k-means

1 Objectif de la classication non-supervisée

On dispose de n observations $X_1, \dots, X_n$ décrites par p variables numériques. Les variables sont généralement standardisées pour éviter qu'une variable domine les autres à cause de son échelle. Notre objectif est de regrouper ces n observations en K groupes (ou classes), de sorte que :

1. Les observations au sein d'un même groupe soient les plus similaires possible (dans un certain sens).
2. Les observations appartenant à des groupes différents soient les moins similaires possible.

Autrement dit, nous cherchons une fonction de classification :

$$C : \{1, \dots, n\} \to \{1, \dots, K\}$$

qui, à chaque observation $i \in \{1, \dots, n\}$ associe une étiquette de groupe $C(i) \in \{1, \dots, K\}$.

Définition : fonction de coût

La qualité d'une partition C est mesurée à l'aide d'une **fonction de coût** W, qui évalue la somme des distances intra-groupes :

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d(X_i, X_j),$$

où $d(X_i, X_j)$ est une mesure de la dissimilarité entre les observations X_i et X_j. Par exemple, on peut utiliser la distance euclidienne.

Plus $W(C)$ est faible, meilleure est la qualité de la partition au sens de la cohésion intra-groupe. Le problème de la classication non-supervisée est donc un problème d'optimisation combinatoire. Il s'agit de trouver la fonction C qui minimise $W(C)$.

Cependant, ce problème est très difficile (voir impossible) à résourdre exactement. En effet, il existe K^n combinaisons possibles (chaque observation pouvant appartenir à un des K groupes).

Il est donc compliqué d'explorer toutes les solutions possibles même pour de petites valeurs de n et de K.

Pour résoudre ce problème, on utilise des **algorithmes gloutons** (*greedy algorithm*) qui procèdent de manière itérative. L'idée est de, premièrement, explorer un sous-ensemble restreint de l'espace des partitions. Ensuite, l'algorithme améliore progressivement la solution de manière itérative. Ceux-ci ne garantissent pas d'atteindre le minimum global, mais plutôt un minimum local qui est souvent une bonne solution en pratique.

Hypothèses

1. Les p variables sont numériques, catégorielles ou ordinales, souvent centées-réduites.
2. Le nombre de groueps K est fixé avant de lancer l'algorithme (choisi par l'utilisateur).

2 Algorithme k-moyennes

L'**algorithme des k-moyennes** (k-means) est une méthode classique de classication non-supervisée. L'objectif est de regrouper n observations en K groupes homogènes, i.e. tels que les observations au sein d'un même groupe soient aussi proches que possible, tandis que celles appartenant à des groupes différents soient aussi éloignées que possible. Cette méthode repose sur une mesure de dissimilarité, généralement la distance euclienne pour des données quantitatives.

Algorithme

Voici les principales étapes de l'algorithme des k-moyennes.

1. Choix du nombre de groupes K : Le nombre de classes doit être choisi à l'avance.
2. Initialisation : On partitionne aléatoirement les n observations en K groupes (ou on choisit aléatoirement K observations comme centres initiaux).
3. Calcul des centroïdes : Pour chaque groupe k, on calcule leur centre de gravité

$$\mu_k = \frac{1}{N_k} \sum_{i:C(i)=k} X_i, \quad \text{pour} \quad k = 1, \dots, K$$

où N_k est le nombre d'observations dans le groupe k.

4. Réaffectation : Chaque observation est affectée au centre le plus proche, i.e. au groupe dont le centroïde minimise la distance à l'observation.
5. Itération : On répète les étapes 3 et 4 jusqu' 'à la statibilisation des groupes, i.e. jusqu'à ce qu'aucune observation ne change de groupe.

La convergence de l'algorithme est garantie en un nombre fini d'itérations, car chaque étape réduit l'inertie intra-groupe, i.e. la somme des distances des observations à leur centroïdes respectif. En revanche, rien ne garantit que l'algorithme atteigne un minimum global de W, il peut converger vers un minimum locale.

Exemple avec une petite visualisation.

Cette algorithme présente plusieurs limites. Tout d'abord, il est sensible à l'initialisation des groupes. Les résultats peuvent varier d'un essai à l'autre en fonction du choix initial des centroïdes. Ce problème peut être résolu en utilisant l'algorithme `kmean++` (cf. link). Deuxièmement, l'algorithme nécessite de connaître à l'avance le nombre de groupes K, ce qui n'est pas toujours évident. Il existe plusieurs critères pour guider ce choix, tels que le coefficient de silhouette, la méthode du coude (**elbow method**) ou encore des critères d'information comme le BIC ou l'AIC dans un cadre probabiliste. Troisièmement, à chaque itération, l'algorithme requiert le recalcul de toutes les distances entre les observations et les centroïdes. Cela peut devenir coûteux en temps de calcul lorsque le nombre d'observations n ou le nombre de variables p est élevé. Enfin, l'algorithme est assez sensible aux valeurs extrêmes. La moyenne est en effet influencée par les observations atypiques, ce qui peut fausser le calcul des centroïdes et engendrer des regroupements incohérents.

3 Algorithme k-médoides

Pour remédier à certaines des limites des k-moyennes, on peut utiliser une variante plus robuste : l'algorithme des k-médoïdes. Contrairement aux k-moyennes, les k-médoïdes utilisent des observations réelles comme représentants. Plus précisément, dans chaque groupe, le médoïde est l'observation qui minimise la somme des distances aux autres observations du même groupe.

Cette méthode présente plusieurs avantages par rapport aux k-moyennes. D'abord, elle est plus robuste aux valeurs extrêmes, car les médoïdes sont moins influencés par les observations atypiques que les moyennes. Ensuite, elle est compatible avec des variables ordinales ou catégorielles, à condition d'utiliser une mesure de dissimilarité appropriée. On peut aussi spécifier une matrice de dissimilarité sur mesure, adaptée à la nature des données.

En revanche, l'algorithme des k-médoïdes partage certains inconvénients avec celui des k-moyennes. Il faut aussi spécifier à l'avance le nombre de groupes K. De plus, son coût computationnel est plus élevé, en particulier si l'on utilise des distances non euclidiennes ou si l'on travaille avec un grand nombre d'observations.

Exemple avec une comparison avec le k-means.