Clustering

→ finding group in data

→ unsupervised learning

↳ There is no target variable

1 Calinski-Harabasz Score

The Calinski-Harabasz score also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters, the higher the score, the better the performances. The CH score for K number of cluster on a dataset is given by:

$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$

$$CH = \frac{\sum_{k=1}^{K} n_k ||\mu_k - \mu||^2}{K-1} \times \frac{N-K}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} ||x_i - \mu_k||^2} \tag{1}$$

where n_k and μ_k are the number of data points and centroid of the k-th cluster, μ is the global centroid, and N is the total number of data points.

2 Davies-Bouldin Score

This score signifies the average "similarity" between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves. A lower Davies-Bouldin score relates to a model with better separation between the clusters. The score is defined as the average similarity between each cluster C_i for $i = 1, 2, \ldots, k$ and its most similar one C_j. In the context of this score, similarity is defined as a measure R_{ij} that trades off:

- s_i, the average distance between each point of cluster i and the centroid of that cluster (also known as cluster diamter).
- d_{ij}, the distance between cluster centroids i and j.

A simple choice to construct R_{ij} so that is non-negative and symmetric is given by:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2}$$

Then the Davies-Bouldin score is defined as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{\boxed{i \neq j}} R_{ij} \tag{3}$$

3 Silhouette Score

The silhouette score is a formal measure of how well a clustering fits the data. The higher the silhouette score, the better. Typically, the score is calculated for each data point separately, and the average is taken as a measure of how well the model fits the dataset altogether.

There are two main components to the silhouette score. The first component measures how well the data point fits into the cluster that it is assigned to. This is defined as the average distance between it and all other data points in the same cluster. The second component measures how well the data point fits into the next nearest cluster. It is calculated in the same way by measuring the average distance between the data point and all of the data points assigned to the next nearest cluster. The difference between these two numbers can be considered as a measure of how well the data point fits into the cluster it is assigned to as opposed to a different cluster. Therefore, when calculated for all data points, it is a measure of how good each data point fits into the particular cluster it has been assigned to.

More formally, given a data point x_i, where a_{x_i} is the average distance between that data point and all other data points in the same cluster, and b_{x_i} is the average distance between the data point x_i and the data points in the next nearest cluster, the silhouette score is defined as follows:

$$s(x_i) = \frac{b_{x_i} - a_{x_i}}{\max(a_{x_i}, b_{x_i})} \tag{4}$$

Notice that since we divide by the maximum of a_{x_i} and b_{x_i}, we end up with a number between -1 and 1. A negative silhouette score means that this data point is actually on average closer to other cluster, whereas a high positive silhouette score means it is a much better fit to the cluster it is assigned to. When we take the average silhouette score across all data points, we will therefore still get a number between -1 and 1, where the closer we are to one the better the fit. Notice that the silhouette score is a general measure of how well a clustering fits the data, so it can be used to determine the number of cluster and to compare different models.