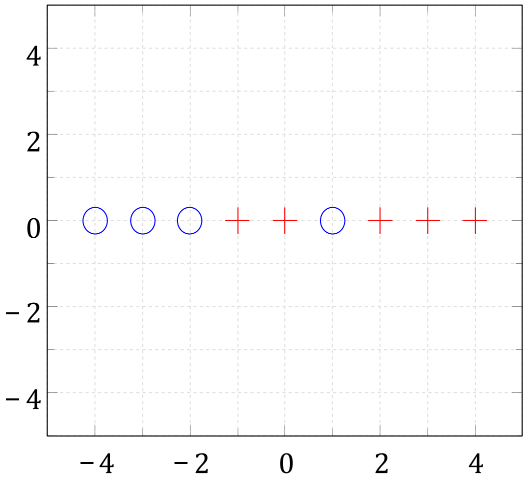
(a) I think **hotdogs → hamburgers** is **not** a strong rule, because its confidence value smaller than min_confidence.

{hotdogs, hamburgers} => 500 / 5000 = 10%
{hotdogs} => 2000 / 5000 => 40%
{hamburgers} => 2500 / 5000 => 50%

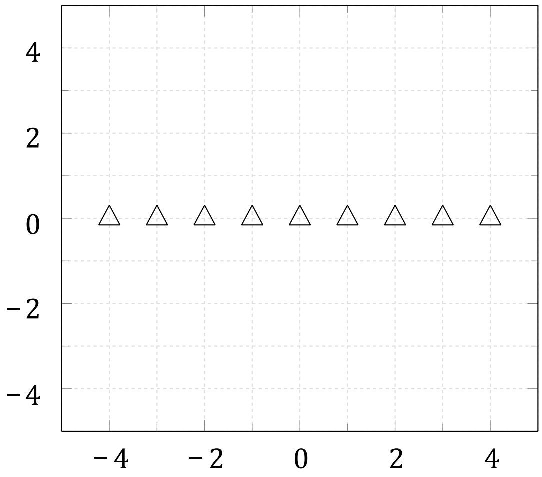
lift = 0.1 / (0.4 * 0.5) = 0.5

(b) lift == 0.5 != 1, so the purchase of hotdogs is **not** independent of the purchase of hamburgers. Hotdogs and hamburgers are **negatively correlated** as the lift is smaller than 1.

8: (10 points)
You use k-nearest neighbor (kNN) to classify a two-class dataset shown below (left). Then you apply your classifier back to the data on which it was trained (or, equivalently, you apply it to new yet identical data) and try to predict the class for each instance in the data (right). For how many instances does it predict the incorrect class when: (1) k = 1 (i.e., using the nearest neighbor) and (2) k = 3 (i.e., using the three nearest neighbors).



(a) Training data. Plus signs and circles denote the two different classes.



(b)Testing dataset (triangles denote instances to classify). The set of testing instances has identical values to the set of training instances. But note that they are *not* the same set of instances.

Answer 8:

When using k = 1, there is 0 incorrect class predicted. Because the testing dataset has identical values to the training dataset, so the nearest neighbor of an instance is the instance with the same value in training dataset. Since they are at the same place, so it can always predict the correct class.

When using k = 3, there is 1 incorrect class predicted, which is the instance at (1,0). By selecting three nearest neighbors, there are two plus signs and one circle. The majority class of 3 nearest training instances is Plus sign, so it will be classified as Plus sign, which is incorrect.

9: (10 points)

Suppose the fraction of undergraduate students (UG) who live off-campus is 10% and the fraction of graduate students (G) who live off-campus is 50%. Now suppose that 20% of the university students are graduate students and the rest are undergraduates.

Answer the following: (1) what is the likelihood, as a fraction or decimal, that a student who lives off-campus is an undergraduate student (i.e., what is P(UG|O), where O denotes off-campus); (2) what is the likelihood, as a fraction or decimal, that a student who lives off-campus is a graduate student (i.e., what is P(G|O), where O denotes off-campus); and (3) based upon your previous two answers, if you discovered that a student lived off-campus, would you—based solely on this information—predict him or her to be an undergraduate student or a graduate student? Use Bayes' Theorem for parts (1) and (2), which states that:

$$P(Y|X) = \frac{P(X|Y)\times P(Y)}{P(X)}$$

For this problem, where Y is a binary variable from the set {0,1}, the specific equations to compute are the following:

$$P(Y=1|X=x) = \frac{P(X=x|Y=1)\times P(Y=1)}{P(X=x|Y=1)P(Y=1)+P(X=x|Y=0)P(Y=0)}$$

$$P(Y=0|X=x) = \frac{P(X=x|Y=0)\times P(Y=0)}{P(X=x|Y=0)P(Y=0)+P(X=x|Y=1)P(Y=1)}$$

Answer 9:

Let Y = 0 → a student is an undergraduate student.
Y = 1 → a student is a graduate student.

Let X = x → a student lives off-campus.

According to the statement:
A student is graduate student who lives off-campus → P(X=x | Y=1) = 0.5
A student is undergraduate student who lives off-campus → P(X=x | Y=0) = 0.1
A student is a graduate student → P(Y=1) = 0.2
A student is a undergraduate student → P(Y=0) = 0.8

(1) What is the likelihood, as a fraction or decimal, that a student who lives off-campus is an undergraduate student (i.e., what is P(UG|O), where O denotes off-campus) ?

According to Bayes' Theorem:
A student lives off-campus is a undergraduate student =
P(Y=0 | X=x) = 0.1*0.8 / (0.1*0.8 + 0.5*0.2) = 0.44444… ~ 44.44%

(2) What is the likelihood, as a fraction or decimal, that a student who lives off-campus is a graduate student (i.e., what is P(G|O), where O denotes off-campus) ?

According to Bayes' Theorem:
A student lives off-campus is a graduate student:
P(Y=1|X=x) = 0.5 * 0.2 / (0.5*0.2 + 0.1*0.8) = 0.55555… ~55.56%

(3) If you discovered that a student lived off-campus, would you—based solely on this information—predict him or her to be an undergraduate student or a graduate student?

Since P(Y=1|X=x) > P(Y=0 | X=x), thus, if I discovered that a student lived off-campus, I would predict him or her to be a **graduate** student.

10: (10 points)
A company asks you for help. They want to select the best clustering algorithm for their data, among a set of many algorithms. They can provide you with a dataset that consists of data (features) about their customers. The dataset has been labeled by a company expert, who labeled each customer record as "HIGH" or "LOW" based on the customer's spending level.
Answer the following: (1) How would you use this data to perform effective clustering (i.e., what parts of the data would you use for training the algorithm and what parts would you use for testing it)? (2) Pleased with your recommendations, the company later returns with a new task. Now they want to select the best classification algorithm, still using the same data as before. How would you use this data to perform classification (i.e., what parts of the data would you use for training the algorithm and what parts would you use for testing it)? What would be your class labels?

Hint: Consider how you would use the features and labels for both clustering and classification. Would they be used the same way for both types of tasks? Would you divide the data into a set with features and a separate set with the labels? Or would you divide the data into two sets of instances, each with features and labels?

Answer 10:

(1) How would you use this data to perform effective clustering (i.e., what parts of the data would you use for training the algorithm and what parts would you use for testing it)?

The goal for clustering is to identify a useful grouping of the instances, that is, we need to discover the labels for each instance by analyzing the data. It should be unsupervised learning. Thus, **I would divide the data into a set with features (no labels) as the training data, and a separate set with the labels as the testing data.**

Using the set with only features, we can group instances based only on information found in the data, like the description of each instance and their relationships. In this case, I would group the data based on the features of the customers and try to label each cluster. Once we finish training the data and get the predicted labels, we can use the set with original labels to compare them with the original labels -- "HIGH" and "LOW", to see if the predicted labels are correct.

(2) How would you use this data to perform classification (i.e., what parts of the data would you use for training the algorithm and what parts would you use for testing it)? What would be your class labels?

The goal for classification is to select correct class for a new instance, that is, we need to assign instances to one of the predefined groups. Unlike clustering, it is supervised learning.

The input data for classification should compose of a feature set and class label. After training the data, we will get a function that maps each instance feature to a class label. Thus, **I would divide the data into two sets of instances, each with features and labels. One set is used for training purpose, another is used for testing purpose.**

In this case, I would train the data to explain what features define a “HIGH” or “LOW”, which means “HIGH” and “LOW” would be my class labels. After I get the learning model, I would pick a customer from the test set and apply the model to see what group he/she belongs to.