# Exercise Sheet #5: Unsupervised Learning

Mhd Jawad Al Rahwanji, Wojciech Ptaś
7038980, 7042843

January 15, 2023

## Problem 3

1. We have two cases:

   - There isn't enough information to tell. We will describe 2 scenarios, one has CL occurring higher, and one has both linkages occurring at the same height.

     I. The data points are on a straight line and the clusters are far apart. In this case,

     $$d_{CL}(\{1,2,3\},\{4,5\}) > d_{SL}(\{1,2,3\},\{4,5\})$$

     II. The data points of the cluster {1,2,3} are on the surface of a sphere and are close together. The cluster {4,5} resembles two data points at the center of said sphere. In this case,

     $$d_{CL}(\{1,2,3\},\{4,5\}) = d_{SL}(\{1,2,3\},\{4,5\})$$

   - We know that for single element clusters,

     $$d_{CL}(\{5\},\{6\}) = d_{SL}(\{5\},\{6\}) = d_{5,6}$$

     Thus, the clusters will fuse at the same height.

2. The cosine distance; It is concerned with measuring the angle between vectors. Imagine 2 vectors that share the same origin and extend in almost the same direction. We will describe 2 scenarios. I) The vectors have equal magnitudes. II) The magnitudes are vastly different. Scenario II is an example of where the cosine similarity would be preferred. It indicates that the vectors are similar when the euclidean distance would indicate otherwise.

   This is particularly useful in information retrieval where we would have 2 word frequency vectors, one of the search query, and one of a document. The objective is to measure their similarity. Euclidean distance would return the difference in length between the documents (few words : few thousand words). Cosine similarity scales with word matches.

3. Some practical considerations for clustering.

   I. As discussed in (3.2), proper hierarchical clustering requires choosing an appropriate dissimilarity measure for the task at hand.

II. As for $k$-means clustering, choosing a suitable $k$ is challenging despite years of domain expertise. We cannot know for sure the number of classes hidden in the data.

III. Whether to standardize or not depends on the dissimilarity measure and feature types & scales. After all, clustering algorithms are distance-based. Therefore, features with large scales pose a concern as they inadvertently drive the cluster estimation.

IV. We could attempt validating the resulting clusters. Albeit challenging, if we have access to some observations we can quantify class-cluster purity. Alternatively, the bootstrap may aid in assessing cluster robustness.