New York University Shanghai

Problem Set 5

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Clustering Algorithms

Please briefly answer the following questions:

(a) Consider the hierarchical clustering method. The training data set of 6 data points with a single covariate X is shown in the following table:

Data Index	1	2	3	4	5	6
X	0	1	3	4	8	9

Please draw the tree for the hierarchical clustering method. If we set the number of clusters to be $k = 3$, which data points does each cluster include under the hierarchical clustering method? We use the Ward's method to calculate the distance between two clusters.

(b) Consider the data sample as part (a).

We apply the $k-$means method with $k = 2$, and the algorithm starts with the initial clusters $\mathcal{C}_1 = \{1, 3, 5\}$ and $\mathcal{C}_2 = \{2, 4, 6\}$. What will be the result of $k-$means clustering algorithm? Please use a figure (or table) to report each step of the algorithm (please clearly specify the clusters and the centers in each step) to reach the final clustering output.

2. Cluster-Then-Predict

In this problem, you will use the data set `Stock.csv`, which contains monthly stock returns from the NASDAQ stock exchange, to do clustering. Then, you will use the clustering results to make some predictions about future stock returns. The variables in this data set are listed as follows:

- $ReturnJan$ = the return for the company's stock during January (in the year of the observation)
- $ReturnFeb$ = the return for the company's stock during February (in the year of the observation)
- $ReturnMar$ = the return for the company's stock during March (in the year of the observation)
- *ReturnApr* = the return for the company's stock during April (in the year of the observation)
- $ReturnMay$ = the return for the company's stock during May (in the year of the observation)
- $ReturnJune$ = the return for the company's stock during June (in the year of the observation)
- $ReturnJuly$ = the return for the company's stock during July (in the year of the observation)
- *ReturnAug* = the return for the company's stock during August (in the year of the observation)
- *ReturnSep* = the return for the company's stock during September (in the year of the observation)
- *ReturnOct* = the return for the company's stock during October (in the year of the observation)
- $ReturnNov$ = the return for the company's stock during November (in the year of the observation)
- *PositiveDec* = whether or not the company's stock had a positive return in December (in the year of the observation). This variable takes value 1 if the return was positive, and value 0 if the return was NOT positive.

Please solve the following questions.

(a) Train and validate L_1-regularized logistic regression models to predict *PositiveDec* (set $t = 0.5$). Report the out-of-sample overall accuracies and ROC-AUCs of your trained models.

(b) Now, please try to cluster the stocks. Do we need to remove the dependent variable *PositiveDec* before running a clustering algorithm? Why or why not?

(c) Next, please standardize the features by subtracting the mean and dividing by the standard deviation. Why should we standardize the features?

(d) Run k-means clustering with 3 clusters for the training set. Report the number of observations in each cluster, and the average return of the stocks in January in each cluster.

(e) Make predictions about which cluster the data in the testing set should belong to. Report the number of observations (of the testing set) in each cluster, and the average return of the stocks in January (of the testing set) in each cluster.

(f) Split the training set into three sub-training-sets according to which cluster each training data point belongs to. Also split the testing set into three sub-testing-sets according to which cluster each testing data point belongs to. Please train and validate cluster-specific logistic regression models (set the threshold value $t = 0.5$ for each cluster). Report the out-of-sample overall accuracies and ROC-AUCs of all your cluster-specific models. Please also compare the overall accuracy of the cluster-specific models for **the entire testing set** with the best logistic regression model built in part (a). Do you think clustering helps improve the prediction accuracy?