CptS 315 Introduction to Data Mining
Midterm Exam 1, Spring 2020

Exam date: Mar 12

Your Name and WSU ID:

Instructions.

- The maximum score of the exam is 100 points.
- Read all the questions before starting to answer. Try to answer those questions, which you think are easy from your perspective first. You can select a subset of questions that are worth 100 points to answer.
- Work efficiently. Most questions don't require much work. If you are spending more than 3 mins on any question, then you should try to re-think about it.
- Keep your answers short and simple.
- **Short Questions (25 points)**

  Please keep your answers short (one sentence). For the True / False questions, please also provide a short justification.

  0. **(2 points)** Incredibly hard question! What is the course number for this data mining class?

  1. **(5 points)** Park-Chen-Yu (PCY) algorithm is always more efficient than Apriori algorithm for finding frequent item-pairs (True/False)

  2. **(5 points)** User-user collaborative filtering and item-item collaborative filtering are duals of each other. Therefore, they both are equally accurate when applied to real-world applications. (True/False)

3. **(5 points)** A classifier trained on more training data is more likely to over-fit (True / False)

4. **(3 points)** Passive learning is label-efficient when compared to active learning protocol (True/False)

5. **(5 points)** If you are given m data points, and use half for training and half for testing, the difference between training error and testing error decreases as m increases. (True/False) Please provide one sentence justification.

- **Frequent Itemset and Association Rule Mining (30 points)**

  6. **(6 points)** Suppose the support of $\{A\}$ is 5, support of $\{B\}$ is 7, support of $\{A, B\}$ is 4, support of $\{B, C\}$ is 3, support of $\{A, C\}$ is 4, and support of $\{A, B, C\}$ is 2. What is the confidence of following association rules?

  6.1 $A \Rightarrow \{B, C\}$

  6.2 $\{A, B\} \Rightarrow C$

  7. **(5 points)** Describe the key property that is exploited by the Apriori algorithm for efficiently computing the frequent itemsets.

8. **(4 points)** Suppose we have a total of 100 items. The number of frequent items equals 10. How many candidate pairs will Apriori algorithm consider for counting in Pass 2?

9. **(5 points)** For Park-Chen-Yu (PCY) algorithm to improve over standard Apriori algorithm, the hash table in Pass 1 must eliminate at least the following fraction of candidate pairs

a) 1/3

b) 1/2

c) 2/3

d) 3/4

10. **(5 points)** The concept of *negative border* is relevant in the context of following algorithm:

a) Apriori

b) Park-Chen-Yu (PCY)

c) Toivonen's Algorithm

d) Savasere-Omiecinski-Navathe (SON) Algorithm

11. **(5 points)** When do you say an itemset I is in the *negative border*?

- **Recommender Systems (20 points)**

12. **(5 points)** What is the key idea behind content-based filtering algorithm to answer the basic filtering question: "will user U like item X?"

a) Look at what items U likes, and then check if X is similar to those items

b) Look at which users like X, and then check if U is similar to those users

13. **(5 points)** Root Mean Squared Error (RMSE) is the appropriate metric to evaluate the effectiveness or predictions of recommendation algorithms. (True/False)

14. **(5 points)** When applying the recommendation algorithms, what do we achieve by normalization of each row of the utility matrix (subtract the mean from rating values)?

15. **(5 points)** Please list two drawbacks for both content-based filtering and collaborative filtering approaches.

- **Machine Learning (25 points)**

16. **(8 points)** Suppose we have a binary classification data with classes $Y \in \{+1, -1\}$ and d features with each feature $f_i \in \{+1, -1\}$. To improve the performance of the classifier, Jana decided to duplicate each feature. Hence, each training example now has $2d$ features with $f_{d+i} = f_i$ for $i = 1, 2, \cdots, d$. This question is about comparing the training problem with *original* feature set and *double* feature set. Assume that there are same number of training examples for both positive and negative class, and in case of ties, you will chose positive class.

For a Perceptron classifier, select all that apply.

a) Test accuracy with original feature set could be higher

b) Test accuracy with double feature set could be higher

c) Test accuracy will be same with both original and double feature set

Please write one sentence justification

Consider the following Perceptron, for which the inputs are the always "1" feature and two binary features $x_1 \in \{0, 1\}$ and $x_2 \in \{0, 1\}$. The output label $y \in \{0, 1\}$. Suppose w_0, w_1, w_2 stands for weights of the three features. The classification decision is made as follows: $y = 1$ if $(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2) > 0$. Otherwise, $y = 0$.

17. **(5 points)** Which of the following choices for the weight vector (w_0, w_1, w_2) can classify y as $y = (x_1 \text{ XOR } x_2)$? XOR is the logical exclusive OR operation, which equals to ZERO when x_1 equals to x_2, and equals to ONE when x_1 is different from x_2.

a) (1, 1, 0)

b) (-2, 1, 1.5)

c) Any weights that satisfy $(-w_1 - w_2) < w_0 < min(0; -w_1; -w_2)$

d) No weights can compute the XOR logical relation

18. **(6 points)** Which of the following choices for the weight vector (w_0, w_1, w_2) can classify y as $y = (x_1 \text{ AND } x_2)$? Here AND refers to the logical AND operation, which equals to ONE when $x_1 = 1$ and $x_2 = 1$, and equals to ZERO for all other combinations.

a) (1, 1, 0)

b) (-1.5, 1, 1)

c) (-2, 1, 1.5)

d) Any weights that satisfy $(-w_1 - w_2) < w_0 < min(0; -w_1; -w_2)$

e) No weights can compute the AND logical relation

19. **(3 points)** As the number of passes over training data increases for perceptron based learning, which of the following are False?

a) training accuracy increases

b) number of mistakes decreases

c) training accuracy decreases

d) number of mistakes increases

Please write one sentence justification

20. **(3 points)** As the number of training examples used to learn a linear classifier are increased, which of the following are False?

a) training accuracy decreases

b) testing accuracy decreases

c) training accuracy increases

d) testing accuracy increases

Please write one sentence justification

Extra Sheet

Extra Sheet