CLASSIFICATION NON SUPERVISÉE, MÉLANGES GAUSSIENS ET ALGORITHME EM

I Algorithme K-means

Notations :

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

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

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

Algorithme K-means :

1. Initialiser les centroïdes.
2. Répéter :
   a. minimiser $J(t)$ par rapport aux $(r_{nk})_{n,k}$

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

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

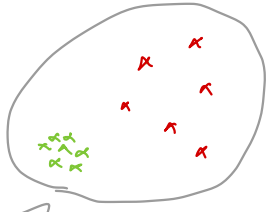
$$\frac{\partial J(t)}{\partial \mu_k(t)} = -2 \sum_{n=1}^{N} r_{nk}(t) (x_n - \mu_k(t))$$

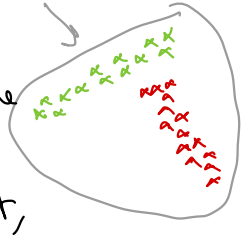
$$= -2 \sum_{n=1}^{N} r_{nk}(t)\left(x_n - \mu_k(t)\right)$$

$$\frac{\partial J(t)}{\partial \mu_k(t)} = 0 \iff \mu_k = \frac{\sum_{n=1}^{N} r_{nk}(t)\, x_n}{\sum_{n=1}^{N} r_{nk}(t)}$$

$$= \frac{1}{N_k} \sum_{\substack{n=1 \\ c_n = k}}^{N} x_n$$

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





Rmq :

- L'algorithme converge rapidement
- Il fonctionne mal si les clusters sont irréguliers, inhomogènes (variances différentes), ou anisotropes
- $J(t)$ n'est pas normalisée, donc on ne peut évaluer la performance de l'algorithme 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 dimension.
- K-means converge vers un minimum local qui dépend de l'initialisation : essayer plusieurs fois avec plusieurs initialisations différentes et sélectionner le meilleur J.
- kmeans++ : initialisation qui espace les centroïdes les uns des autres le plus possible. Ça accélère et augmente la perf.

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