← Go Back to Unsupervised Learning

Course Content

FAQ - Hierarchical Clustering and PCA

1. 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.

2. 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.

3. 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.

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

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 data size is very very big.

5. I get the following error when I try to use SilhouetteVisualizer with AgglomerativeClustering.

```
AttributeError: 'AgglomerativeClustering' object has no attribute 'predict'
```

Why does this throw the error and how to debug this?

The SilhouetteVisualizer function from yellowbrick library is designed for visualizing the cluster of K-means only as you can read from its documentation. If you want to find silhouette score for clusters obtained using hierarchical clustering, then you have to use sklearn function of silhouette score which is given here.

A sample code is given below

```
score = silhouette_score(X, HCmodel.labels_, metric='euclidean')
```

where X is a scaled dataset, and HCmodel is the agglomerative model fit on the dataset.

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 to choose the optimal number of clusters from the dendrogram in Hierarchical Clustering?

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 dendrogram height is maximum as it corresponds to distinct and homogeneous 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.

If the chosen number of clusters does not seem meaningful from the cluster profiles or does not align with domain knowledge, then one should choose different values of k (number of clusters) and repeat the process until the cluster profiles obtained are meaningful and align with the domain knowledge.