Homework 4

Yu-Ru Lin

University of Pittsburgh
INFSCI 2160: Data Mining

yurulin@pitt.edu

2018-02-28

Homework 4 I

You will use two datasets, “stock price” (modified from dataset Dow Jones Index), and the “congress vote” data for this assignment.

- Prepare your report using R Markdown and save into *.html file. Submit your *.html and your original*.Rmd via courseweb.
- Due: 2018-03-13 11.59pm

Task 1: analyze the data “stock price”

Data description: stock_price_description.txt

Data csv: stock_price.csv

Homework 4 II

Original source: http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index

The objective of the analysis is to group stocks together if they have similar trends in price variance.

1. Use PCA to reduce the dimension of stock-price information. Generate a screeplot and determine the number of principle components based on this plot. Plot the loadings for first principal component.
2. Generate a scatterplot to project stocks on the first two principal components.
3. Generate an MDS map to plot stocks on a two-dimensional space.

Homework 4 III

4. Use k-means and hierarchical clustering to group stocks. Specifically, you will generate 8 MDS maps for the stocks and color the stocks based on different clustering methods (k-means, h-clustering with single-link, h-clustering with complete-link, h-clustering with average-link) and different number of clusters ($k = 3$, $k = 6$). For each hierarchical clustering method, generate a dendrogram.
   hint: Standardize the data before performing clustering

Task 2: analyze US Senator Roll Call Data
The objective is to identify and visualize the clustering patterns of senators' voting activities.

Homework 4 IV

Download and load the data from:

http://www.yurulin.com/class/spring2018_datamining/data/roll_call

These are data for Senate roll call votes for the 101st through 113th Congresses (as of March 2015). Each row corresponds to a voter in the US Senate. The first nine columns of the data frame include identification information for those voters, and the remaining columns are the actual votes. See the codebook for the 101st Congress for explanation of what is contained in each of the first nine columns at:

http://www.voteview.com/senate101.htm

1. Create a senator-by-senator distance matrix for the 113th Congress. Generate an MDS plot to project the senators on the two dimensional space. Use shapes or colors to differentiate the senators' party affliation

Homework 4 V

2. Use k-means and hierarchical clustering to group the senators, and color the senators on the MDS plots based on the clustering results (you will use k-means, h-clustering with single-link, h-clustering with complete-link, h-clustering with average-link and k=2).
3. Compare the clustering results with the party labels and identify the party members who are assigned to a seemly wrong cluster. Specifically, based on the k-means results, which Republicans are clustered together with Democrats, and vice versa? And based on the three variants (single-link, complete-link and average-link), which Republicans are clustered together with Democrats, and vice versa?

Homework 4 VI

4. Compute the purity and entropy for these clustering results with respect to the senators' party labels. You will generate a 2x4 table as follows:

	k-means	hclust-single	hclust-complete	hclust-average
Purity				
Entropy				

5. Based on your observation on both measures and mis-classified members, choose two clustering methods that generate the most meaningful results and explain why.