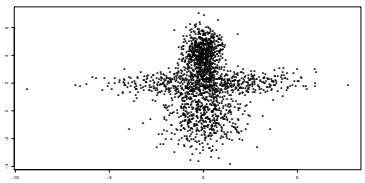
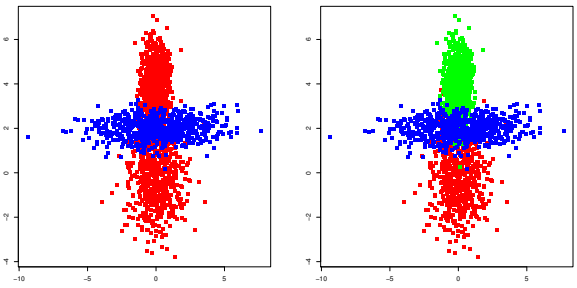
Clasification non supervisée

1 Introduction

- On dispose de données concernant n individus. Questions :
  Existe-t-il des individus qui se ressemblent ? Comment définir leur ressemblance ? Y-a-t-il plusieurs groupes d'individus ? Combien de groupes ?
- La classification non supervisée désigne un ensemble de méthodes dont le but est de trouver une partition des individus en K classes à partir d'observations sur ces individus. On sous-entend donc que les observations, bien que collectées lors d'une même expérience, ne sont pas toutes issues de la même population homogène, mais plutôt de K populations.
  Exemples pratiques :
  — trouver des "profils types" de clients d'une banque à partir des transactions bancaires de n individus.
  — établir une typologie de patients en K groupes en fonction de leur historique médical (puis étudier la réponse à un traitement en fonction du groupe des patients)
  — text mining : retrouver des textes proches...
- Supervisé ou non ?
  Dans le cas non supervisé, on fait une recherche à l'aveugle ie, l'appartenance des individus à un groupe n'est pas connue et c'est ce que l'on cherche à obtenir.
  En supervisé on a un échantillon d'apprentissage (individus pour lesquels on connait le groupe) et on cherche à construire une règle de classement pour prédire l'appartenance de nouveaux individus.
- En français on parle de classification non supervisée, en anglais on parle de clustering. Et attention en anglais, le terme classification fait référence à la classification supervisée !
- But : partitionner les individus en K classes. C'est un problème d'optimisation : on cherche la "meilleure" partition au vu d'un certain critère. En effet, il est impossible de tout explorer car le nombre de partitions de n individus en K classes est approximativement (pour n grand) $K^n/K!$ (nombre de Stirling). Par exemple il y a $\approx 8e + 46$ partitions de $n = 100$ individus en $K = 3$ classes.
  Les stratégies développées sont donc itératives :
  — on part d'une partition réalisable
  — étape $(m + 1)$: on cherche une partition P^{m+1} à partir de la partition P^m telle que P^{m+1} soit meilleure.
  Questions sous-jacentes :
  → passer d'une partition à la suivante
  → critère qualité ?
  → choix du nombre de classes K.
- Différents types de clustering :
  - par agglomération : CAH
  - par partitionnement : K-means
  - par modélisation : modèle de mélange (cours Apprentissage an prochain)

Problème qui peut être difficile : les groupes peuvent être assez bien définis et séparés, ou au contraire imbriqués/sans frontières claires.





2 Notion de dissimilarité et distances

2.1 Dissimilarité ou distances entre individus

On va chercher à regrouper des individus qui se ressemblent.

Définition. Une dissimilarité d est une fonction qui à tout couple d'individus $(x_i, x_{i'})$ associe une valeur dans $\mathbb{R}^+$ qui vérifie $d(x_i, x_{i'}) = d(x_{i'}, x_i)$ et $d(x_i, x_{i'}) = 0$ ssi $x_i = x_{i'}$.
Une distance est une dissimilarité qui possède en plus l'inégalité triangulaire.

Quelles données a-ton sur les n individus ? quantitatives, qualitatives ? plus compliquées : courbes, réseaux, textes .. ?

En fonction du type de données, il existe une multitude de manière de calculer des distances ou dissimilarité entre objets.

Remarque. Attention, il est bien plus important de choisir la distance entre individu que l'algorithme de classification, c'est l'étape cruciale ! Voici les distances et dissimilarités les plus courantes.

On travaille sur un tableau de données X de taille (n, p). La ligne x_i renseigne sur l'individu i et la colonne x^j sur la jème variable (voir ACP).

1. Cas où les p variables sont quantitatives.
   - La distance euclidienne est la plus souvent utilisée :

$$\begin{aligned} d(x_i, x_{i'}) := d^2(x_i, x_{i'}) &= \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \\ &= (x_i - x_{i'})' M (x_i - x_{i'}) \text{ avec } M = I \end{aligned}$$

   - Comme en ACP, on peut préférer travailler sur les données normalisées (ie réduites), ce qui revient à travailler sur la distance précédente avec $M = diag(1/s_1, \ldots, 1/s_p)$ où s_j est l'écart-type empirique de la j-ème variable.
   - La distance dite de Mahalanobis peut aussi être utilisée pour atténuer la structure de corrélation entre les variables. Dans ce cas M est la matrice de variances-covariances empirique de X.
   - La distance L^1, plus robuste aux valeurs extrêmes :

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$$

2. Cas où les p variables sont qualitatives.
   - Il existe de nombreuses dissimilarités basées sur le nombre de caractères communs aux deux individus, sur le nombre de caractères possédés par l'un mais pas l'autre ... (voir dans la littérature, les indices de concordance, Jacard, Dice)
   - basée sur les profils lignes (cf AFC) : voir TP. On utilise alors les premières composantes principales comme variables quantitatives à la place des variables qualitatives.
3. Mélange de variables qualitatives et quantitatives : soit on rend tout qualitatif en regroupant les variables quantitatives sous forme de classe ; soit on effectue une AFC des variables qualititatives (ou ACM si plus de 2 variables qualitatives) et on utilise les premières composantes principales comme variables quantitatives à la place des variables qualitatives. Dans le package FactoMineR, une fonction a été crée exprès pour les données mixtes (qualitatives et quantitatives). C'est la fonction FAMD.

Remarque. Parfois on dispose (au lieu d'une matrice X de données) directement de la matrice des distances ou dissimilarités entre les n individus.

2.2 Inertie

On cherche une partition des individus en classes homogènes, bien séparées entres-elles, ce qui revient à regarder l'inertie intra et inter classe.

Définition. (Cas de la distance euclidienne)
Soit $\mathcal{C} = \{C_1, \ldots, C_K\}$ une partition des individus $\{x_1, \ldots, x_n\}$ en K classes. On note $g = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ le centre de gravité du nuage de points et $g_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ le centre de gravité de la classe C_k. On définit les inerties suivantes :

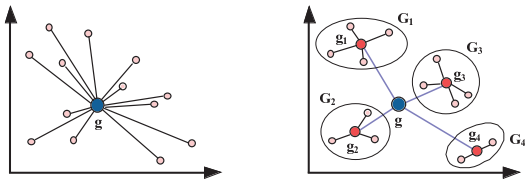
— Inertie totale : $I_T = \frac{1}{n} \sum_{i=1}^{n} d^2(x_i, g)$

— Inertie inter-classe : $I_{inter} = \frac{1}{n} \sum_{k=1}^{K} |C_k| \times d^2(g_k, g)$ mesure la dispersion des centres des classes

— Inertie intra-classe : $I_{intra} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} d^2(x_i, g_k)$ mesure la dispersion des points dans une même classe.

Propriété 1. Propriété de Huygens

$$I_T = I_{inter} + I_{intra}$$



Objectif : avoir des classes homogènes, bien séparées entres-elles, ce qui revient à :

$$\text{Minimiser l'inertie intra-classe} \iff \text{Maximiser l'inertie inter-classe}$$

3 Algorithme de type kmeans

3.1 Algorithme kmeans

Initialisation : choisir K classes et K centres de classes g_k
tant que le clustering n'est pas stabilisé ou tant que $(T < T_{max})$

- pour chaque individu i , on cherche dans quelle classe l'affecter. On cherche donc la classe dont le centre est le plus proche (au sens de la distance ou dissimilarité choisie) de l'individu i :

$$k^* = \underset{k}{\operatorname{argmin}}\ d(x_i, g_k)$$

  et on affecte l'individu i dans la classe C_{k^*}.
- Mise à jour du centre de chaque classe.

Remarque. Cet algorithme est implémenté dans la plupart des logiciels pour des variables quantitatives avec la distance euclidienne. Dans ce cas, le centre de classe est le centre de gravité de la classe. L'algorithme K-means peut être généralisé à des variables qualitatives mais alors le centre de gravité n'a plus de sens, et il est remplacé par le mode ou un autre représentant de la classe (voir K-médoids).

Propriété 2. A chaque étape l'inertie intra-classe diminue. L'algorithme converge vers une partition réalisant un minimum local de l'inertie intra-classe. En général une dizaine d'itérations suffisent. Mais la solution est très sensible à l'initialisation du point de départ.

3.2 Choix des centres de classes initiaux

— K individus choisis aléatoirement
— K individus choisis grâce à des connaissances complémentaires
— essayer plusieurs points de départ : répétition de l'algorithme $Nsim$ fois (avec des centres choisis aléatoirement) et sélection de la meilleure des $Nsim$ partitions (ie celle avec la plus faible inertie intra-classe)

3.3 Choix du nombre de classes K

C'est la difficulté de toute méthode de classification non supervisée !
— Critère fondé sur les inerties : recherche d'un coude dans la décroissance de l'inertie intra-classe (on s'arrête si l'inertie intra-classe ne décroit quasiment plus entre deux itérations)
— Critère de silhouettes : La silhouette d'une classification est un graphe indiquant à quel point chaque individu est bien assigné à sa classe (voir littérature par exemple wikistat pour plus de détails). Pour un individu i, une silhouette $s(i)$ proche de 1 signifie que l'individu est bien classé, une silhouette négative que l'individu est mal placé. Le nombre de classe retenu est celui qui maximise la moyenne des silhouettes pour les n individus.

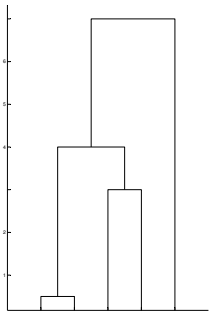
3.4 Variantes de l'algorithme K-means

Algorithme des K-médoides : le centre d'une classe est un médoid, ie le point de la classe le plus proche de tous les autres. Un médoid fait donc partie des données.
L'algorithme est moins sensible aux valeurs atypiques mais il est plus couteux en mémoire et temps de calcul car il faut calculer des distances pour trouver le médoid.

4 Classification ascendante hiérarchique

La classification ascendante hiérarchique, notée CAH, a pour objectif de construire une suite de partitions emboitées des données (on parle aussi de hiérarchie).

Exemple. Représentation par dendrogramme d'une hiérarchie indicée



4.1 Principe de la classification hiérarchique

Principe CAH : Classification ascendante hiérarchique
— A l'étape initiale, les n individus constituent des classes à eux seuls.
— On calcule les distances 2 à 2 entre individus, et les deux individus les plus proches sont réunis en une classe.
— La distance entre cette nouvelle classe et les $n-2$ individus restants est ensuite calculée, et à nouveau les 2 éléments (classes ou individus) les plus proches sont réunis.
— Ce processus est réitéré jusqu'à ce qu'il ne reste plus qu'une unique classe constituée de tous les individus.
→ Besoin d'une distance entre individus (voir section kmeans), mais aussi d'une distance entre classes.

Remarque. Une autre stratégie pourrait être de partir de la classe constituée de tous les individus, puis de procéder par divisions succesives jusqu'à obtenir des classes réduites à un élément. On parle de classification descendante hiérarchique, mais c'est moins utilisée dans la littérature.

4.2 Mesures d'agrégation entre classes

Voici plusieurs mesures d'aggrégation classiques entre deux classes A et B. Ces mesures utilisent la distance ou dissimilarité qui a été choisie pour comparer deux individus.

Définition. Cas où les données sont quantitatives et d^2 est la distance euclidienne. On peut alors calculer les centres de gravité de chaque classe et définir les distances suivantes entre deux classes A et B :

— mesure entre barycentres g_A et g_B :

$$D(A,B) = d^2(g_A, g_B)$$

— mesure de Ward = mesure de l'augmentation d'inertie :

$$D(A,B) = \frac{card(A)card(B)}{card(A)+card(B)} d^2(g_A, g_B)$$

La fusion de deux classes s'accompagne toujours d'une augmentation de la variabilité intra-classe, mais on veut que cette augmentation soit la plus petite possible. La mesure de Ward est la stratégie la plus courante dans le cas d'une distance euclidienne entre individus car elle permet à chaque étape de l'algorithme, de limiter la hausse de l'inertie intra-classe.

Quelle distance/dissimilarité inter-groupes si on ne dispose pas de la distance euclidienne ?

Définition. Disssimilarités entre deux classes A et B :

— (single linkage algorithm) mesure d'agrégation du lien minimum :

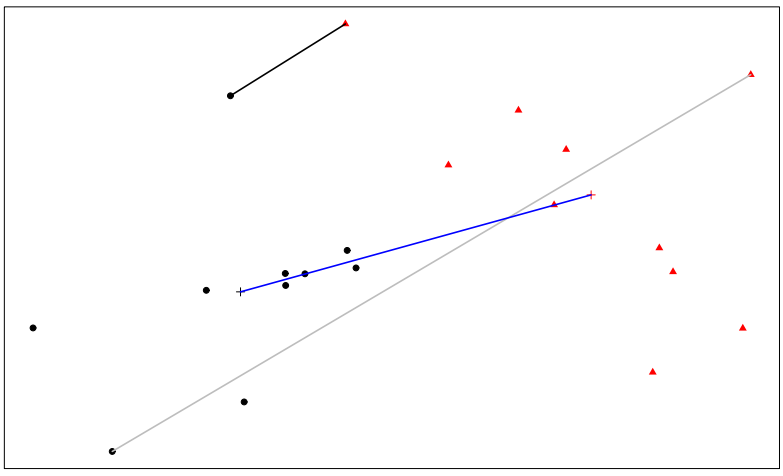
$$D(A,B) = \min_{i\in A, j\in B} d(i,j)$$

— (complete linkage algorithm) mesure d'agrégation du lien maximum :

$$D(A,B) = \max_{i\in A, j\in B} d(i,j)$$

— (average linkage algorithm) mesure d'agrégation du lien moyen :

$$D(A,B) = \frac{1}{card(A)card(B)} \sum_{i\in A, j\in B} d(i,j)$$



Lien minimum : ne prend en compte qu'une seule observation par groupe (la plus proche) ce qui peut créer des "paquets" (grande variabilité intra-classe).
Lien maximum : deux groupes sont proches si toutes les observations dans la réunion sont relativement proches. Crée des petits groupes homogènes (grande variabilité inter-classe).
Lien moyen représente un compromis entre les deux.

4.3 Représentation du résultats d'une CAH par un dendogramme

La classification hierarchique permet une représentation des résultats de classification sous la forme d'un arbre ou dendrogramme. La hauteur d'une branche est proportionnelle à la distance (ou dissimilarité) entre les deux objets regroupés. Dans le cas du saut de Ward, c'est la perte de variance inter-classes.
Une fois un nombre de classes sélectionné par l'un des critères proposés dans la suite, une coupure de l'arbre fournit la répartition des individus en classes.

4.4 Coupure du dendrogramme ie choix du nombre de classes

Par construction, l'algorithme CAH fournit une classification des données en K classes, pour tout K entre 1 et n. Ainsi, il n'est pas nécessaire de préciser a priori le nombre de classes que l'on souhaite. Ce choix peut être réalisé a posteriori, en considérant le dendrogramme. Les hauteurs des branches étant proportionnelles à la distance inter-classes, on peut choisir une classification en "coupant" l'arborescence lorsque les branches sont jugées trop grandes. En effet une grande branche indique que l'on regroupe des classes qui ne sont pas homogènes.
Sinon (voir k-means) :

- Critères fondés sur les inerties.
- Critère de silhouettes.

5 Utilisation combinée K-means et CAH :

Comparaison K-means/CAH :

- L'algorithme K-means nécessite de se fixer un nombre K de classes.
- La CAH construit la matrice des distances (n, n) et donc n'est pas adaptée si n trop grand.

Dans la pratique, on peut combiner les deux, notamment lorsque n est grand :

- si n grand : on fait les K-means en demandant un grand nombre de classes (typiquement $K = n/10$)
- puis on fait une CAH sur les centres des classes. On détermine alors un K "optimal".
- puis on ré-exécute un K-means avec le K "optimal" et pour initialisation les centres des classes obtenues à l'étape précédente.