Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Course Work #5: *Unsupervised Learning*

Problem 1 (). **Principal Component Analysis**

The first principal component is the direction of maximum variance in the data. Show that this first principal component also minimizes the residual sum of squares, which is here the Euclidean squared distance between the projected data points and the original data points.

Solution.

1. We define w as a unit vector along the first principal component. The distance of the projection of a data point x_i to zero is given by $x_i * w$ (recall that data is centered around 0). The coordinate of the projection is given by $(x_i * w)w$. We are interested in the distance between the data point x_i and this projection, which can be computed with Pythagoras' theorem.

$$||x_i - (x_i * w)w||^2 + ||x_i * w||^2 = ||x_i||^2$$
$$\iff ||x_i - (x_i * w)w||^2 = ||x_i||^2 - ||x_i * w||^2$$

Adding up those squared distances over all data points (depending on the PC):

$$\begin{aligned} RSS(w) &= \sum_{i=1}^{n}(||x_i||^2 - ||x_i * w||^2) \\ &= \sum_{i=1}^{n}||x_i||^2 - \sum_{i=1}^{n}||x_i * w||^2 \end{aligned}$$

We now aim at minimizing this RSS. The first term does not depend on w and we can thus ignore it for minimization. Due to the sign, we end up with maximizing the second term.

$$\text{argmax}_w \sum_{i=1}^{n}||x_i * w||^2 = \text{argmax}_w \frac{1}{n}\sum_{i=1}^{n}||x_i * w||^2$$

Since, $Var(X) = E(X^2) - E(X)^2$

$$\rightarrow \frac{1}{n}\sum_{i=1}^{n}||x_i * w||^2 = (\frac{1}{n}\sum_{i=1}^{n}x_i * w)^2 + Var(x_i * w)$$

Problem 2 (). **Hierarchical Clustering**

Suppose that for a particular data set, we perform hierarchical clustering using single linkage and using complete linkage. We obtain two dendrograms.

1. - At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ fuse. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also fuse at a certain point.Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?
   - At a certain point on the single linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ fuse. On the complete linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

   Explain your reasoning for both the cases.

2. Name and explain one other choice of dissimilarity measure for hierarchical clustering apart from the Euclidean distance metric. Give an example where your stated dissimilarity measure would be a better than the Euclidean distance metric.

3. What are some practical considerations that one needs to take into account when applying clustering on the data? Describe a total of four practical considerations with at least one consideration for hierarchical and at least one for K-Means clustering.

Solution.

1. - **Simple Answer:** The clusters will most likely fuse at a higher point in case of complete linkage since Complete and Single linkage consider the maximum resp. minimum distance between clusters and by definition $Max > Min$,

     Alternate Answer: There is not enough information to answer this. This is because we do not know what distance metric is used. Note that the cluster numbers (i.e. $\{5\}$ and $\{6\}$) are just serial numbers and not the values of the clusters. Unless we know what distance metric is used, we can not compute the distances. And If we can not compute the distances, we can not make a claim about the height they would fuse at.
   - In this case, we know that these clusters would fuse at the same height. We do not know what the height is but we know that these clusters only contain a single observation, therefore the score using single linkage and using complete linkage would be the same even though we do not know the actual scoring metric. Therefore the height at which they fuse will be the same.

2. **Correlation-based distance (but we allow any other reasonable answer)**: Considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance.

   - **For Hierarchal Clustering**
     (a) What dissimilarity measure should be used?
     (b) What type of linkage should be used
     (c) Where should we cut the dendogram in order to obtain clusters?
   - **For K-Means Clustering**: How many clusters should we look for in the data?

Problem 3 ().   K-Means

1. - Show that the following equation holds :

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

   where $|C_k|$ denotes the number of observations in cluster C_k, and $\bar{x}_{kj}$ the mean for feature j in cluster C_k. Argue on the basis of this identity, that the K-means clustering algorithm decreases the objective

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\}$$

   at each iteration.

   - Explain in your own words (a) what equality you are proving and (b) what you can conclude from it.

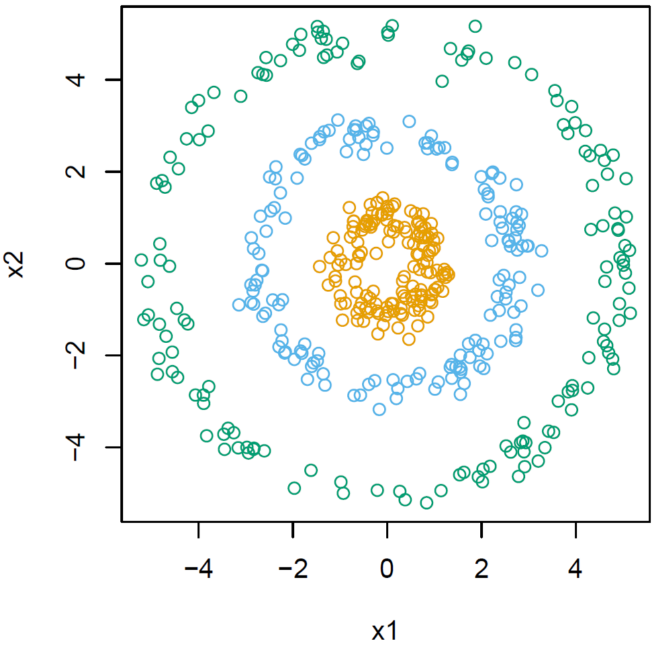

Figure 1: 2 dimensional data for task 3 (from ESL Fig. 14.29).

2. Consider the data plot shown in Figure 1 where each colour denotes one cluster.
   - Can we use k-means clustering to correctly cluster the data points? Why or why not?
   - If you should use hierarchical clustering for this data, which linkage (complete, average, single or centroid) would do best and why?

Solution.

1.

$$\begin{aligned}
\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 &= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}((x_{ij}-\bar{x}_{kj})-(x_{i'j}-\bar{x}_{kj}))^2\\
&= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2-2(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})+(x_{i'j}-\bar{x}_{kj})^2\\
&\text{for each element in } C_k \text{ we have one of the first and the last terms}\\
&= \frac{|C_k|}{|C_k|}\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2\\
&+ \frac{|C_k|}{|C_k|}\sum_{i'\in C_k}\sum_{j=1}^{p}(x_{i'j}-\bar{x}_{kj})^2\\
&- \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})\\
&= 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2-\frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij})(x_{i'j}-\bar{x}_{kj})\\
&= 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2-0
\end{aligned}$$

As mentioned in the lecture, we compute the cluster centroids at each step and specifically minimize the distances from the cluster centroids thus minimizing the RHS of this formula. In turn, the LHS is also minimized.

2. (a) No, k-means cannot be used in this setting because k-means algorithm tends to find spherical clusters in the data.

   (b) Single-linkage would do best because it works on the same principle as k-nearest neighbours. As the data in the given figure has clusters in circles and the distance between the data points belonging to different circles is more than the distance between the data points lying in the same circle, single linkage would be the best to use in this setting. Both, complete and average linkage although being the most popular ones, won't do well here.