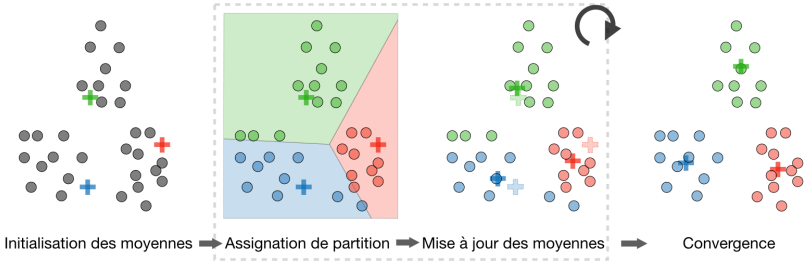
2.2.2 Partitionnement k-means

On note $c^{(i)}$ la partition du point i et μ_j le centre de la partition j.

❒ **Algorithme** – Après avoir aléatoirement initialisé les centroïdes de partitions $\mu_1,\mu_2,...,\mu_k \in \mathbb{R}^n$, l'algorithme k-means répète l'étape suivante jusqu'à convergence :

$$c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||^2 \quad \text{et} \quad \mu_j = \frac{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}}}$$



❒ **Fonction de distortion** – Pour voir si l'algorithme converge, on regarde la fonction de distortion définie de la manière suivante :

$$J(c,\mu) = \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

2.2.3 Regroupement hiérarchique

❒ **Algorithme** – C'est un algorithme de partitionnement avec une approche hiérarchique qui construit des partitions intriqués de manière successive.

❒ **Types** – Il y a différents types d'algorithme de regroupement hiérarchique qui ont pour but d'optimiser différents fonctions objectif, récapitulés dans le tableau ci-dessous :

Ward linkage	Average linkage	Complete linkage
Minimize within cluster distance	Minimize average distance between cluster pairs	Minimize maximum distance of between cluster pairs

2.2.4 Indicateurs d'évaluation de clustering

Dans le cadre de l'apprentissage non-supervisé, il est souvent difficile d'évaluer la performance d'un modèle vu que les vrais labels ne sont pas connus (contrairement à l'apprentissage supervisé).

❒ **Coefficient silhouette** – En notant a et b la distance moyenne entre un échantillon et tous les autres points d'une même classe, et entre un échantillon et tous les autres points de la prochaine partition la plus proche, le coefficient silhouette s d'un échantillon donné est défini de la manière suivante :