Joseph Barbati
CS378
March 26, 2017

Assignment 4 – Clustering

Dataset1.csv

Info

This was pulled from the output of a classifier about knowledge gained.

Attribute Information:

STG (The degree of study time for goal object materails), (input value)
SCG (The degree of repetition number of user for goal object materails) (input value)
STR (The degree of study time of user for related objects with goal object) (input value)
LPR (The exam performance of user for related objects with goal object) (input value)
PEG (The exam performance of user for goal objects) (input value)
UNS (The knowledge level of user) (target value)

Very Low: 50
Low:129
Middle: 122
High 130

Dataset2.csv

Info

Each record (row) is data for a week. Each record also has the percentage of return that stock has in the following week (percent_change_next_weeks_price).

Attribute Information (taken from UCI website):

quarter: the yearly quarter (1 = Jan-Mar; 2 = Apr=Jun).
stock: the stock symbol (see above)
date: the last business day of the work (this is typically a Friday)
open: the price of the stock at the beginning of the week
high: the highest price of the stock during the week
low: the lowest price of the stock during the week
close: the price of the stock at the end of the week
volume: the number of shares of stock that traded hands in the week
percent_change_price: the percentage change in price throughout the week
percent_chagne_volume_over_last_wek: the percentage change in the number of shares of
stock that traded hands for this week compared to the previous week
previous_weeks_volume: the number of shares of stock that traded hands in the previous week
next_weeks_open: the opening price of the stock in the following week
next_weeks_close: the closing price of the stock in the following week
percent_change_next_weeks_price: the percentage change in price of the stock in the
following week days_to_next_dividend: the number of days until the next dividend
percent_return_next_dividend: the percentage of return on the next dividend

Joseph Barbati
CS378
March 26, 2017

I ignore the second third attributes as the second I would consider the class and the third I would consider irrelevant.

Preprocessing and Arguments

I didn't preprocess the data as both datasets were complete, no missing data.

As for the arguments, besides the required ones (dataset file, k, output file), I added a fourth argument, class_col, which is the column number of the class attribute. For dataset1.csv, the column of the class is 6 and for dataset2.csv it is 2. This number should NOT be zero-based as I subtract one to account for that in the program itself.

Results

Cluster (index)	Weka	Mine
0	24%	11%
1	40%	40%
2	24%	25%
3	11%	23%

Figure 1 – Clustering Breakdown for dataset1.csv. Four clusters, total number of tuples is 258.

Cluster (index)	Weka	Mine
0	.2995, .392, .528, .670, .223	.626, .431, .504, .791, .661
1	.365, .386, .516, .246, .678	.366, .382, .515, .243, .672,
2	.335, .235, .312, .335, .242	.327, .295, .232, .439, .238
3	.626, .431, .504, .791, .6614	.302, .339, .634, .589, .222

Figure 2 – Final Clusters for dataset1.csv.

My cluster[0] and Weka's cluster[3] are almost identical as well as both cluster[1]'s which is why there are similar amounts of records in both of those clusters.

In terms of performance, k-means is okay for small datasets but very slow for big datasets. Runtime is O(kNM) where N is the number of tuples in the data and M is the number of iterations until centroids converge. In Weka, I tried a couple more Clustering algorithms and most seemed slower than k-means for the small datasets I used, but they also did a better job clustering as they didn't rely on randomly generated centroids. K-means doesn't do well with oddly-shaped clusters or clusters of different sizes so it's likely the clustering results I got could be heavily improved upon with other algorithms, especially density-based algorithms like DBSCAN.