Classification non supervisée
Mélanges Gaussiens & ALGO EM

I. Algorithme k-means.

Notations :

- Observations $x_1, \_\_, x_N \in \mathbb{R}^D \longmapsto c_1, \_\_, c_N \in [\![1, K]\!]$
  (N obs, K classes, D features)
- $c_n(t)$: classe de l'individu n après t itérations.
- $r_{nk}(t)$: assignation $\begin{cases} 1 & \text{si } c_n(t) = k \\ 0 & \text{sinon} \end{cases}$ $\begin{pmatrix} n \in [\![1, N]\!] \\ k \in [\![1, K]\!] \end{pmatrix}$
- $\mu_k(t)$: centroid / prototype de classe k

$$J(t) = \sum_{n=1}^{N} \| x_n - \mu_{c_n}(t) \|^2$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}(t) \cdot \| x_n - \mu_k(t) \|^2$$

$$R = \begin{array}{c} \\ ind_1 \\ \vdots \\ ind_n \end{array} \begin{array}{c} c_1 \quad c_2 \quad \cdots \\ \begin{bmatrix} \text{———} \; 1 \; \text{——} \\ \text{—} \; 1 \; \text{———} \\ \vdots \end{bmatrix} \end{array} \begin{array}{l} \\ \Sigma = 1 \\ 1 \\ \vdots \end{array}$$

Algorithme K-means :

1. Initialiser les centroids
2. Répéter :

   a. minimiser $J(t)$ par rapport aux $(r_{nk})_{n,k}$

$$c_n(t+1) = \underset{1 \leq j \leq k}{\operatorname{argmin}} \| x_n - \mu_j(t) \|$$

   b. miniser $J(t)$ par rapport aux $(\mu_k)_k$

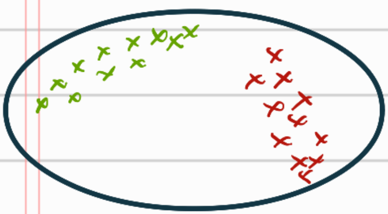
$$\frac{\partial J(t)}{\partial \mu_k(t)} = -2 \sum_{n}^{N} r_{nk}(t) \left(x_n - \mu_k(t) \right)$$

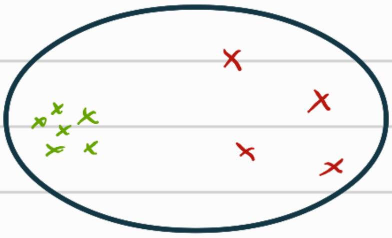
$$= -2 \cdot \sum_{n=1}^{N} \left(x_n - \mu_k(t) \right) \cdot r_{nk}(t)$$

$$\frac{\partial J(t)}{\partial \mu_k} = 0 \longleftrightarrow \mu_k(t) = \frac{\sum^{N} r_{nk}(t) \cdot x_n}{\sum^{N} r_{nk}(t)} = \frac{1}{N_k} \sum^{N} x_n$$

$$\rightarrow \mu_k(t+1) = \frac{1}{N_k} \cdot \sum_{\substack{n=1 \\ c_n=k}}^{N} x_n$$

Rmq : . l'algorithme converge rapidement
. Il fonctione mal quand les clusters sont irréguliers, inhomogènes (variances différentes), ou anisotropes.





. Il n'est pas normalisé, donc on ne peut évaluer la performance de l'algo de façon standardisée pour toutes les applications
(on sait qu'on veut J petit, mais pas à quel point)

. En particulier en grande dimension, J peut être très grand $\longrightarrow$ on peut utiliser une ACP en amont pour réduire la dim.

. K-means converge vers un min local qui dépend de l'initialisation : essayer plusiers fois avec plusiers initialisations différentes & sélectionner le meilleur J.

. kmeans ++ : initialisation qui espace les centroides les uns des autres le + possible.
↳ accélerer & augmenter la perf.

. très sensible aux outliers.
. ne prend pas en compte l'incertitude proche des frontières entre les clusters.