CptS 315 Introduction to Data Mining
Final Exam, Spring 2020

Exam date: May 8

Your Name and WSU ID:

Instructions.

- Please write your name and WSU ID in the above space.
- 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.
- Work efficiently. Most questions don't require much work. If you are spending more than 5 mins on any question, then you should try to re-think about it.
- Keep your answers short and simple.
- **Short Questions (20 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 your favorite data mining algorithm?

  1. **(3 points)** In association rule mining, generating the frequent itemsets is the most computationally-expensive step (True/False)

  2. **(5 points)** When deploying anomaly detection systems in practice, we want to have low false-positive rate (True/False)

  3. **(5 points)** Isolation Forest algorithm for anomaly detection need supervised training data (True/False)

4. **(5 points)** Which of the following methods can achieve zero training error on *any* linearly separable dataset?

a) Decision Tree

b) 15-nearest neighbors

c) Perceptron

- **Frequent Itemset Mining and Recommender Systems (20 points)**

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

6.1 $A \Rightarrow B$

6.2 $B \Rightarrow A$

7. **(4 points)** Suppose 50 percent of the item pairs have non-zero counts. Which of the following methods is preferable for counting item pairs in main memory?

a) Triangular matrix method

b) Tabular method

Please write one sentence justification

9. **(5 points)** Suppose you have a real-world application with 1 billion items and 100 billion baskets. You have access to lot of of parallel computing resources. Which of the following frequent itemset mining algorithms will you employ?

a) Apriori algorithm

b) Park-Chen-Yu (PCY) algorithm

c) SON algorithm

d) Toivonen algorithm

Please write one sentence justification

10. **(5 points)** What is the key idea behind collaborative 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

- **Multi-Class Perceptron (10 points)**

11. **(10 points)** Suppose we are training a Perceptron for a three-class (*good*, *bad*, *ugly*) problem. Each training example has 4 features. The weights are currently.

w_{good} = (-1, -1, -1, -1) for class *good*
w_{bad} = (-1, +1, +1, -1) for class *bad*
w_{ugly} = (-1, -1, -1, -1) for class *ugly*

Consider the training example $x = (-1, +1, +1, +1)$ with correct label *good.*

a) Which classification label is predicted for the training example x with the current weights?

b) What are weights (w_{good}, w_{bad}, w_{ugly}) after the update that incorporates the training example using a learning rate of 1?

- **Decision Trees, Nearest Neighbor Classifiers, and Clustering (20 points)**

12. **(4 points)** What strategies can help reduce over-fitting in decision trees?

a) Pruning

b) Enforce a minimum number of examples in leaf nodes

c) Make sure each leaf node is one pure class

d) Enforce a maximum depth for the tree

14. **(4 points)** The depth of a learned decision tree can be larger than the number of training examples. (True/False) Give a short justification if true and a contradiction otherwise.

15. **(4 points)** The training error of 1-NN (nearest neighbor) classifier is ZERO. (True/False) Please provide one sentence justification.

16. **(4 points)** Suppose we have 1 billion data points. Which clustering algorithm among *Hierarchical Agglomerative Clustering (HAC)* and *K-Means* is more computationally-efficient to apply? Why?

17. **(4 points)** The K-Means clustering algorithm will automatically find the number of appropriate clusters in the data (True / False) Please provide one sentence justification.

- **Miscellaneous (15 points)**

  18. **(5 points)** In Isolation Forest algorithm for anomaly detection, what is the relation between path lengths for normal data instances and path lengths for anomalous data instances?

  20. **(5 points)** Consider a decision tree built from an arbitrary training data. Suppose the class label can take 2 values: +1 and -1 (binary classification).

  a) What is the maximum training set error (expressed as a percentage) that any data set could possibly have?

b) Construct a simple example dataset that achieves this maximum training set error. (It must have less than or equal to 2 features)

21. **(5 points)** Which of the following are true for Bagging?

a) In bagging, we choose random subsamples of the input training examples with replacement

b) In bagging, we choose random subsamples of the input training examples without replacement

c) Bagging is ineffective with perceptron, because all of the learners learn exactly the same decision boundary

d) Bagging only works with decision trees

e) Bagging only works when the classifiers in the ensemble have diversity

- **Bloom Filter and Computational Advertising (15 points)**

23. **(7 points)** You are given a Bloom filter that consists of $m = 11$ memory bits and two hash functions h_1 and h_2 defined as follows: $h_1(x)$=$3x$ mod m and $h_2(x)$=$2x$ mod m, where x is a given stream element. Assume that all m bits of the Bloom filter are initially set to 0.

Show the Bloom filter bits following the insertion of the the following three elements: 7, 12, 9. Show result of Bloom filter after each insertion.

24. **(8 points)** Suppose we apply the BALANCE algorithm with bids of 0 or 1 only, to a situation where advertiser A bids on query words x and y, while advertiser B bids on query words x and z. Both have a budget of 2 dollars. Identify the sequences of queries that will certainly be handled optimally by the algorithm, and provide a one line explanation.

(a) y, z, y, y

(b) x, y, z, x

(c) y, y, x, x

(d) x, y, y, y

Extra Sheet

Extra Sheet