FAQ's - Hierarchical Clustering and PCA

1. Is scaling necessary before doing any Unsupervised Learning model?

Yes, not only Unsupervised Learning models but also some Supervised and Ensemble models. Because scaling brings all the features under one single scale. This way model can give equal weightage to all features and data points. K Means uses Euclidean distance measure so scaling is crucial. PCA tries to get the features with maximum variance, and the variance is high for high magnitude features and skews the PCA towards high magnitude features, so scaling is a necessary pre processing step.

2. How do we know whether K-means or hierarchical clustering is appropriate to use in a given business problem?

The k-means algorithm is used when it is already known in advance how many clusters have to be formed, also k-means is suitable if your data is well separated into spherical-like clusters. On the other hand, hierarchical clustering is density-based clustering in which nearby points are joined to form clusters. It gives you a dendrogram from which you can figure out how many clusters should be formed. Hierarchical clustering is computationally expensive so it will not perform well when the data size is very big.

3. PCA explained how to reduce dimensions on the data set, However, when solved, there is no info on which dimensions (columns) are the most important and which we can remove. How can I know that information?

Using PCA, the columns are completely changed. PCA finds a new relationship between the columns. So, we use it to get a higher score. If the focus is to find feature importance, then PCA fails as it completely transforms the columns and new columns will not mean the same as the older ones, so we cannot name them. Feature importance after PCA will make no sense at all.

After transforming columns using PCA we try to take the minimum number of columns through which we can achieve the maximum score.

4. How do we choose the correct distance to use in clustering algorithms?

There is no single distance that will give the best results with all data and all problem statements. The type of distance you use depends on the data and the problem statement. Generally normalized Euclidean distance is used. However, when data size is very large and high dimensional, Manhattan distance is found to perform better computationally. The type of distance to use is decided by the problem at hand.

5. What then do we do with the clusters after interpretations?

What you do with clusters after interpretation depends upon the problem you are trying to solve. Clustering will give you groups that are similar in some aspects. This can be used for recommendations, understanding customers, market campaigning, etc. Again, what you do after interpretation of clustering depends upon what problem you are trying to solve.

6. Are correlated features a problem when it comes to clustering? If so can you point me to why this might be a problem?

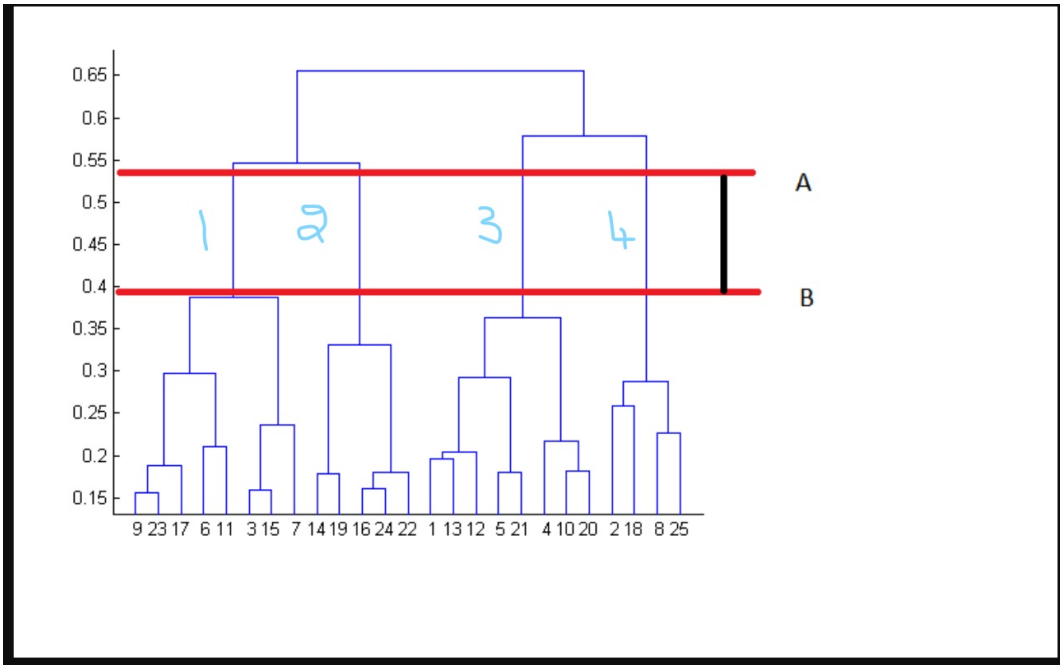
First of all, you have to decide which variables you should be used for clustering, once you have done that, it is better to choose only those variables among them, such that no two variables have a correlation of more than 0.7 (magnitude).

Correlation does not have a negative impact on clustering but removing correlated variables helps to reduce the dimension and can be computationally efficient when you have a very large number of observations.

7. How do we select the optimal number of clusters from the Dendrogram?

Choosing the optimal number of clusters is a fairly subjective matter, and the best method to identify the optimum number of clusters is to use a combination of metrics and domain expertise. The dendrogram is one of the most common ways for estimating the appropriate number of clusters for Hierarchical clustering if we don't have domain expertise.

The optimal number of clusters from a dendrogram can be obtained by deciding where to cut the cluster tree. Generally, the cluster tree is cut where the dendrogram height is maximum as it corresponds to distinct and homogeneous clusters.



Please refer to the above plot for an example, the distance between the red horizontal lines denotes the maximum distance that has been traversed without intersecting any cluster. So, the ideal number of clusters, from the above dendrogram can be 4 clusters. However, one should also do cluster profiling and check if the cluster profiles are meaningful and have variability, for which domain knowledge is needed.

We can get a different number of clusters from the dendrogram at different heights. But the standard approach to picking the appropriate number of clusters from the dendrogram is to check the maximum height of the vertical line formed.

8. How does interpretability of data becomes difficult after performing PCA?

Principal components are linear combinations of the features from the original data, as principal components become orthogonal to each other but they are not as easy to interpret. For example, it is difficult to tell which are the most important features in the dataset after computing principal components.

Although dimensionality reduction is useful, it comes at a cost. Information loss is a necessary part of PCA. Balancing the trade-off between dimensionality reduction and information loss is unfortunately a necessary compromise that we have to make when using PCA.

Necessary Libraries

from sklearn.decomposition import PCA

********** HAPPY LEARNING **********