CSE 347/447 Data Mining: Midterm Exam

3:00 PM to 5:00 PM, April 13, 2022

Exam Invigilator: Lifang He & Kai Zhang

Standard and General Requirements

Please read the following instructions carefully before starting the exam.

- This exam is an individual effort. Please do not discuss or share your answers with others.
- Keep silent during exam. Do not disturb others.
- If you have any difficulty with the exam form, please raise your hand and wait for help.
- Write readably and clearly! Answers that cannot be read can obviously not result in any points and unclear formulations can be misunderstood.
- Assumptions outside of what is stated in the question must be explained. Any assumptions made should not alter the given question.
- Responses with short answer, however, a bulleted list is insufficient. Need to balance brevity and completely answering the question in your own words.
- **Late submission is not accepted**.

Good luck !

0 Affirmation of Academic Integrity

I affirm that I understand Lehigh's Code of Conduct and Academic Integrity expectations; I understand that all suspected violations (including cheating, collusion, and plagiarism) will be submitted to the Office of Student Conduct & Community Expectations; and I pledge that the work I submit will be my own.

Enter your full name as your signature: _____

1 Multiple-Choice Questions (15 pts)

Note: Unless instructed otherwise, there is only ONE answer to each question.

1. Which of the following statements is NOT TRUE?
   A. Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
   B. Noise is never desirable or interesting.
   C. Leave-one-out cross validation is K-fold cross validation, where $K = 1$.
   D. Model performance evaluation and hyperparameter tuning cannot be done using the same cross validation loop.
   Answer: C. For option C, it should be $K = N$, where N is the number of samples. For option D, they should be done using different for-loops. Remember that the cross validation splits dataset into two sets, you can either use it for hyperparameter tuning (training/validation) or model performance evaluation (training/testing).

2. The discriminating between spam and ham e-mails is a classification task.
   A. True

B. False
Answer: A.

3. Clustering algorithms can be used for data summarization and classification.
   A. True
   B. False
   Answer: A. Note that clustering algorithms can be used for classification when ignoring label information, but classification algorithms cannot be used for clustering as they need the label information which are not available in clustering.

4. Linear regression can be used both for binary classification and regression.
   A. True
   B. False
   Answer: A. When it is a binary classification, linear regression can be used because there is no difference between binary classification and binary regression.

5. Which one is NOT a sample of classification problem?
   A. To predict the category to which a customer belongs to.
   B. To predict whether a customer switches to another providers/brand.
   C. To predict the amount of money a customer will spend in one year.
   D. To predict whether a customer responds to a particular advertising campaign or not.
   Answer: C. Option C is a regression problem.

6. Which of the following statements are TRUE about Logistic Regression? (**select all that apply**)
   A. Logistic regression can be used both for binary classification and multi-class classification.
   B. Logistic regression is analogous to linear regression but takes a categorical/discrete target field instead of continuous numeric values.
   C. Logistic regression cannot predict continuous outcomes.
   Answer: A, B, C.

7. Which of the following examples is/are a sample application of Logistic Regression? (**select all that apply**)
   A. Customer's propensity to purchase a product or halt a subscription in marketing applications.
   B. Likelihood of a patient having cancer.
   C. Estimating the blood pressure of a patient based on her symptoms and biographical data.
   Answer: A, B. Option C is a regression problem.

8. Which one is TRUE about the K-Nearest Neighbor (KNN) algorithm?
   A. KNN is a classification algorithm that takes a bunch of labelled points and uses them to learn how to label other points.
   B. KNN algorithm can be used to estimate values for a continuous target.
   C. KNN algorithm can be used for clustering task.
   Answer: A. Similar to the explanation given in Q3: KNN is a supervised method that cannot be used for clustering task.

9. Which statement is NOT TRUE about K-means clustering?
   A. K-means divides the data into non-overlapping clusters without any cluster-internal structure.
   B. The objective of K-means is to form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.
   C. As K-means is an iterative algorithm, it guarantees that it will always converge to the global optimum.
   Answer: C. Iterative algorithm is easily to fall into the local optimum trap.

10. Confusion matrix is used to measure _____ of data mining models.
    A. Speed
    B. Scalability
    C. Accuracy
    Answer: C

11. "Clustering" information can be obtained through data mining using which of the following methodologies?
A. KNN
B. Neural networks
C. All of the above
D. None of the above
Answer: B. As we mentioned in class that neural network can handle clustering task by using similarity loss, just similar to traditional machine learning models. This is also one of the most attractive parts of deep learning.

12. "Clustering" information can be obtained through data mining using which of the following methodologies?
A. Hierarchical clustering
B. Neural networks
C. All of the above
D. None of the above
Answer: C

13. "Classification" information can be obtained through data mining using which of the following methodologies?
A. Decision tree
B. Neural networks
C. All of the above
D. None of the above
Answer: C

14. With a 7x7 input and 3x3 filter kernel, when stride is 3, which padding we should use?
A. Pad = 0
B. Pad = 1
C. Pad = 2
Answer: B. $(N - F + 2P)/S + 1$. Replace the options to this equation, you will see that only pad 1 can get an integer output size.

15. For a 2D convolution, when filter is 3x3, input feature size is 5x5, stride is 2, the output feature size will be:
A. 3x3
B. 2x2
C. 7x7
Answer: B. $(N - F/stride) + 1 = ((5 - 3)/2) + 1 => 2$

2 Short-Answer Questions (20 pts)

(a) (**2pts**) Discuss how you might map correlation values from the interval $[-1, 1]$ to the interval $[0, 1]$.
Answer:
Solution 1: Min-max normalization (**see Lec. 3, pp. 33**), where $new\_min = 0$, and $new\_max = 1$. $x_{new} = \frac{x - min_x}{max_x - min_x}(new\_max_x - new\_min_x) + new\_min_x = \frac{x-(-1)}{1-(-1)}(1 - 0) + 0 = \frac{x+1}{2}$.
Solution 2:

$$x_{new} = \begin{cases} corr & \text{if } corr > 0 \\ 0 & 0 \leq corr < 0 \end{cases}$$

Solution 3: Use the Sigmoid function (**see Lec. 14, pp. 55**), as its output value is between 0 and 1.

Note that the type of transformation that you use might depend on the application. For the first solution, it is a general way to do the data normalization. For the second solution, it is often used for time series clustering as time series with relatively high positive correlation should be put together. For the third solution, it can be used to transform the data to the probability - e.g., for the entropy loss, we need to restrict input as probabilities, we can apply sigmoid function to transform it.

(b) (**2pts**) Explain why we do not need to consider the eigenvector corresponding to the smallest eigenvalue for spectral clustering.
Answer: The eigenvector corresponding to the smallest eigenvalue is a constant vector with all the same elements, and there is no discriminative power.

(c) (**2pts**) Assume the data dimension is $d = 1$, what is the complexity of the KNN algorithm as a function of the number of elements in the training set (n), and the number of elements (m) to be classified?
Answer: $O(nm)$.

(d) (**2pts**) Discuss issues that are important to consider when employing a Decision Tree-based classification algorithm.
Answer:

- How to split nodes (binary split, multiway split).
- How to evaluate how good splits are (GINI-meassure, entropy).
- Stopping conditions.

(e) (**3pts**) Discuss when to use (linear) hard-margin SVM, (linear) soft-margin SVM and nonlinear SVM?
Answer:

- Hard-margin SVM: Linear separable case.
- Soft-margin SVM: Nearly linear separable case but with few noisy data (try to find a line to separate, but tolerate one or few misclassified data points).
- Nonlinear SVM: Linear non-separable case or not linearly separable case (try to find a non-linear decision boundary by kernel trick).

(f) (**3pts**) Explain the purpose of splitting data into training, validation, and test data.
Answer: The training set is used to train the model, the validation/test set is used to validate it on data it has never seen before. In particular, the validation set is used to select hyperparameters in a model, and the test set is to test the ability of a model to predict new data or evaluate the generalization ability of a model.

(g) (**2pts**) To train a deep neural network, suppose we have 1000 training samples, and the batch size is 200, how many iterations are required to complete 1 epoch?
Answer: 1000/200 = 5 (Iterations per epoch = Number of training samples / Batch size).

(h) (**4pts**) What are the main advantages and disadvantages of Deep Learning (please give at least two points for each)?
Answer:

- **Pros:**
  - Features are automatically deduced and optimally tuned for desired outcome.
  - The deep learning architecture is flexible to be adapted to different applications and data types.
  - Massive parallel computations can be performed using GPUs and are scalable for large volumes of data.
  - ...
- **Cons:**
  - It requires very large amount of data in order to perform better than other techniques.
  - It is extremely expensive to train due to complex data models. Moreover deep learning requires expensive GPUs and hundreds of machines. This increases cost to the users.
  - There is no standard theory to guide us in selecting right deep learning tools as it requires knowledge of topology, training method and other parameters.
  - ...

3 Data Types in Data Mining (10 pts)

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). (Example: Age in years. Answer: Discrete, quantitative, ratio).

(a) Angles as measured in degrees between 0° and 360°.
Answer: Continuous, quantitative, ratio

(b) Bronze, Silver, and Gold medals as awarded at the Olympics.
Answer: Discrete, qualitative, ordinal

(c) Number of patients in a hospital.
Answer: Discrete, quantitative, ratio

(d) Military rank.
Answer: Discrete, qualitative, ordinal

(e) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
Answer: Discrete, qualitative, nominal

4 Similarity/Distance Measures (15 pts)

1. (**10pts**) Compute the cosine measure $\frac{\langle X,Y\rangle}{\|X\|\|Y\|}$ using the raw frequencies between the following two sentences:
S1: "The sly fox jumped over the lazy dog."
S2: "The dog jumped at the intruder."
Show the intermediate steps as well as the answer.
[*Hint*] The lexicon/vocabulary here is {the, sly, fox, jumped, over, lazy, dog, at, intruder}. Convert S1 and S2 to frequency vectors. And then computer the similarity measure.

   You may need to use the following approximations:

N	2	3	5	6	7	8	10
$\sqrt{N}$	1.4	1.7	2.2	2.5	2.7	2.8	3.2

   Answer: The unique words in both sentences are: the, sly, fox, jumped, over, lazy, dog, at, intruder.
For sentence 1, the frequency of each word is:
the: 2; sly: 1; fox: 1; jumped: 1; over: 1; lazy: 1; dog: 1; at: 0; intruder: 0.
For sentence 2, the frequency of each word is:
the: 2; sly: 0; fox: 0; jumped: 1; over: 0; lazy: 0; dog: 1; at: 1; intruder: 1.
Therefore, frequency vector of sentence 1 is: S1 = [2, 1, 1, 1, 1, 1, 1, 0, 0] and frequency vector of sentence 2 is: S2 = [2, 0, 0, 1, 0, 0, 1, 1, 1]. We can get:

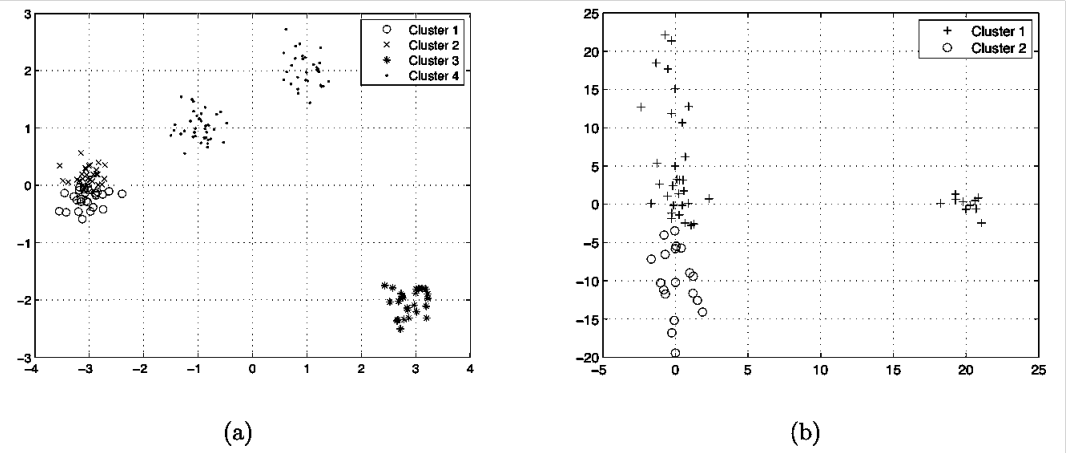
$$Cosine(S1, S2) = \frac{\langle S1, S2\rangle}{\|S1\|\|S2\|} = \frac{6}{\sqrt{10} \times \sqrt{8}} = 0.67$$

2. (**3pts**) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, would be more appropriate for comparing the genetic makeup of two organisms. Explain why. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)
Answer: Jaccard is appropriate for this situation, since we need to perceive what number of qualities these two organisms share.

3. (**2pts**) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance or the Jaccard coefficient? Explain why. (Note that two human beings share $> 99.9\%$ of the same genes.)
Answer: You would use Hamming distance, since it focuses on their differences.

5 Clustering: K-Means [10 pts]

In the figures below, two bad clusterings based on K-means are shown. What is the main reason for the bad results, and what can be done to address the problems?



Answer:

(i) (**5pts**) In figure (a), the problem is caused by bad initial values for the clusters (**see Lec. 5, pp. 33**), as you can see that cluster 1 and cluster 2 are too close but each of them are not broken down into smaller pieces. The easiest solution is to make several runs with different initial values and choose the clustering with the smallest mean distance to the centroids.

(ii) (**5pts**) In figure (b), the problem is caused by the difference in size and density of the two clusters (**see Lec. 5, pp. 34-35**). The problem can be solved by using bisecting K-means (i.e., divisive hierarchical clustering) – similar to K-means, we first initialize K centroids. After which we apply regular K-means with $K = 2$. We keep repeating this bisection step until the desired number of clusters are reached.

6 Classification: KNN [15 pts]

After a data mining course the results of the exam was recorded along with some data about the students. The results can be found in the table below. (GPA is the Grade Point Average.)

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
1	555 - 3452	Java	No	3.1	Yes
2	555 - 6294	Java	No	2.0	No
3	555 - 9385	C++	Yes	3.5	Yes
4	555 - 9387	Python	Yes	2.5	Yes
5	555 - 9284	Java	Yes	3.9	No
6	555 - 0293	C++	No	2.9	No
7	555 - 9237	Java	No	1.9	No
8	555 - 3737	Python	Yes	3.2	Yes

(a) (**10pts**) Describe the design of a K-Nearest Neighbor (KNN) classifier to predict if a student will fail or pass the exam.
[*Hint*]: Think about which variables should be considered as data features; which distance measures should

be used in case of different data types.

Answer: The ID and Phone number are unrelated to a students capacity to pass the exam so they are discarded. The language category has three different values C++, Java and Python. Since the language values are nominal rather than ordinal we will consider the distance between two languages to be 1 if they are different and 0 if they are the same.

Passed all assignments is a binary category this means that we can easily use the same idea as for language i.e. if they are different the distance is 1 and if they are the same the distance is 0. The GPA is a quantitative category that ranges from 1.9 to 3.9. For the distance in this dimension we can just take the absolute value of the difference in GPA. To make each category have about the same weight we can multiply both the language and assignments distances with 2.

We end up with the following distance measure:

$$dist(X,Y) = 2 \times diff_{lang}(X,Y) + 2 \times diff_{assign}(X,Y) + |gpa(X) - gpa(Y)|$$

For the value of K both 1 and 3 are reasonable (a larger value for K would be too high since there are so few data points). We will select 3 to make the classifier less sensitive.

(b) (**5pts**) Use your KNN classifier to predict whether the following student (who overslept and missed the original exam) will pass the re-exam.

ID	Phone number	Language	Passed all assignments	GPA	Passed exam
9	555 - 6295	C++	Yes	3.0	?

Answer: We calculate all the distances using our distance formula:

dist(1,9) = 2 + 2 + 0.1 = 4.1

dist(2,9) = 2 + 2 + 1.0 = 5.0

dist(3,9) = 0 + 0 + 0.5 = 0.5

dist(4,9) = 2 + 0 + 0.5 = 2.5

dist(5,9) = 2 + 0 + 0.9 = 2.9

dist(6,9) = 0 + 2 + 0.1 = 2.1

dist(7,9) = 2 + 2 + 1.1 = 5.1

dist(8,9) = 2 + 0 + 0.2 = 2.2

Students 3, 6 and 8 are the three closest neighbours this gives us two votes for yes and one for no. So our prediction is that the student will pass the assignment.

7 Classification: Bayes Rule (15 pts)

Given a training dataset below, predict the class label of $x = [1 \;\; 1]$ using Naïve Bayes model.

$$X = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \qquad y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

[*Hint*] The formula of Naïve Bayes model is

$$
\begin{aligned}
P(y_i|x) &= \frac{P(x|y_i)P(y_i)}{P(x)} \\
&\propto P(x|y_i)P(y_i) \\
&\approx \prod_j [P(x_j|y_i)]P(y_i)
\end{aligned}
$$

Answer: **See Lec. 10, pp. 34-36 for this example**. Based on the Naïve Bayes formula, we have:
$P(y = 0|x = [1 \;\; 1]) = P(x_1 = 1|y = 0) \times P(x_2 = 1|y = 0) \times P(y = 0) = 0.167 \times 0.5 \times 0.6 = 0.05$
$P(y = 1|x = [1 \;\; 1]) = P(x_1 = 1|y = 1) \times P(x_2 = 1|y = 1) \times P(y = 1) = 1 \times 0.5 \times 0.4 = 0.2$
So, the class label of $x = [1 \;\; 1]$ is 1.

8 Bonus Question for Extra Credit (5 pts)

Assume the forward propagation of the convolution on X to Y could be written into $Y = AX$, and $\frac{\partial L}{\partial X}$ and $\frac{\partial L}{\partial Y}$ denote the gradients of loss L w.r.t. X and Y, respectively. For the back propagation, could you figure out whether $\frac{\partial L}{\partial X} = B\frac{\partial L}{\partial Y}$ also corresponding to a convolution on $\frac{\partial L}{\partial Y}$ to $\frac{\partial L}{\partial X}$? If yes, write down the solution of B for this convolution. If no, explain why.
[*Hint*] For $Y = AX$, we have the following gradient: $\frac{\partial Y}{\partial X} = A^{\mathrm{T}}$, where $^{\mathrm{T}}$ denotes the transpose of the matrix.

Answer: Since the back-propagation follows the chain rule, we have $\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial X} = A^{\mathrm{T}}\frac{\partial L}{\partial Y}$, so we can get $B = A^T$.
Note: When solving this problem, the two key points are: chain rule and matrices are multipliable. Below lists the sizes of relevant matrices: $Y : m \times n$; $A : m \times r$; $X : r \times n$; $\frac{\partial L}{\partial Y} : m \times n$; $\frac{\partial L}{\partial X} : r \times n$.