Clustering

Clustering is a member of unsupervised learning. It could be divided into three categories roughly: prototype-based clustering, density-based clustering and hierarchal clustering.

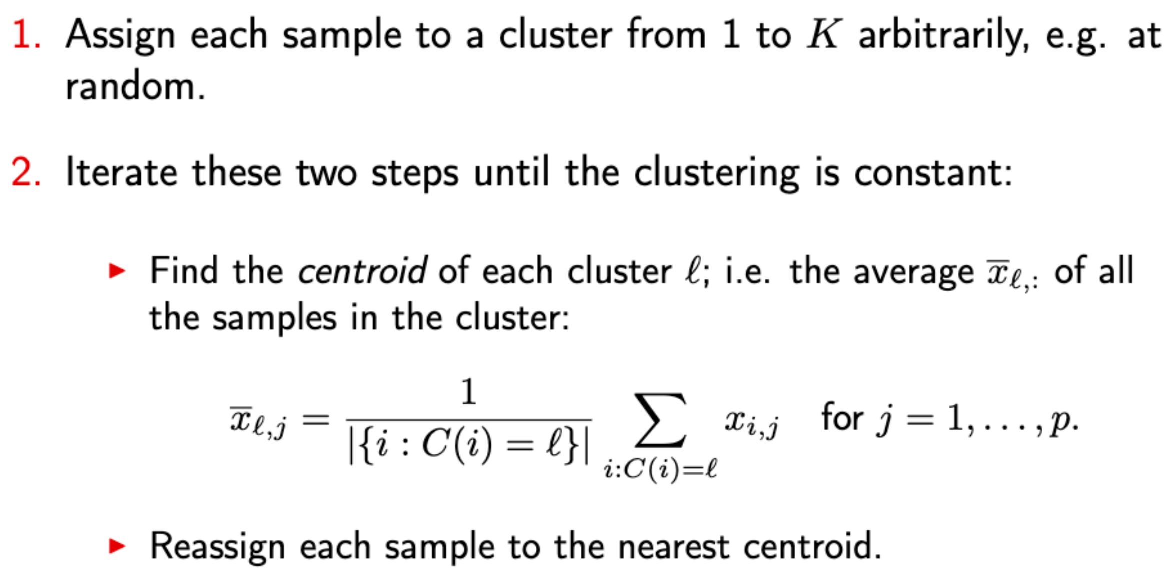
- Prototype-based clustering: we expect that clustering structure could be described by a set of prototypes. (K-means, K-medoids, Learning vector quantization, Gaussian mixture model)
- Density-based clustering: we expect that clustering structure could be described by the density of samples. (DBSCAN)
- Hierarchal clustering: tree clustering structure.

K-Means

1. Algorithm

The goal of this method is to maximize the similarity of samples within each cluster:

$$W(C) = \frac{1}{2}\Sigma_{k=1}^{K}\Sigma_{C(i)=k}\Sigma_{C(j)=k}d(x_i, x_j)$$



1. Assign each sample to a cluster from 1 to K arbitrarily, e.g. at random.
2. Iterate these two steps until the clustering is constant:
   - Find the *centroid* of each cluster ℓ; i.e. the average $\overline{x}_{\ell,:}$ of all the samples in the cluster:

     $$\overline{x}_{\ell,j} = \frac{1}{|\{i : C(i) = \ell\}|} \sum_{i:C(i)=\ell} x_{i,j} \quad \text{for } j = 1, \ldots, p.$$

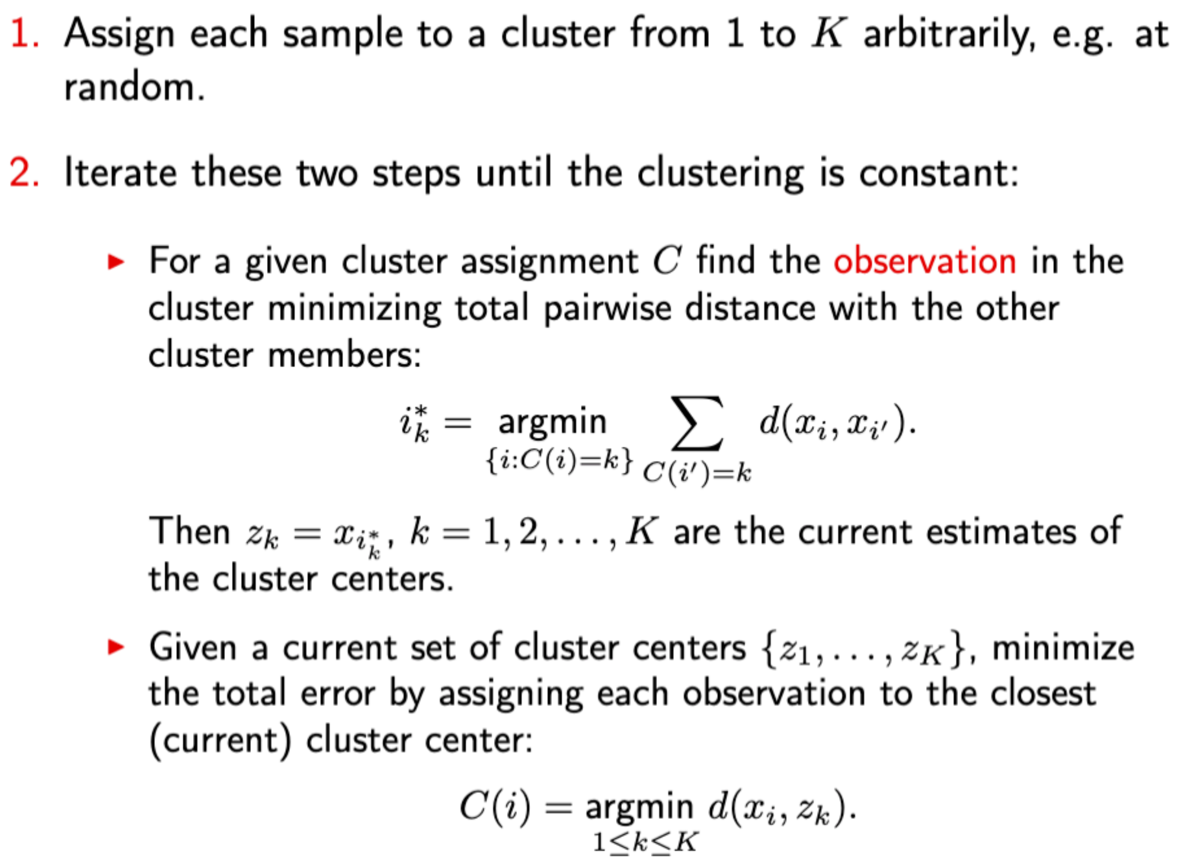
   - Reassign each sample to the nearest centroid.

2. Additionals

- Different initialization could yield a different result.
- In practise, we usually start from many random initializations and choose the one which minimizes the objective function.
- K could be tuned via cross validation.
- Euclidean distance is not the only choice for k-means, sometimes we could also try to use correlation distance.

K-Medoids

1. Algorithm



1. Assign each sample to a cluster from 1 to K arbitrarily, e.g. at random.
2. Iterate these two steps until the clustering is constant:
   - For a given cluster assignment C find the observation in the cluster minimizing total pairwise distance with the other cluster members:
     $$i_k^* = \operatorname*{argmin}_{\{i:C(i)=k\}} \sum_{C(i')=k} d(x_i, x_{i'}).$$
     Then $z_k = x_{i_k^*}$, $k = 1, 2, \dots, K$ are the current estimates of the cluster centers.
   - Given a current set of cluster centers $\{z_1, \dots, z_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center:
     $$C(i) = \operatorname*{argmin}_{1 \le k \le K} d(x_i, z_k).$$

2. Additionals

Comparing with k-means, k-medoids clustering is more explanable because the centroid is required to be one of the observations. And one only needs pairwise distances for K-medoids rather than the raw observations.