Clustering
Teacher: Mauro Sozio

1 Clustering

In this exercise, 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 behaviour in the stock market. 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" behaviour. For this exercise 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.

You should send us your Jupyter notebook including the answers to the questions and the code in Python. The answer for each question should not exceed 20 lines (the lines of code do not count). The answers for all lab sessions and the project should be sent together until the date specified on the website. You can form groups of 1-2 people that will work together on all labs and the project. The composition of the group cannot be changet throughout the course.

For this question, you should use the tutorial that has been included in the zip file and change it accordingly. You are allowed to use any library for computing the SSE.

Questions. Questions give the same amount of points.

1. You should run the k-means algorithm on the stock data. Compute the sum of squared errors (SSE) for the clustering you obtained, while using the default values of the parameters for k-means and report the SSE. Report the SSE you obtained.

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 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 (increasing its value means better or worse results?) b) whether increasing or decreasing the value of the parameter should always improve the results or not necessarily c) report the SSE you obtained.

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. It is fine to describe a cluster as a technology cluster if most of the stocks deal with technology, for example. Motivate your answers.