Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

- **Unsupervised learning** task;
- Contrary to classification, we don't feed the algorithm with labeled data, instead we want to **find patterns** and **group** the data;

There are many clustering methods:

- **Hierarchical** clustering - agglomerative;
- **Partition-based** clustering - k-means** and k-medoids;
  - All clusters have the same size;
- **Model-based** clustering - EM (Expectation-Maximization);
  - Clusters can have different sizes, usually one big and one small;
- **Density-based** clustering - DBSCAN
  - **DBSCAN - Density-Based Spatial Clustering of Applications with Noise** - forms clusters based on the density of the data points - density is the same for all clusters;

Assessing Clustering Quality

- The goal is to try to balance the individual clusters' **cohesion** and **separation**;
- **Cohesion** - how similar are the objects within a cluster; also known as **intra-cluster similarity**;
  - Estimation of the **diameter** or **radius** of the cluster - calculate distance between the **centroid** and the **farthest** point in the cluster;

- The most common measure of cohesion if the quadratic error, usually called **Mean Squared Error** (MSE) - the mean of the squared distances between each point and the centroid: $MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$;
- The **Mean Absolute Error** (MAE) is also used - the mean of the absolute distances between each point and the centroid: $MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$;
  * Less sensitive to outliers - should be used when outliers are abundant;

- **Separation** - how different are the objects in different clusters; clusters should be **far apart**, without overlapping;
  - The most common measure of separation is the **distance between the centroids** of the clusters;
  - **Single-linkage** - the distance between the **closest** pair of points in different clusters;
  - **Ward's distance** - the distance between the **farthest** pair of points in different clusters;

  We are looking for **compact and well-separated clusters**.

- The **Dunn Index (DI)** is a measure of the **compactness** of the clusters - the ration between the minimum separation between clusters and cohesion of the largest cluster: $Dunn = \frac{\min_{i \neq j} d(C_i, C_j)}{\max_i d(C_i)}$;
  - **The higher the value, the better the clustering**;
- The **Davies-Bouldin index** is a similar measure, measuring how compact the clusters are compared to the separation between them: $DB = \frac{1}{k}\sum_{i=1}^{k}\max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(C_i, C_j)}\right)$;
  - **The lower the value, the better the clustering**;
- **SSE - Sum of Squared Errors** - the sum of the squared distances between each point and the centroid: $SSE = \sum_{i=1}^{n}(x_i - \mu)^2$;
  - **The lower the value, the better the clustering**;
- The **Silhouette coefficient** is a measure of how well each object lies within its cluster - the mean distance between a sample and all other points in the same cluster, divided by the mean distance between a sample and all other points in the next nearest cluster: $s = \frac{b-a}{\max(a,b)}$;
  - **Better** if s is **closer to 1**;
  - **The higher the value, the better the clustering**;
  - $-1 \leq s \leq 1$;

- $s = 0$ - the sample is **very close** to the neighboring clusters;
- $s = 1$ - the sample is **far away** from the neighboring clusters;
- $s = -1$ - the sample is assigned to the **wrong** clusters.