CS 5525: Data Analytics I

Homework 5

Due Date: December 2nd, 2016 (4:00PM) **Total: 100 Points**

Problem 1. Baye's Theorem **(10 Points; 2 + 3 + 5)**

Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. It is also given that one-fifth of the college students are graduate students and the rest are undergraduates.

(a) What is the probability that a student who smokes is a graduate student?

(b) Is a randomly chosen smoker more likely to be a graduate or undergraduate student?

(c) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

Problem 2. K-means Clustering **(25 Points; 10+15)**

PART I.

Consider the following set of one-dimensional data points: {0.1, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9}.

(a) Suppose we apply k-means clustering to obtain three clusters, A, B, and C. If the initial centroids are located at {0, 0.25, 0.6}, respectively, show the cluster assignments and locations of the centroids after the first three iterations. Compute the SSE of the k-means solution (after 3 iterations).

(b) Apply bisecting k-means (with k=3) on the data. First, apply k-means on the data with k=2 using initial centroids located at {0.1,0.9} Next, compute the SSE for each cluster (make sure you indicate the SSE values in your answer). Choose the cluster with larger SSE value and split it further into 2 sub-clusters. You can choose the two data points with the smallest and largest values as your initial centroids. For example, if the cluster to be split contains data points (0.20, 0.40, 0.60, and 0.80), then the centroids should be initialized to 0.20 and 0.80. Show the clustering solution produced obtained applying bisecting k-means.

PART II. WEKA – K-means Clustering

Load iris.arff file into Weka. Click on the *Cluster* tab and choose "SimpleKMeans" algorithm for clustering and set "numClusters" to 3. Select "Classes to cluster evaluation" and click on the "Ignore attributes" and select "class". Start the clustering.

(a) How many instances were clustered incorrectly? Provide the confusion matrix.
(b) How many instances are in cluster2? How many of these instances were incorrectly clustered and which cluster they should belong to?
(c) Right-click on the result list and click on “visualize cluster assignments”. Set the x-axis to instance_numbr and y-axis to sepallength. Change the color to class. Which type of iris flower has all instances clustered correctly?

Problem 3. WEKA – Association Rule-Learner

(15 Points)

This exercise aims to familiarize you with an association rule algorithm using Weka. We want to discover association rules in a given dataset by invoking the Apriori algorithm. First thing is to make sure that all the attributes in the dataset are of type nominal, in other words there should not be any numerical attribute(s) in the dataset and if there are any, they should be discretized before you can perform the association rule algorithm.

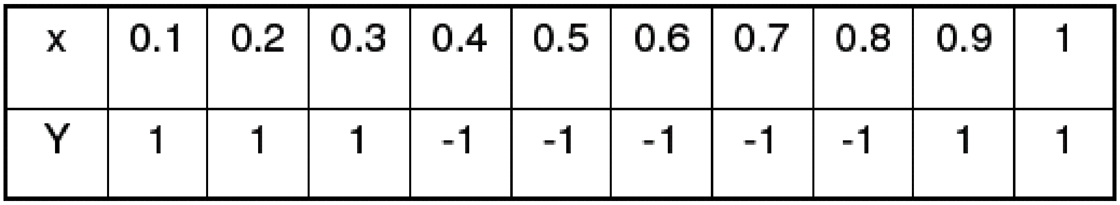
(a) Load zoo.arff into Weka. Find the numerical attribute(s) in this dataset and choose the right filter under *Filter* to discretize the numerical attribute(s). Report the attribute(s) and the filter you used.
(b) Now, go to the *Associate* tab and choose Apriori algorithm and start the algorithm. List 4 interesting rules.
(c) Which rule(s) are going to be always true according to the algorithm and the confidence?

Problem 4. Anomaly Detection **(10 Points; 3 + 7)**

(a) What are the problems with the nearest-neighbor based anomaly detection?
(b) Why does one have to go for the relative density based approach rather than the traditional density based approach for anomaly detection? Explain with a simple example containing two clusters along with an outlier.

Problem 5. Ensemble Methods **(15 Points)**

Let us consider the Boosting algorithm in which all the data points are initially given uniform weights. Given below is a dataset with feature X and response Y.



x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	-1	-1	-1	-1	-1	1	1

We will now consider 3 weak classifiers (Hypotheses) H1, H2 and H3.

H1: X<=0.35 ➔Y=1, else Y= -1

H2: X<=0.75 ➔Y= -1, else Y=1

H3: X<=0.3 or X>=0.95 ➔ Y=1, else Y= -1

Compute the weights of all the instances after the first round of boosting for each of the above mentioned weak classifiers. Use the initial set of (uniform) weights for each classifier. Do not perform normalization of the weights at the end of the iteration.

Problem 6. Regression Analysis **(25 Points: 10+10+5)**

The Energy Efficiency Dataset (ENB2012_data.xlsx) in this problem is a public dataset from UCI (https://archive.ics.uci.edu/ml/datasets/Energy+efficiency). In this data, the energy analysis is performed using 12 different building shapes simulated in Ecotect. They are different in several aspects, such as including the glazing area and the orientation. In the dataset, there are 768 samples and 8 features. In this problem, we will use 7 features, and the goal is to predict the cooling load (Y), which is converted to a binary variable in this problem.

(a) Fit a multivariate linear regression model using 5-fold cross validation to predict the cooling load Y. In the model, please use the backward elimination method to select the best feature subset. In your solution, for each step, please provide the explicit formulation of the fitted linear model and specify which features are considered in the model.

(b) Fit a logistic regression model using 5-fold cross validation to predict the cooling load Y. Give the explicit formulation of the fitted model.

(c) Provide the confusion matrix along with the accuracy to evaluate the performance of each model.

To implement the linear/logistic regression model, you can use the R/Matlab/python or other existing packages. Please submit the code along with your solutions to get full points.