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 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 $k = 8$. This lab will not be evaluated.

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.

Questions.

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 selecting the initial centroids randomly and using the default values for the other parameters.

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 try to understand: a) how do 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.

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.

4. implement the k-means++ algorithm, that is, the algorithm that selects the initial centroids in a more clever way than random. To this end, we recommend to use a pseudo-random number generator and to partition the interval $[0, 1]$ into n "smaller" intervals, where n is the number of input points so that 1) each interval correspond to one input point; 2) the probability that a random number falls into one of the smaller intervals is equal to the probability of selecting the corresponding input point in k-means++.