Exploratory Data Analysis - Written Exam

André Plancha
andre.plancha@hotmail.com
MALMÖ UNIVERTITY
Spring 2025

Computer Science: Applied Data Science

March 29, 2025

1. Due to large sizes of data sets in practice one need to do dimension reduction thereafter study the pattern by various clusterings techniques. Dimension reduction means lost of certain information whence less accuracy of clustering results in general. In this question, we consider the Iris data set as four dimensional data set with the ground truce label: setosa, versicolor, virginica and use the k-means for clustering.

a) First determine three clusters of Iris data set by k-means, and even compute the percentage of correct classified observations.

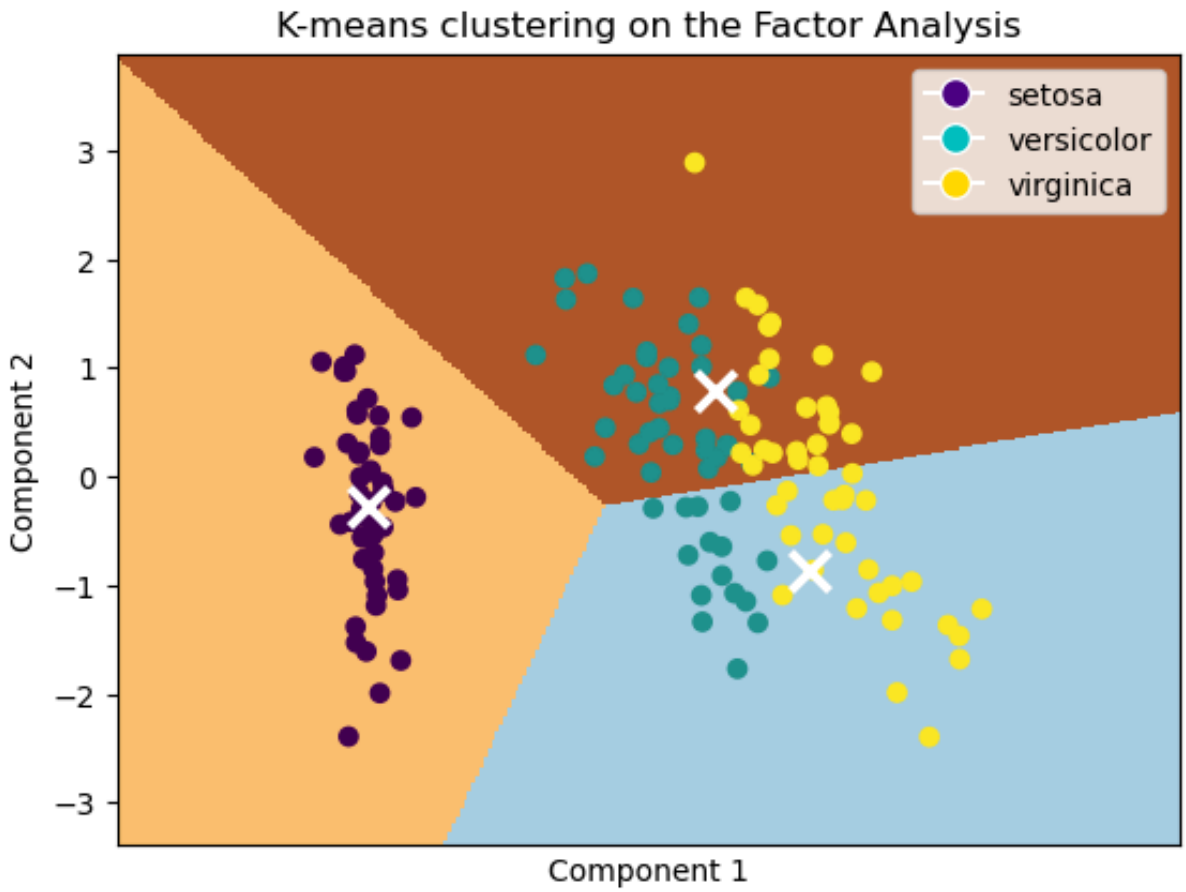
K-means clustering is an iterative algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Using scikit-learn's `KMeans`, the following was obtained:

	Cluster 1	Cluster 2	Cluster 3
setosa	0 (0%)	50 (100%)	0 (0%)
versicolor	47 (77%)	0 (0%)	3 (8%)
virginica	14 (23%)	0 (0%)	36 (92%)
Total	61	50	39

From this, we can see that the clusters obtained are able to separate the classes quite well, with 100% of setosa on one cluster, 77% of versicolor on another, and 92% of virginica on the last one. This leads us to a total of 82% of correct classified observations.

b) Now apply first the factor analysis method to reduce the Iris data set to two dimensional, then determine three clusters of the reduce data set by k-means. What is then the percentage of correct classified observations? Even visualize the reduced data set with original label respectively by the clustering.

Factor analysis uses the correlation structure amongst observed variables to model a smaller number of unobserved, latent variables known as factors. Using scikit-learn's `FactorAnalysis` and scikit-learn's `KMeans`, the following was obtained (the color in the background represents the cluste division):



	Cluster 1	Cluster 2	Cluster 3
setosa	0 (0%)	50 (100%)	0 (0%)
versicolor	15 (38%)	0 (0%)	35 (57%)
virginica	24 (62%)	0 (0%)	26 (43%)
Total	39	50	61

From this, we can clearly see that the algorithm was able to separate setosas from the other classes, but it had a hard time separating versicolors from virginicas. This leads us to a total of $\approx 72.67\%$ of correct classified observations. This is a decrease in the accuracy of the classification, which is expected when reducing the dimensionality of the data.

2. Consider the minimum spanning trees method (MST) and lung cancer data set `LungA`.

a) First apply MST to find two clusters, then find the percentage of correct clustered genes using the original label of cancer cell: `Normal, Cancer (Small-cell lung carcinomas, Nonsmall cell lung carcinomas)` as ground truth.

As I interpreted it, this questions asks us to:

1. Transform `LungA` into a undirected weighted complete graph, where the weight of the edge between two nodes is the euclidean distance between them;

2. Transform that graph into its minimum spanning tree, using for example Kruskal's algorithm;
3. Remove the longest edge from the tree, thus dividing it into two clusters;
4. Compare the clusters obtained with the original labels.

`LungA` is a dataset that seems to have 203 observations and 3312 unnamed numerical features, and each obsevration has a label between `AD`, `SQ`, `COID`, `NL`, and `SCLC`. Grouping `AD`, `SQ`, `COID` and `SCLC` as "Cancer", using the described steps above and using SciPy's `minimum_spanning_tree` and `connected_components`, the following was obtained:

	Cluster 1	Cluster 2
Cancer	185 (92%)	1 (100%)
Normal	17 (8%)	0 (0%)
Total	202	1

From this, we can see that the algorithm was not able to separate the normal cells from the cancer cells, since the smaller cluster has only one observation Cancer observation. Because of the nature of this algorithm, we can conclude that this observation might be an outlier, showing the method's robustness in outlier detection.

b) Apply MST once again to find three clusters, then compare the correct clustered genes using the original label: `Nonsmall cell lung carcinomas, Normal, Small-cell lung carcinomas` as ground truth.

This time, `AD`, `SQ` and `COID` were grouped as "Nonsmall cell lung carcinomas", `SLC` as "Small-cell lung carcinomas", and `NL` as "Normal". Using the same steps as before, the following was obtained:

	Cluster 1	Cluster 2	Cluster 3
Nonsmall cell lung carcinomas	178 (89%)	1 (100%)	1 (100%)
Normal	17 (8%)	0 (0%)	0 (0%)
Small-cell lung carcinomas	6 (3%)	0 (0%)	0 (0%)
Total	201	1	1

The same conclusion can be drawn from this result, where the algorithm was not able to separate the observations based on the labels given, reinforcing the idea that the algorithm is robust to outliers. Aditionally, it's important to note that, since the method is a divise clustering method, cluster 3 is a subset of the previous cluster 1, and increasing the number of clusters will result in this cluster 1 being divided into smaller clusters.

c) Finally, test to find five clusters comparing the correct clustered genes using the label: `AD, COID, SQ, NL, SCLC` as ground truth. Even comment on these clustering results.

This time, `AD`, `COID`, `SQ`, `NL` and `SCLC` were used as labels. Using the same steps as before, the following was obtained:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AD	138 (70%)	1 (100%)	0 (0%)	0 (0%)
COID	16 (8%)	0 (0%)	3 (100%)	1 (100%)
NL	17 (9%)	0 (0%)	0 (0%)	0 (0%)
SMCL	6 (3%)	0 (0%)	0 (0%)	0 (0%)
SQ	21 (11%)	0 (0%)	0 (0%)	0 (0%)
Total	198	1	3	1

As in a) and b), the algorithm was not able to separate the observations based on the labels given. This might also be because there's a big imbalance between the number of `AD` observations, which make up a grand majority of the dataset, making the division difficult to obtain from the data.

3. Model based clustering in chapter 6 (section 6.5) is based on geometrical properties of clusters, e.g. balls and ellipsoids in 3-dim. In this question, you are asked to randomly generate data sets to check the capability of MBC: two balls of different sizes, two ellipsoids of different size with symmetry axis parallel with coordinates axis, three ellipsoids also different sizes with arbitrary symmetry axis.

a) Randomly generate those sub data sets so that they are disjoint (far away from each other), then apply model based clustering to find clusters even compare the estimated clusters with original labels.