3.2.1 K-means

The goal here is to place each data point in a group whose center it is closest to, and then adjust the group centers to better represent the points in them

The first clustering method we will consider is the so called K-means or Lloyd's algorithm. To derive the algorithm, we fix a maximum number of groups $K \in \mathbb{N}$ and consider two variables: a clustering variable $C := (C_1, \ldots, C_K)$ (represents the clusters or groups of data points) and a mean variable $m := (m_1, \ldots, m_k)$. Here $C_k \subset \mathcal{D}$ (Each C_k is a subset of $\mathcal{D}$ (the whole dataset)) for $k = 1, \ldots, K$ are subsets of the data points with $\bigcup_{k=1}^{K} C_k = \mathcal{D}$ (the union of all clusters should include all the data points: every data point must belong to one of the clusters and no point is left out) and $C_i \cap C_j = \emptyset$ for $i \neq j$ (clusters are disjoint, no data point can belong to more than one cluster). These subsets are interpreted as the k-th group. Additionally, $m_k \in \mathbb{R}^d$ for $k = 1, .., K$ and m_k is the mean of the k-th group.

The K-means algorithm is an iterative method that alternates between updating the variables C and m, holding the other constant

Given a clustering $C = (C_1, \ldots, C_K)$ the k-th mean m_k is updated as the mean of the data point in C_k :

3.2.1

$$m_k \leftarrow \frac{1}{|C_k|} \sum_{\substack{i=1,\ldots,N \\ x_i \in C_k}} \forall k = 1, \ldots, K$$

where $|C_k| := \#\{i = 1, \ldots, N | x_i \in C_k\}$ is the number of data points in the k-th group. This corresponds to the reconstruction rule 2

(The mean of each cluster is updated using the average of all points in that cluster, where $|C_k|$ represents the number of data points in cluster C_k)

On the other hand, fixing the means $m = (m_1, \ldots, m_K)$, the k-th group C_k is defined as the set of all points that are closer to m_k than to any other m_j for $j \neq k$.

To achieve this, we assign each data point to the nearest mean:

3.2.2

$$k_i \in \arg\min_{k \in \{1,\ldots,K\}} ||x_i - m_k||, \quad i = 1, \ldots, N,$$

resolving any ambiguity by choosing a group assignment (e.g., randomly). This is the assignment rule 1'.

(Here each data point is reassigned to the nearest cluster center. For each point x_i find the

cluster k whose mean m_k is closest to x_i. If multiple clusters are equidistant, a random assignment can be made)
Using this assignment to the nearest mean, we can now update the clustering as follows

3.2.3

$$C_k \leftarrow \{x_i | i = 1, \ldots, N, k_i = k\}$$

(Update cluster C_k to contain all points x_i, such that the assigned cluster index $k_i = k$)
By definition $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\bigcup_{k=1}^{K} C_k = \mathcal{D}$.
The K-means algorithm iterates through the steps 3.2.1 to 3.2.3 for a certain number of iterations or until the means and clustering change little or not at all, see:

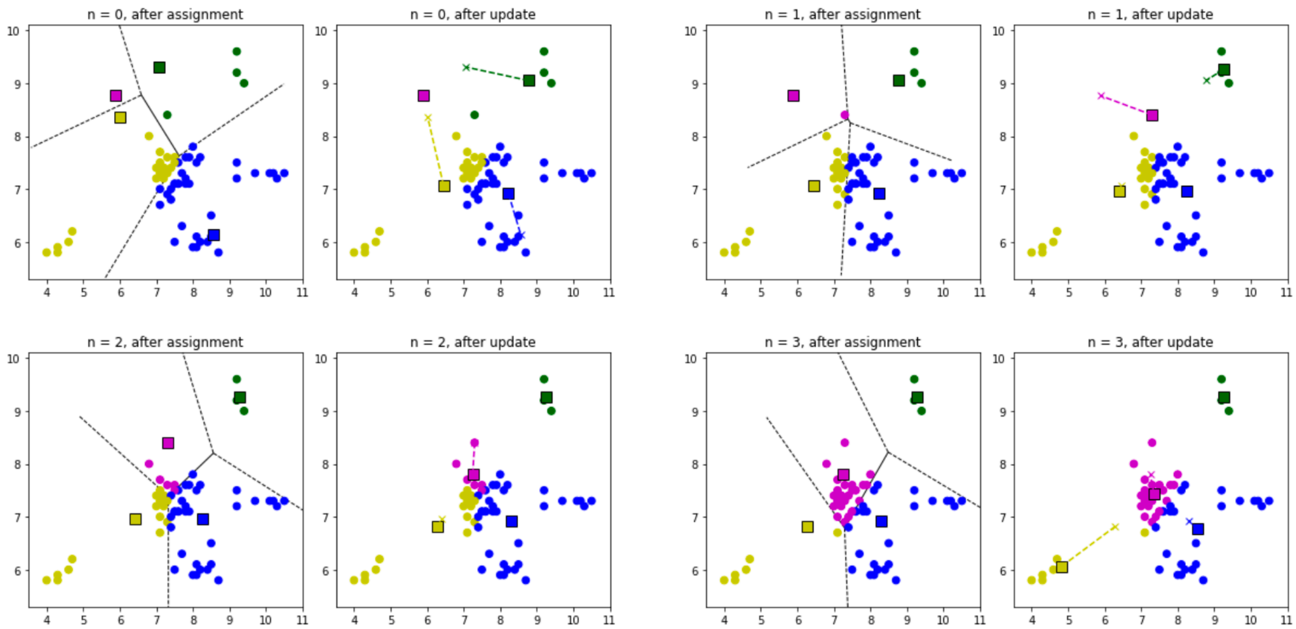

Figure 3.1: Four iterations of the K-means algorithm for $K = 4$.

The general assignment rule 1 for arbitrary points $x \in \mathbb{R}^d$ is obtained exactly as in 3.2.2 via

$$x \mapsto k(x) \in \arg \min_{k \in \{1, \ldots, K\}} ||x - m_k||$$

where in case of ambiguity, a k in the argmin is chosen (e.g. randomly)

Remark 3.2.1 (Voronoi Cells)

The discrete clustering $C = (C_1, \ldots, C_K)$ can also be replaced by so-called **Voronoi cells** of the means $m = (m_1, \ldots, m_K)$. These are defined by

3.2.5

$$V_k := \{x \in \mathbb{R}^d | \ \ ||x - m_k|| \leq ||x - m_j|| \, \forall j = 1, \ldots, K\}, k = 1, \ldots, K$$

The cell V_k consists of all points that are closer to m_k than to any other means, see Figure blow.

- A voronoi cell V_k is the set of all points in the space ($\mathbb{R}^D$) that are closer to the centroid m_k than to any other centroid m_j for $j \neq k$
- If you imagine placing K cluster centers (means), the space around each center is divided into regions, where each region contains all the points that are closest to that specific center. These regions are called **Voronoi cells**.
  Note that the interiors of these Voronoi cells are not necessarily disjoint! For example, if two means are the same, the cells coincide.
  In fact, the K-means algorithm 3.2.1 to 3.2.3 has the property that it decreases the so-called clustering energy $E(C, m)$ (how good the clustering is) at each step. This energy is defined as

  $$E(C, m) := \sum_{k=1}^{K} \sum_{\substack{i=1,...,N \\ x_i \in C_k}} ||x_i - m_k||^2$$

  But what does this formula mean?
  The clustering Energy $E(C, m)$ measures how well the centroids represent the data points in each cluster

1. For each cluster C_k compute the squared distance between each data point x_i in that cluster and the cluster's mean m_k
2. Sum these squared distances across all clusters
   Why squared distance?
   It penalizes points that are far from the centroid more than closer ones, encouraging tighter clusters
   A **lower energy** means that the clusters are well formed (points are close to centroids)
   A **higher energy** means that the clustering isn't very good (points far from their assigned centroids)

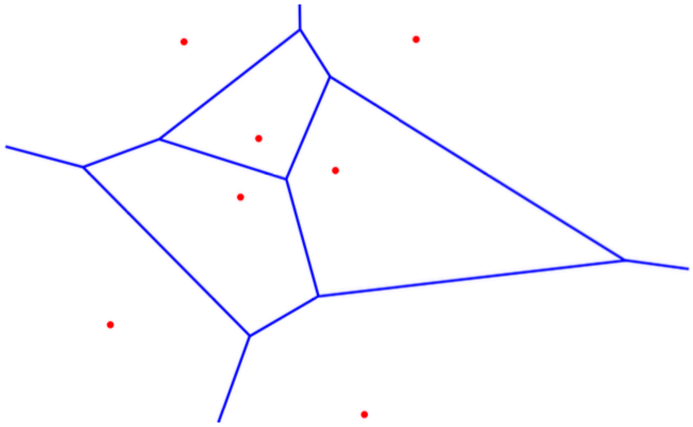

Figure 3.2: Voronoi cells regarding the red data points.

Note that the dependence on the clustering variable C is quite complicated. Finally, C is neither a vector nor a number but a collection of K disjoint subsets of $\mathcal{D}$. Thus, one cannot use a gradient-based algorithm to minimize E with respect to C.
Instead, the following holds:

Proposition 3.2.1

For all clusterings C and for m defined as in 3.2.1, it holds that

$$E(m, C) \leq E(n, C)$$

for all other n.
This means: If we fix the clustering C and update the centroids m, the energy $E(m, C)$ will always decrease or stay the same. This means updating the centroids to the mean of the assigned points is always an improvement!

For all means m and for $C := (C_1, \ldots, C_K)$ defined in 3.2.2 and 3.2.3, it holds that

$$E(m, C) \leq E(m, D)$$

for all other clusterings $D = (D_1, \ldots, D_K)$ with $\bigcup_{k=1}^{K} D_k = \mathcal{D}$ (all points in dataset must be included in one of the clusters of D_k and $D_i \cap D_j = \emptyset$ for $i \neq j$ (no clusters overlap; each data point belongs to exactly one cluster)
This means: If we fix the centroids m and update the clusters C by assigning points to the closest centroid, the energy will always decrease or stay the same. This means the reassignment step (placing each point in the closest cluster) improves or maintains clustering quality

The K-means algorithm also has some disadvantages. On the one hand, it struggles with more complex data, as illustrated in the Figure in the next chapter. Another problem is that K-means does not provide information about the certainty of the group assignment. It would be desirable to obtain a probability vector with K probabilities for each data point.