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.

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 using the default values of the parameters for k-means and report the SSE.

2. You should then try to decrease the SSE as much as possible (while keeping $k = 8$) by changing the parameters accordingly. Explain your choice of the parameters and explain why you expect that that choice should give better results.

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.

4. Normalize your data, that is, divide each vector corresponding to a stock by its L2 norm. We recall that the L2 norm of a vector x with n dimensions is equal to: $\sqrt{\sum_{i=1}^{n} x_i^2}$. Run k-means and look again at the clustering you obtained. Did you get better results? That is, could you find more meaningful labels? If your results are better, try to explain why normalizing could have helped. If not, try to explain what could be the benefit of normalizing the data.

5. Use any of the techniques we mentioned during our course (or any other idea you come up with) so as to try to get better results. In particular, you could try to improve the SSE while keeping the same number of clusters or try to obtain more meaningful labels for the clusters (with the same or a slightly different number of clusters). Even if your technique is not successful, it is important to motivate why you expect that such a technique could improve the results.