Davies-Bouldin Index

<https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index>

Obtain D\_ij by finding the ratio of d-bar\_i + d-bar\_j to d\_ij where d-bar\_i is the average distance between every data point in cluster i to the centroid of i, d-bar\_j is the average distance between every data point in cluster j to the centroid of j, and d\_ij is the Euclidean distance between the centroids of i and j. For a given cluster, we find the maximum D between it and another cluster. We do this for every cluster, and average the scores to find the Davies-Bouldin index for the entire scheme. The smaller the index, the better the clustering.

Silhouette Score

<https://en.wikipedia.org/wiki/Silhouette_(clustering)>

Apply two functions a and b to a data point i. The output of the function a characterizes the average distance of i to other points in its cluster. The output of the function b characterizes the average distance of i to all other points in the next closest cluster. Apply a third function s to i using the outputs of a and b to obtain a score between -1 and 1 representing how fit i is for its assigned cluster. Average the silhouette scores of every data point in the set to obtain a global score.

Elbow method

<https://en.wikipedia.org/wiki/Elbow_method_(clustering)>

Try x number of cluster, evaluate the percentage of variance explained under this scheme, try x + 1 ... x + n clusters, evaluate the percentage of variance explained under those schemes. Plot percentage of variance explained as a function of x. Find the value of x at which the graph curves most dramatically. Choose this value of x to be the optimal number of clusters for the data.

Silhouette, DB Comparison

<https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering/#:~:text=For%20example%2C%20the%20Davies%2DBouldin,centroid%20belonging%20to%20another%20cluster>.

The silhouette score is more computationally demanding, but provides more information about the a clustering scheme by assessing the quality of each data point’s assignment individually. The Davies-Bouldin index zooms out and assesses intra-cluster similarity and inter-cluster difference.

Overfitting

<https://www.youtube.com/watch?v=T9NtOa-IITo>

Linear discriminant analysis is a process similar to PCA, but instead of maximizing variance it maximizes separation between categories. Overfitting is a problem wherein LDA separates data with too much fidelity relative to the training set, and consequently does a poor job classifying new data points.