AI for Finance

IE471 / IE 412

Hands-on Practice on Session #2

[Session 2] Within Stock Classification via Clustering

Main TA for this session: Youngbin Lee (e-mail: young@unist.ac.kr)

Report Description

In this project, you can implement the process of stock classification using k-means clustering on your own.

Follow the steps below:

1. **Selecting variables and pre-processing data**
   A. Use variables from 'financial_statements.csv'.
   B. You can run code as provided without modifying and skip this part in report.
2. **Clustering (2 pts)**
   A. Use k-means clustering, with your own number of clusters ($k \neq 6$) and random state.
   (You can choose k regardless of the Silhouette scores. However, for the convenience of later analysis, one of [4,5,7] is recommended. Random state affects initialization of cluster centroids, which may change your clustering results.)
3. **Clustering results (2 pts)**
   A. Compare clusters with scatterplot of 2 or 3 features.
4. **Analysis (4 pts)**
   A. Compare financial properties between clusters with boxplot and mean table.
   B. Compare market behaviors between clusters with correlation table and scatterplot.

Extension (2 pts, optional) :

In extension, you can implement the process of stock classification using different clustering method on your own.

Follow the steps below:

1. **Selecting variables and pre-processing data**
   A. Use variables from 'financial_statements_extension.csv'.
   B. You should be careful about feature scaling before clustering because variables have different distributions.
2. **Clustering (choose one of the following)**
   A. Use one of algorithms from scikit-learn package.[1] (except k-means)
   B. Perform dimensionality reduction and clustering.[2] (PCA & k-means)
3. **Clustering results**
   A. Compare clusters with scatterplot of 2 or 3 features.
4. **Analysis**
   A. Compare financial properties between clusters with boxplot and mean table.
   B. Compare market behaviors between clusters with correlation table and scatterplot.

[1] https://scikit-learn.org/stable/modules/clustering.html

[2] If you choose Option B, the number of variables should be large (e.g., 10 or more variables). With a large number of variables, it would be better to reduce dimension before clustering because Euclidean distances tend to become inflated. So, you could perform PCA to lower dimensions (e.g., 2 or 3) using PCA and then apply k-means.