CSCE 587 Big Data Analytics Spring, 2017 In Class Midterm Exam

Name ______________________________

1. [15/20 points] Association Rules:
   a. Explain how the *a priori* principle can help reduce the search for frequent item sets.

   In constructing K+1 item sets from K item sets, no need to consider K-length itemsets that are not frequent. The a priori principle tells us no superset of an infrequent set can be frequent.

   b. If we have data on 21 unique items, how many items sets are there if we exhaustively enumerate them? 2^{21}

   c. The *confidence* and lift measures are used to filter out association rules.
      i. Which of these two is a symmetric measure? Lift is symmetric, confidence is not
      ii. [Grad/Honors only] Both are ratios. Explain the conceptual difference in these two types of ratios. $A \to B$ $s(A,B)/s(A) \cdot s(B)$ vs $s(A,B)/s(A)$

Lift measures the probability of A+B occurring together ~~as the ratio~~ comparing it with the expected value if A+B were independent. ⇒ correlation

Confidence measures the probability of A + B occurring together, compared to the probability of A, it does not measure A's dependence on B.

2. [10 points] K-means clustering: In using the "elbow" method, we employ a plot depicting different values for different choices of k. Explain what those values represent?

The values are the within-sum-of-squares. It measures the coherence of the clusters.

3. [15/20 points] T-Test:
   a. What determines the Degree of Freedom (DoF) in a two sample T-test? **EXPLAIN!**

   The DoF depends on the number of observations of the two samples, i.e. n1 + n2 − 2 where n1 = # of observations in S1
   n2 = # of observations in S2

   b. What is the null hypothesis in the case of the two sample t-test?

   $\mu_{S1} = \mu_{S2}$

   c. [Grad/Honors only] The p-value is used as a threshold. What is the definition of the p-value?

   The p-value is the probability of ~~observing the value~~ finding the observed or more extreme value when the null hypothesis is true.

4. [15 points] Model Evaluation: In the case of Decision Trees, the classical approach is to partition the data set into a training set, a test set, and a validation set. Explain the role of each of these data sets.

Training set role:

This is the data used to learn the model.

Test set role:

This is the data use to provide an unbiased evaluation of the final model

Validation set role:

This data set is used to detect overfitting

5. [20 points] Logistic Regression:
   a. What does the exponent of the regression coefficient, exp(β_1), represent in a logistic regression model?

      The odds ratio

   b. If the probability of passing this test 75%, what are the odds of passing this test compared to not passing this test?

      3:1

6. [10 points] Linear Regression: If the mean of the residual is close to zero, does that mean we have a good fit? Explain.

Not necessarily, If a linear model is not appropriate then the fit well not be good.

7. [5 points] Model Evaluation: The "leave-one out" approach is an extreme case of *k*-fold cross validation. List two drawbacks to choosing "leave-one out" as opposed to something like 10-fold cross validation.

1) stratification is not possible, not necessarily a problem

2) computationally very expensive

3) Tends over estimates performance of model.