1 Clustering

Clustering is an unsupervised learning techniques that aims to finds clusters or groups in the data such that observations in the same group are similar to each other while observations in different groups are different from each other.

Clustering is useful for an exploratory analysis of the data and also useful in problems like customer segmentation.

1.1 K-means Clustering

A very powerful technique that organizes the data into K distinct groups such that each observation will fall into exactly one group and when all the groups are combined, they cover the entire data set. K is determined beforehand and is the number of clusters we are going to make.

The fundamental idea of clustering is to reduce the within cluster variation. Let C_k denote the set containing the indices of the points falling in the cluster k and $W(C_k)$ be the within cluster variation for cluster k. Then,

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \sum_{k=1}^{K} W(C_k)$$

Using Euclidean distance as a measure of the intercluster distance between two points, we can redefine the optimization summing across all the dimensions of the data as

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \sum_{k=1}^{K} \frac{1}{|C_k|} \left\{ \sum_{i_1,i_2 \in C_k} \sum_{j=1}^{p} (x_{i_1 j} - x_{i_2 j})^2 \right\}$$

where $|C_k|$ is the number of observations in the cluster C_k. Considering all possible partitions is impossible for large n and we use the following algorithm to obtain the local optimum.

1.1.1 Algorithm

We repeat the following until some predefined convergence criteria

1. Random assign a cluster in $1, \ldots, K$ to each observation in the data.
2. Repeat the following till convergence
   (a) Calculate the centroid for each cluster where centroid is a p dimensional vector whose each component is the average of the components of all the points that fall in the considered cluster.
   (b) Assign each observation the cluster index of the centroid that is closest to the given observation (using Euclidean distance).

It is usually a good idea to center and standardize the variables first so that the individual magnitudes and variances don't affect the Euclidean distances drastically.
To see why using distance from centroid is a good replacement for the pairwise distance, consider the following equation

$$\sum_{i_1,i_2 \in C_k} (x_{i_1} - x_{i_2})^T (x_{i_1} - x_{i_2}) = \frac{1}{2} \sum_{i \in C_k} \sum_{j \in C_k} (x_i - x_j)^T (x_i - x_j)$$

where the right side allows for all possible pairs including the ones where the indices might repeat. Continuing to expand the right hand side,

$$\begin{aligned}
\sum_{i\in C_k}\sum_{j\in C_k}(x_i-x_j)^T(x_i-x_j) &= \sum_{i\in C_k}\left(|C_k|(x_i^Tx_i)-2x_i^T\sum_{j\in C_k}x_j+\sum_{j\in C_k}x_j^Tx_j\right)\\
&= \sum_{i\in C_k}\left(|C_k|(x_i^Tx_i)-2|C_k|x_i^T\bar{x}+\sum_{j\in C_k}x_j^Tx_j\right)\\
&= \sum_{i\in C_k}\left(|C_k|(x_i^Tx_i)-2|C_k|x_i^T\bar{x}\right)+\sum_{i\in C_k}|C_k|x_i^Tx_i\\
&= \sum_{i\in C_k}\left(2|C_k|(x_i^Tx_i)-2|C_k|x_i^T\bar{x}\right)\\
&= 2|C_k|\left\{\left(\sum_{i\in C_k}(x_i^Tx_i)\right)-|C_k|\bar{x}^T\bar{x}\right\}\\
&= 2|C_k|\left\{\left(\sum_{i\in C_k}(x_i^Tx_i)\right)-2|C_k|\bar{x}^T\bar{x}+|C_k|\bar{x}^T\bar{x}\right\}\\
&= 2|C_k|\left\{\left(\sum_{i\in C_k}(x_i^Tx_i)\right)-2(\sum_{i\in C_k}x_i^T\bar{x})+|C_k|\bar{x}^T\bar{x}\right\}\\
&= 2|C_k|\left\{\sum_{i\in C_k}x_i^Tx_i-2x_i^T\bar{x}+\bar{x}^T\bar{x}\right\}\\
&= 2|C_k|\left\{\sum_{i\in C_k}(x_i-\bar{x})^T(x_i-\bar{x})\right\}
\end{aligned}$$

$$\text{Thus, } \frac{1}{|C_k|}\left\{\sum_{i_1,i_2\in C_k}\sum_{j=1}^{p}(x_{i_1j}-x_{i_2j})^2\right\}=\left\{\sum_{i\in C_k}(x_i-\bar{x})^T(x_i-\bar{x})\right\}$$

Thus, the quantity we set out to minimize for each cluster is indeed the sum of distance of each point from the centroid of the cluster, which means the cluster is to be chosen based on the closest centroid to minimize the overall intra cluster distance.

The optimum found by K-means clustering is local which makes it important to run the algorithm with different random initializations to get the minima.

Elbow Curve is a plot between the total intra cluster distance vs the number of cluster K and is a visual method to obtain the optimum number of clusters to use. The elbow shape refers to the fact that if there is indeed clusters present in the data, the plot will see a sharp decline in the intra cluster distance for some k.
Note that the curve is going to always keep decreasing as in the limiting case when we have n points and n clusters, the total distance will be zero. Hence, the number of clusters must be carefully chosen.

1.2 Hierarchical Clustering

K-means suffers from the disadvantage that the number of clusters needs to be specified beforehand. Hierarchical does not require such a consideration beforehand. here we dicsuss the **bottom-up** or **agglomerative clustering** approach. Hierarchical clustering is visualized using a **dendogram** which is a tree like diagram draw upside down. Starting from the bottom, branches are originate from the individual data points and slowly start merging as we move

upward. The earlier the branches merge, the similar the data points are and vice versa. (Be careful to not judge the similarity from the proximity on the horizontal axis)

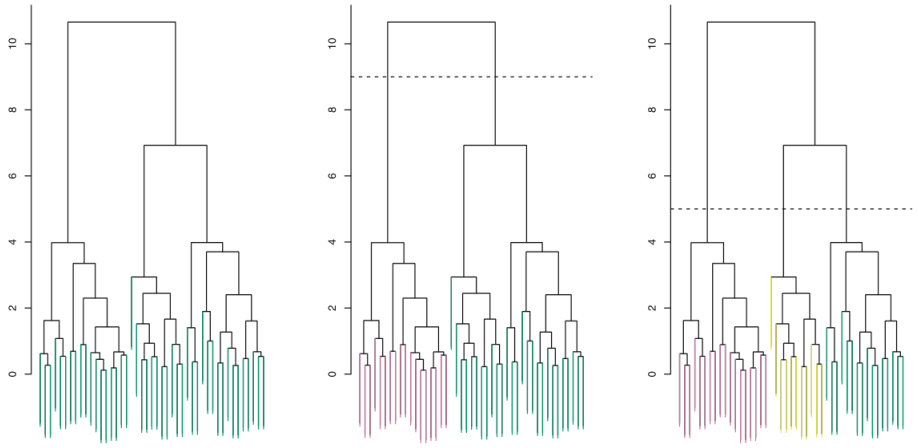

Figure 1: Visualization of dendogram. The two curves on the right colour the different clusters obtained based on the height at which we decide to cut.

The number of clusters is simply determined by the height at which we made the cut. The middle figure in figure 1 shows a cut at height 9 which results in two branches and thus two clusters.
Changing the height from the highest value to the lowest value will result in 1 and n clusters respectively. Thus, we do not need to specify the number of clusters beforehand but it is rather to be chosen by us based on the height. This height can be seen similar to K in K-means clustering.

The inherent nesting of the clusters as indicated visually by the dendogram may not be possible in every data set and there will be cases when K-means clustering may be superior.

1.2.1 Algorithm

Hierarchical Clustering is performed in a bottom-up approach. Start with n clusters where each observation is it's own cluster. Define a dissimilarity measure between each pair of observation. This can be Euclidean distance as well. Now, cluster the observations that are least dissimilar into the same group, which will give us $n - 1$ clusters. Again use the dissimilarity measure to group two similar observations until the total number of clusters is 1.

Consequently, there will be cases when we need to determine the dissimilarity between a group and an observation or a pair of groups. This is done using **linkage**. Four types of linkages used are **complete, average, single** and **centroid**. Average and complete linkages are preferred over single linkages, and all three are more popular than centroid linkage. Average and complete likages will usually give balanced dendograms.

Following are teh descriptions of individual types of linkages

- Complete
  Maximal intercluster dissimilarity. Take the maximum of the pairwise dissimilarity between observations of cluster A and cluster B.
- Single
  Minimum intercluster dissimilarity. Take the minimum of the pairwise dissimilarity between observations of cluster A and cluster B.

Single linkage can result in extended trailing clusters in which single observations fuse one at a time.

- Average
  Mean intercluster dissimilarity. Take the average of the pairwise dissimilarity between observations of cluster A and B.

- Centroid
  Take the dissimilarity between the centroid of cluster A and B. Centroid linkages can result in undesirable inversions.

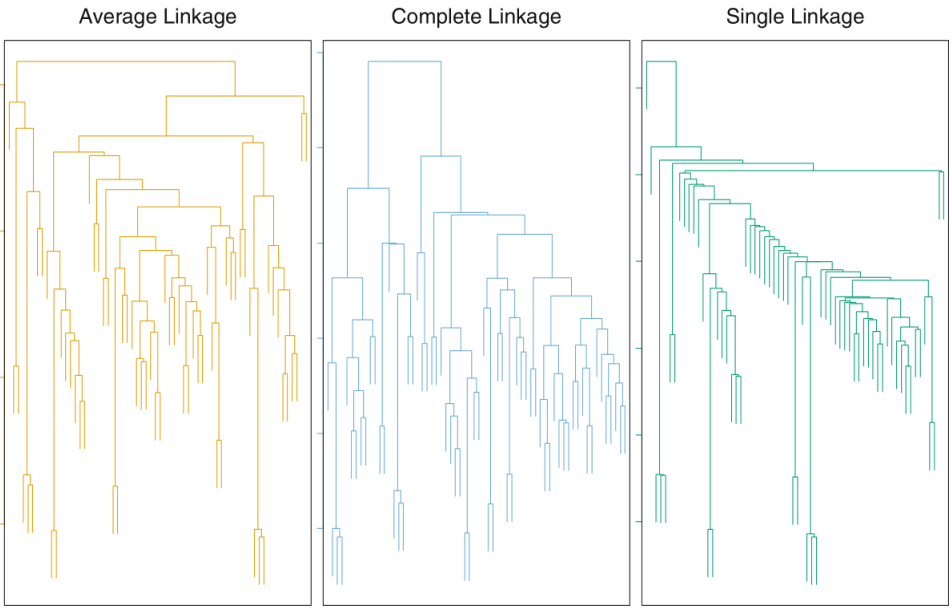

Figure 2: Visualization of dendograms obtained for different choice of linkages.

The algorithm can be summarized as follows

1. Start with n clusters where each observation is it's own cluster. Compute $n(n-1)/2$ pairwise dissimilarity measures between all pairs.

2. For $i = n, n-1, \ldots, 2$

   (a) Compute the pairwise dissimilarity between all i clusters and take the two clusters with the least dissimilarity (or highest similarity). Fuse them into a single cluster. The dissimilarity measure is also indicative of the height in the dendogram where the two clusters fuse.

   (b) Recompute the pairwise dissimilarity between the $n-1$ clusters.

1.2.2 Dissimilarity Measures

So far, euclidean distance has been considered as the defacto dissimilarity measure. In some cases, this may not work well if the magnitude of the observations vary significantly between the different predictors. In such cases, correlation based measures can be preferred since they will group observations with similar behaviour together and not focus on magnitude.
This can be useful in retail behaviour when we want to check profiles based on the whether similar products are purchased rather than how many of them are purchased.

1.2.3 Key Considerations

Following are a set of general rules when doing clustering

- centering and bringing the variables to the same scale is useful when measuring Euclidean distance for the obvious reason of not letting magnitudes affect the distances.
- Different types of clustering approaches should be explored to check which performs the best. This is important as in unsupervised learning, the structure of data is not known beforehand and it is important to explore multiple hypothesis.
- Several choices of similarity measures and linkages can be explored for further understanding of data.
- Clustering can be non robust and thus the results should be "validated" by performing clustering on multiple subsets of data to assure stability.
- In come cases, the hard cluster assignment of K-means and hierarchical clustering may not be useful. Probabilistic models like mixture models can be explored in this case.