Some metrics to evaluate the quality of clustering

Some metrics presented here have been seen in class, sometimes with a different expression. Do not worry, both expression are then equivalent, and you can use the one you prefer ;)

1 Inertia of the clustering

This metric is also called the "with-cluster sum of squares criterion". Suppose a clustering with m centroids $\mathbf{c} = \{c_j\}_{j=1}^m$ over points $\{x_i\}_{i=1}^n$. The inertia is computed as follows:

$$\sum_{i=1}^{n} \underbrace{\min_{c_j \in \mathbf{c}}(||x_i - c_j||^2)}_{\text{dist. of } x_i \text{ to its centroid}} \quad .$$

The smaller is this value, the more "dense" are the clusters, and the highest is the probability that the clustering has created meaningful clusters. This metric is also the one used by K-Means ++ during its initialization.

2 The rand index adjusted for chance (ARI)

The ARI is based on the rank index (RI), which is computed in the following way. Consider all the pairs of points: $\{(x_i, x_j)\}_{i \neq j}$ and compute

- a = number of pairs classified in the same cluster for both the true clustering and the predicted one.
- b = number of pairs classified in different clusters for both the true clustering and the predicted one.
- c = number of pairs classified in the same cluster in true clustering and in different clusters in the predicted one.
- d = number of pairs classified in different clusters in true clustering and in the same cluster in the predicted one.

the RI is then

$$\text{RI} = \frac{a+b}{a+b+c+d}.$$

Intuitively, the rank index computes the proportion of pairs properly grouped (in a) or discriminated (in b).

The ARI is an adjustment of the rank index, such that the metric lies in $[-1, 1]$. Random clusters have an ARI value close to 0, while perfect clusters (up to permutations) have ARI of 1:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[RI]}{\max(\text{RI}) - \mathbb{E}(\text{RI})}.$$

Note that the true clustering should be known for this metric, it is thus a metric for supervised clustering!

Consider this small example to perfectly understand the metric:

	x_1	x_2	x_3	x_4
True clusters	0	0	1	0
Predicted clusters	0	0	0	1

- a = 1 (x_1, x_2)
- b = 1 (x_3, x_4)
- RI = 2/6 = 1/3
- $\mathbb{E}$[RI] = 1/2
- max(RI) = 1
- ARI = -1/3

In this example, the predicted clustering has worse RI than expected for random clustering! It seems to be a very bad clustering choice!

3 Silhouette score

Let $x_i \in C_j$, the silhouette score of this point is based on the two following measures:

a = the mean intra-cluster distance: $\frac{1}{|C_j|} \sum_{x \in C_j} \text{dist}(x, x_i)$;

b = the mean nearest-cluster distance: $\frac{1}{|C_k|} \sum_{x \in C_k} \text{dist}(x, x_i)$, where $C_k \neq C_j$ is the second nearest cluster to x_i.

The silhouette score of x_i is then

$$S(x_i) := \frac{b - a}{\max(a, b)},$$

and the general silhouette score is the mean of the scores of each x_i,

$$S := \sum_{i=1}^{n} \frac{S(x_i)}{n}.$$

This metrics lies in $[-1, 1]$ and the closer this value is to 1, the more separated are the clusters.

4 Sources

1. The sklearn API for metrics: `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics`, consulted Thu. 2020-10-15.

2. The sklearn description of clustering: `https://scikit-learn.org/stable/modules/clustering.html`, consulted Thu. 2020-10-15.