CSCE/STAT 587 Big Data Analytics Fall, 2016 In Class Final Exam

Name _____

1. [15 points] Time Series Analysis: The following equation can be used to model a time series:

$$Y_t = T_t + S_t + R_t, \quad t=1,...,n$$

   a. [5] T_t is called the trend term. Explain what concept this term represents. (Is this term periodic?)

   b. [5] S_t is called the seasonal term. Explain what concept this term represents. (Is this term periodic?)

   c. [5] What do you think the term R_t represents?

2. [25 points] HDFS: Circle one answer per question.
   a. When is the data from a data-file distributed to data nodes in HDFS?
      i. When the data file is loaded to HDFS.
      ii. When the map portion of a map-reduce job is run.
      iii. When the reduce portion of a map-reduce job is run.

   b. When are key-value pairs used in map-reduce?
      i. As a return value from the map function.
      ii. As a return value from the reduce function.
      iii. As a return value from both the map and reduce functions.

   c. What best describes the relation between HDFS running on a linux machine?
      i. HDFS can be examined from the linux file system
      ii. The linux file system can be examined from HDFS
      iii. HDFS and linux file systems occupy completely separate name space.

   d. Directory information in HDFS is:
      i. Distributed among the data nodes
      ii. Stored in the name node
      iii. Held by task tracking nodes
      iv. Held by job tracking nodes

   e. What best describes the file model in HDFS?
      i. Allows concurrent file updates
      ii. Write once, read many
      iii. Allows append, but only by one process at a time

3. [20 points] Which methods are appropriate for categorical data? Circle your answers.
   a. T-test
   b. Logistic regression
   c. Association rules
   d. Decision Trees
   e. Time series analysis
   f. K-means clustering
   g. Linear regression
   h. Naïve Bayes
   i. rmr2
   j. Anova

4. [5 Points] Which classifiers handle nonlinear data and discontinuities in the input data well? Circle your answer(s).
   a. Decision Trees
   b. Logistic regression
   c. Naïve Bayes

5. [10 points] Explain the difference between lift and leverage in the context of association rules.

6. [25 points]This semester we have explored a number of data analytic methods. For each task below, name at least one method covered this semester that addresses the task.

Task	**Method**
I want to group items by similarity.	
I want to discover relationships between items	
I want to determine the relationship between the outcome and the input variables	
I want to assign (known) labels to objects	
I want to find the structure in a temporal process	

7. [10 points **Graduate Students Only**] SVM kernel functions effectively map from what space to what space? Names these two spaces.