CLUSTERING
– Theory and Concepts

Sources:

- Course book Unit 2
- Igual & Segui (2017) Unit 7
  + Jupyter Notebook
- Runkler (2012) Unit 9

1. Distance Measures

Intro: Clustering is the problem of grouping points by similarity. We therefore have to develop a concept of similarity. How can we construct the distance in our n (= examples) $\times$ m (= features) data matrix?

L_k distance:

$$d_k(x_A, x_B) = \sqrt[k]{\sum_{i=1}^{m} |x_{A,i} - x_{B,i}|^k} = \left[\sum_{i=1}^{m} |x_{A,i} - x_{B,i}|^k\right]^{1/k}$$

The value of k can be between 1 and ∞.

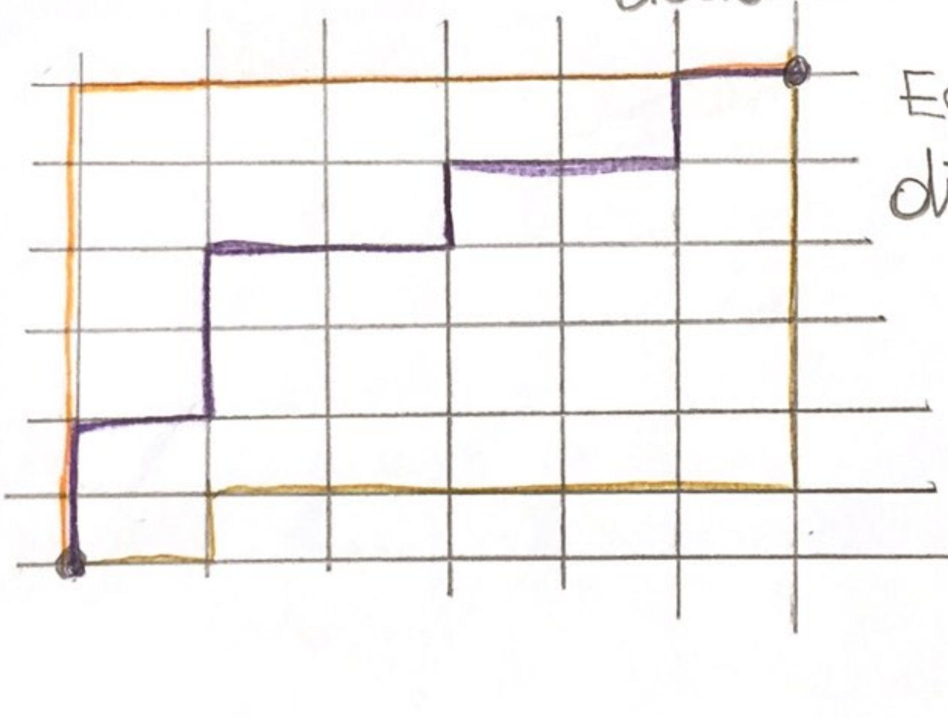
$k=1$ Manhattan distance

$$d_1(x_A, x_B) = \sum_{i=1}^{m} |x_{A,i} - x_{B,i}|$$

$k=2$ Euclidean distance

$$d_2(x_A, x_B) = \sqrt{\sum_{i=1}^{m} (x_{A,i} - x_{B,i})^2}$$

more weight to larger deviations



Equal L1 distance

Example: Distance of points $p_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $p_2 = \begin{bmatrix} 2 \\ 1.99 \end{bmatrix}$ from origin.

$k = 1$ $\quad |2-0| + |0-0| = 2$

$\quad |2-0| + |1.99-0| = 3.99$

$k = 2$ $\quad \sqrt{(2-0)^2 + (0-0)^2} = 2$

$\quad \sqrt{(2-0)^2 + (1.99-0)^2} = 2.82136$

$k = 1000$

$\quad \sqrt[1000]{(2-0)^{1000} + (0-0)^{1000}} = 2$

$\quad \sqrt[1000]{(2-0)^{1000} + (1.99-0)^{1000}} = 2.00001$

$k = \infty$ $\quad$ Component $|x_{A,i} - x_{B,i}|$ with highest value, so for p_1 it is 2 and for p_2 also 2.

A distance measure is a metric if it satisfies.

(i) Positivity $d(x_A, x_B) \geqslant 0$

(ii) Identity $d(x_A, x_B) = 0$ if $x = y$

(iii) Symmetry $d(x_A, x_B) = d(x_B, x_A)$

(iv) Triangle Identity $d(x,y) \leqslant d(x,z) + d(z,y)$

For instance,

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

is not a distance measure.

$$d(x,y) = 1 - \frac{\arccos(\cos(x,y))}{\pi}$$

is a distance measure.

2. Metrics to Measure Clustering Quality

Intro: How do we measure the quality of the clustering result? Here we look at two approaches

(i) Rand index

(ii) Silhouette coefficient

The Rand index is defined as follows

$$R = \frac{a+b}{a+b+c+d} \in [0,1]$$

$S = \{o_1, \dots, o_n\}$

$X = \{X_1, \dots, X_r\}$ partition into r subsets

$Y = \{Y_1, \dots, Y_s\}$ partition into s subsets

Here

- a is # of pairs of elements in S that are in the same subset of X and Y
- b is # of pairs of elements in S that are in different subset of X and Y
- c is # of pairs of elements in S that are in same subsets in X but in different subsets in Y
- d is # of pairs of elements in S that are in different subsets in X but same subsets in Y

Adjusted Rand index ensures that index is close to 0 for random labeling and 1 when clusterings are identical. ⟶ sklearn.metrics

V-measure is another such performance metric

Drawback : These metrics require knowledge of groundtruth classes, while in practice this information is almost never available

Silhouette Score :

$$\text{Silhouette}(i) = \frac{b - a}{\max(a,b)} \in [-1,1]$$

- a is mean distance to the other instances in the same cluster
- b is the mean distance to the next closest cluster

Here

+1 means that instance well inside its own cluster and far from other clusters

0 means that instance is close to boundary

-1 means that instance may be in wrong cluster

⟶ sklearn.metrics

On sklearn under 2.3 "Clustering" in 2.3.10 "Clustering performance evaluation" you find more information on possible metrics.

3. Clustering Techniques

Intro: We distinguish between <u>soft partition</u> algorithms (= assigning a probability to datapoint belonging to a cluster) and <u>hard partition</u> algorithms.

Two families of clustering techniques:

(i) Partitional algorithms

random partition + refine it iteratively

(ii) Hierachical algorithms

hierachical structure: bottom-up or top-down

A typical hard partition algorithm is K-means clustering.

K-Means Clustering

n samples in k disjoint clusters

$c_i = 1, \dots, k$

with μ_i the mean of the clusters

= cluster centroids

$$\arg\min \left[\sum_{j=1}^{k} \sum_{x \in C_j} \overbrace{\underbrace{d(x, \mu_j)}_{\|x - \mu_j\|_2^2}}^{\text{inertia = within-cluster sum of squares}} \right]$$

⟶ sklearn.cluster.KMeans

input: n_clusters

K-Means clustering is an example of an <u>expectation maximization (EM)</u> algorithm.

Hierachical Clustering

The hierachy of clusters is represented as a tree. The tree is usually called a dendrogram.

Top-down : all data in single cluster, divide the cluster

Bottom-up : each data point in single cluster; join pair of clusters

Linkage criterion determines the metric for cluster merging:

- Maximum or complete linkage minimizes the maximum distance between observations of pairs of clusters
- Average linkage
- Ward linkage (=minimizes sum of squared differences within clusters)

sklearn.cluster.AgglomerativeClustering

parameters:

- linkage = 'average'
- n_clusters
- connectivity: defines which are the neighboring samples in the dataset → imposed via a <u>connectivity matrix</u> (only has elements at intersection of row and column with indices that should be connected)