COSC 757 Data Mining Assignment 2

Instructions: This is an individual assignment. Use Blackboard to submit your answers on the due date (no hard copies please). Late submissions will receive a zero grade.

Experimentation with Classification: Choose a dataset that is well suited for classification. You can use any dataset that you would like to classify. A good number of datasets can be found in the UCI machine learning data repository but feel free to use any dataset that you want. Make sure that you select a dataset that has a class variable. Then use a tool such as R, Weka, or RapidMiner to classify the dataset. The specific requirements for the assignment are as follows:

- Choose a dataset that is of interest to you and is well suited for classification
- Describe the dataset
- Perform EDA to understand the variables of the dataset.
- Research at least 3 different classification algorithms. There are many algorithms available for R, Python, Weka, RapidMiner, and KNIME.
    - A good resource for R can be found at the Data Mining Algorithms in R Wikibook
        - http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification
    - Also, the caret package in R would is a great place to start experimenting with classification methods.
- Explain the algorithms that you are using.
- Design an experiment using training and testing (holdout method), cross-validation, or the bootstrap method. Use a statistical test to validate your partition of the data.
- Compare the results of three or more classification methods using the same experimental setup using one or more classification evaluation methods discussed in class. The metrics that you choose are up to you and can include accuracy, error rate, sensitivity, specificity, precision, recall, and F measure.
- Write a report that describes your experiment and results. The report should be in either ACM or IEEE conference paper format and should include an introduction section that details the dataset and the objectives of the analysis; a methodology section that explains the approach that you are using to mine the dataset including the steps used to preprocess the data, the classification algorithms and parameters, experimental setup (e.g. holdout, cross validation, bootstrapping), accuracy metrics (e.g. precision, recall, f-measure, etc...); a results section that shows the results of your analysis and any interesting patterns that you found; and a conclusion section that summarizes your results, discusses the limitations of your approach, and any difficulties that you had with your experiment.
    - Links to format templates:
    - http://www.ieee.org/conferences_events/conferences/publishing/templates.html
    - https://www.acm.org/publications/proceedings-template