Clustering
Teacher: Mauro Sozio

1 Clustering

In this assignment, we cluster stocks in the stock market by using the k-means algorithm. In particular, you are provided with a dataset (available on the moodle website) which specifies for each of 30 stocks the percentage change in price of that stock in each given week, for a total of 25 weeks. In our dataset, some stocks might deal with technology, some other with oil, etc. We will try to group together stocks with similar price trends in the stock market. In other words, in a same cluster we would like to have stocks whose price changes by similar amounts every week. This can be used for coming up with successful investment policies. We will see that stocks related to the same market (e.g. technology) have often "similar" price trends. For this assignment, we recommend $k = 8$.

Input File Format. The first line of the file specifies the weeks considered in our dataset, while the rest of the lines specifies the data. In each line, the first element specifies the name of the stock. We use ',' as a separator. For this task, you should consider all continuous-ordinal attributes and ignore the rest of the attributes.

Write your answers in the Jupyter Notebook. Make sure to explain your answers.

Questions.

1. You should run the k-means algorithm on the stock data, while using init='random' and the default values for the other parameters. Compute the sum of squared errors (SSE) for the clustering you obtained and include it in your report.

2. You should then try to decrease the SSE as much as possible (while keeping $k = 8$) by changing some of the parameters accordingly. To this end, select two parameters (numeric or not) that you think should impact the results the most. For each parameter explain : a) how you expect that changing that parameter would affect the results (e.g. if numeric, increasing its value means better or worse results?) b) whether changing the value of the parameter should always improve the results or not necessarily.

3. Then look at the clustering you obtained and try to label each cluster with a topic. For example: cluster of technology stocks, oil stocks, etc. Don't expect your clustering to be perfect. In particular, you might have different kinds of stocks in a given cluster, while you might not be able to label all clusters. We expect that you should be able to label at least three clusters with a topic. It is fine to describe a cluster as a technology cluster if most of the stocks deal with technology, for example. Explain your answers.

What to submit. You should send us your Jupyter notebook with the code in Python, as well as the answers to your questions.