Quelques éléments de classification non supervisée.

F. Panloup
LAREMA-Université d'Angers

—

INTRODUCTION

—

Classification non supervisée : définition/exemples

- Ensemble de méthodes ayant pour objectif de dresser une *typologie* à partir de n observations et p paramètres.
- Typologie : Caractérisation des sous-populations de notre échantillon.
- Exemple 1 : Clients d'une banque. Objectif : Caractériser les K profils types (K à déterminer en général) en fonction des informations associées au client (transactions bancaires, catégorie professionnelle,...).
- Exemple 2 : Cohorte de patients dont on a p mesures physiologiques. Objectif : Etablir des groupes en fonction de ces caractéristiques pour ensuite étudier la réaction au traitement de chacun de ces groupes.

Classification non supervisée : Etape préliminaire ?

- Etablir des typologies peut être un objectif en soit ou une pré-étape dans le processus d'apprentissage.
- Dans l'exemple 1, c'est plutôt un objectif en soit.
- Dans l'exemple 2, l'objectif est supervisé : prédire la réaction au traitement mais cette étape d'homogénéisation peut rendre plus efficace la partie supervisée.
- Remarque : la classification peut être trop dirigée (comme son nom l'indique). Par exemple, dans la recherche sur le cancer, associer "de force" un label (tel que la survie à 2 ans par exemple) à un ensemble de caractéristiques n'est pas forcément judicieux.

K-Means

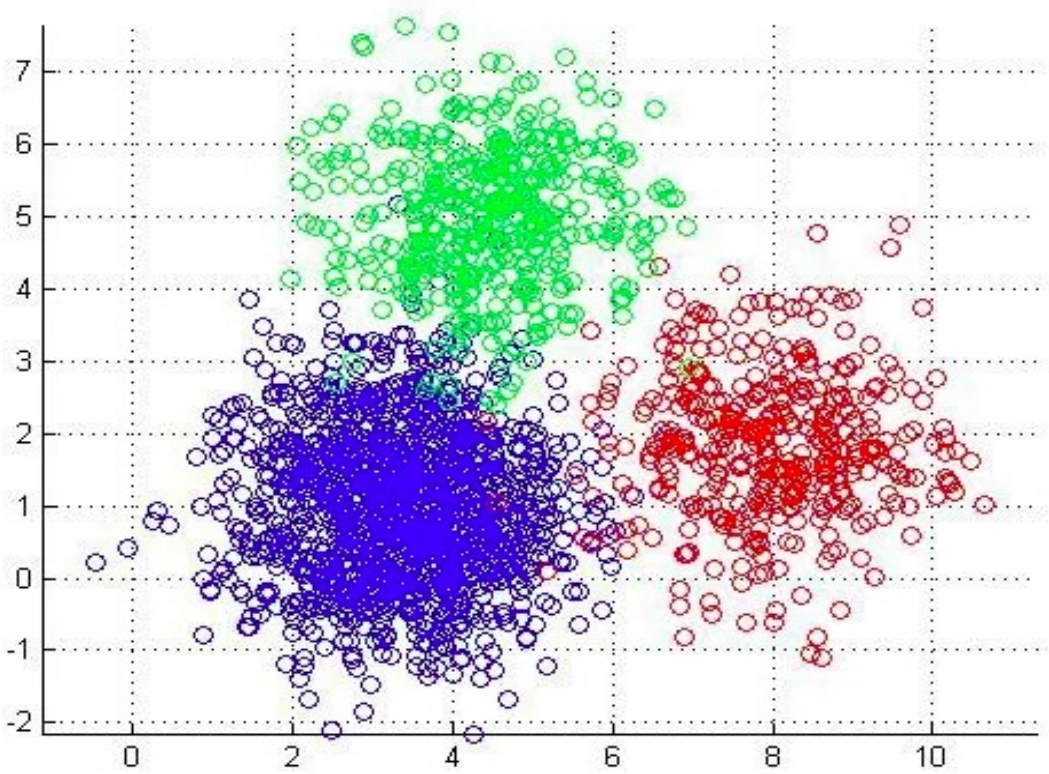

Figure: 3 groupes

Objectif

Pour une distance d donnée (euclidienne par exemple) et un nombre de types/groupes K déterminer le minimiseur global de la fonction :

$$F(c_1, \ldots, c_K) = \frac{1}{n} \sum_{k=1}^{n} \min_{j=1}^{K} d(x_i, c_j)^2.$$

où

- les x_i désignent les observations. On supposera ici que $x_i \in \mathbb{R}^p$.

- $c_1, \ldots, c_K$ vivent dans $\mathbb{R}^p$ et désignent les *centroïdes* : A chaque observation, on associe un tel c_j. Il s'agit du point de l'espace parmi les $c_1, \ldots, c_K$ le plus proche de x_i.

Pour la distance euclidienne, $d(x_i, c_j)^2 = \|x_i - c_j\|_2^2 = \sum_{k=1}^{p} (x_i(k) - c_j(k))^2$.

Algorithme

Notons $(c_1^\star, \dots, c_K^\star)$ ce minimiseur. On constitue alors K-groupes de la manière suivante :

$$G_k = \{x_i, \quad d(x_i, c_k^\star) = \min_{j=1}^{k} d(x_i, c_j^\star)\}.$$

Méthode déterministe d'approximation de $(c_1^\star, \dots, c_K^\star)$. On initialise en $(\mu_1, \dots, \mu_K) \in \mathbb{R}^p$ puis à chaque étape :

- on associe les points de l'échantillon à leur centre le plus proche.
- On définit K-nouveaux centres en prenant les barycentres de chacun des groupes.
- La suite $n \mapsto F(c_1^{(n)}, \dots, c_K^{(n)})$ est décroissante. Elle converge donc vers un minimum **local**.

K-means par descente de gradient stochastique

Lorsque le nombre de points est élevé, la réalisation de chaque étape peut être coûteuse. On peut alors remplacer par une version stochastique de l'algorithme :

- A chaque étape, on tire un point au sort x_i parmi les n (uniformément). On calcule son centre C_i.

- On remplace C_i par $C_i = C_i - \gamma(x_i - C_i)$ (descente de gradient stochastique associée à la "distorsion"...)

Nombre de clusters ?

En pratique, comment choisir K ? Il faut observer la valeur de la fonction F en fonction de K. Idéalement, on observe ce type de phénomène :

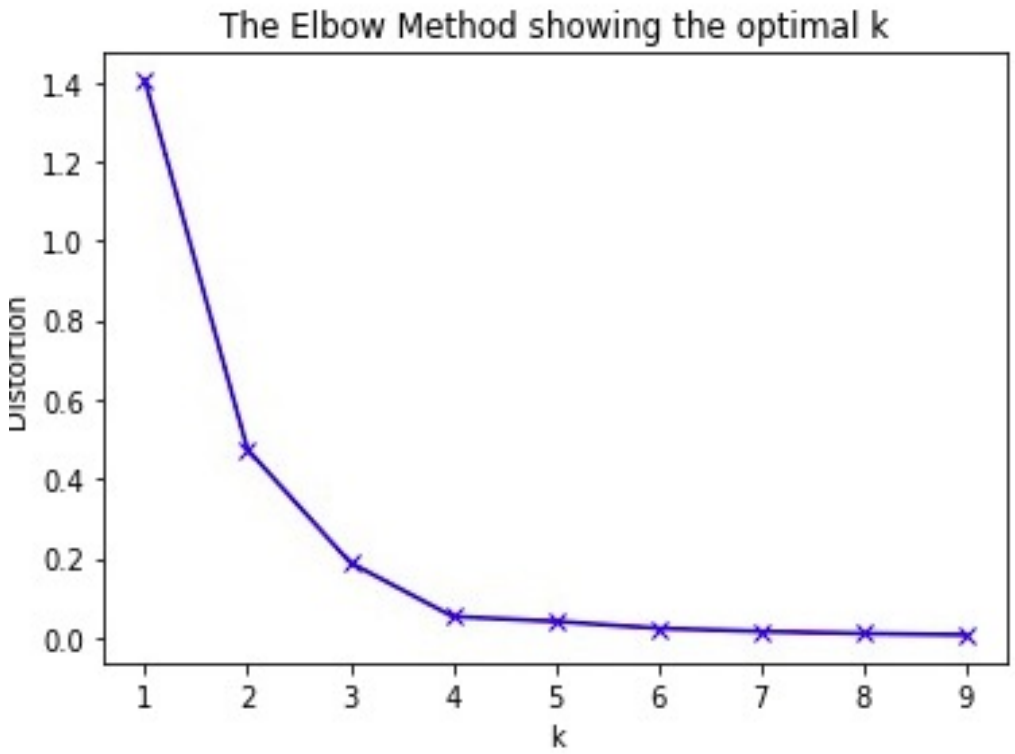

Figure: Méthode du coude

Classification hiérarchique ascendante

- on fabrique des groupes de manière itérative.
  - Etape 0 : On calcule les distances deux à deux entre individus, et les deux individus les plus proches sont réunis en une classe.
  - Etape k : La distance entre cette nouvelle classe et les $n-2$ individus restants est ensuite calculée, et à nouveau les deux éléments (classes ou individus) les plus proches sont réunis.
- Les K-means et la classif. hiérarchique ascendante peuvent être utilisés conjointement (voir `http://www2.agroparistech.fr/IMG/pdf/ClassificationNonSupervisee-AgroParisTech.pdf` pour un poly très complet sur le sujet).

Compléments sur l'ACP

Même si l'Analyse en Composantes Principales a des objectifs différents de la classification non supervisée (réduction de dimension par exemple), nous choisissons ici d'ajouter quelques compléments sur le sujet.

- ACP Sparse : Le principe de l'ACP Sparse est de chercher le meilleur sous-espace vectoriel au sens des moindres carrés parmi les sous-espaces vectoriels qui sont engendrés par des vecteurs propres ayant "peu de coordonnées allumées".
- Par exemple, pour un seul vecteur v, le problème consiste à trouver

$$v^\star = \mathrm{Argmax}_{\{v, \|v\|=1, |v|_0 \leq r\}} v^T \mathbf{X}\mathbf{X}^T v.$$

- En pratique, c'est un peu plus compliqué...
- Avantage : il y a moins de coordonnées à estimer, la variance est donc plus petite.

ACP à noyau

Theorem

Tout algorithme pour vecteurs qui puisse ne s'exprimer qu'en termes de produits scalaires entre vecteurs peut être effectué implicitement dans un espace de Hilbert en remplaçant chaque produit scalaire par l'évaluation d'un noyau défini positif sur un espace quelconque.

- Pour faire de l'ACP à noyau, il faut donc montrer que l'ACP peut s'exprimer uniquement à l'aide du produit scalaire.

Avant l'ACP, le calcul de distance

Soient $x_1, x_2 \in \mathbb{R}^d$:

$$d_{\text{eucl.}}(x_1, x_2)^2 = \langle x_1, x_1 \rangle - 2\langle x_1, x_2 \rangle + \langle x_2, x_2 \rangle.$$

Soit K un noyau défini positif et Φ l'application associée. On peut alors définir :

$$d(\Phi(x_1), \Phi(x_2))^2 = K(x_1, x_1) - 2K(x_1, x_2) + K(x_2, x_2).$$

Exemple : si $K(x_1, x_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{|x_1 - x_2|^2}{2\sigma^2})$, alors

$$d(\Phi(x_1), \Phi(x_2))^2 = \frac{2}{\sigma\sqrt{2\pi}} - 2\frac{2}{\sigma\sqrt{2\pi}} \exp(\frac{|x_1 - x_2|^2}{2\sigma^2}).$$

ACP et produit scalaire

Trouver la k-ième direction de l'ACP, c'est résoudre le problème suivant : déterminer

$$w = \mathrm{Argmin}_{w \perp (w_1, \dots, w_{k-1})} \frac{1}{n} \sum_{i=1}^{n} \frac{\langle x_i, w \rangle^2}{\|w\|^2}.$$

C'est donc bien un problème qui s'exprime en fonction du produit scalaire. D'un point de vue fonctionnel, résoudre l'ACP revient à déterminer

$$f = \mathrm{Argmin}_{f \in \mathcal{H}, f \perp (f_1, \dots, f_{k-1})} \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)^2}{\|f\|_{\mathcal{H}}^2}.$$

où $\mathcal{H}$ est l'ensemble des fonctions $f(x) = \langle w, x \rangle$ muni du produit scalaire $\langle f, \tilde{f} \rangle_{\mathcal{H}} = \langle w, \tilde{w} \rangle_{\mathbb{R}^d}$.

ACP à noyau

On peut transposer ce problème à un espace de fonctions différent via l'astuce du noyau. Par exemple, si

$$K(x, y) = \langle x, y \rangle^2.$$

Alors, on peut montrer que l'espace $\mathcal{H}$ associé est l'ensemble des fonctions f de la forme

$$f_S(x) = x^T S x$$

où S est une matrice symétrique, muni du produit scalaire

$$\langle f_{S_1}, f_{S_2} \rangle_{\mathcal{H}} = \langle S_1, S_2 \rangle_F = \sum_{i,j} S_1(i,j) S_2(i,j).$$

Et comment fait-on ?

Par le théorème "du représentant",

$$f_k = \sum_{i=1}^{n} \alpha_{i,k} K_{x_i}$$

où $K_{x_i} : x \mapsto K(x_i, x)$. On a $\langle K_{x_i}, K_{x_\ell} \rangle_{\mathcal{H}} = K(x_i, x_\ell)$. Ainsi,

$$\|f_k\|^2_{\mathcal{H}} = \sum_{i,\ell} \alpha_i, \alpha_{\ell,k} K(x_i, x_\ell) = \alpha_k^T \mathbf{K} \alpha_k$$

où $\mathbf{K} = (K(x_i, x_\ell))_{i,\ell}$ et

$$\sum_{i=1}^{n} f_k(x_i)^2 = \sum_{i=1}^{n} (\sum_{i=1}^{n} \alpha_{i,k} K(x_i, x_k))^2 = \ldots = \alpha_k^T \mathbf{K}^2 \alpha_k$$

de sorte que

$$\alpha_k = \mathrm{Argmin}_\alpha \frac{\alpha^T \mathbf{K}^2 \alpha}{\alpha^T \mathbf{K} \alpha}.$$

sous la contrainte $\alpha_k^T \mathbf{K} \alpha_j = 0$ pour tout $j \leq k - 1$.

Et comment fait-on ?

On diagonalise alors $\mathbf{K}$ (la matrice de Gram) : $K = U\Delta U^T$ avec $\Delta_1 \geq \Delta_2 \ldots \geq \Delta_n \geq 0$. Ensuite, on pose $\beta_k = \mathbf{K}^{\frac{1}{2}}\alpha$ et on cherche

$$\mathrm{Argmin}_\beta \beta^T \mathbf{K}\beta \quad (= \mathrm{Argmin}_\alpha \frac{\alpha^T \mathbf{K}^2 \alpha}{\alpha^T \mathbf{K}\alpha}).$$

sous les contraintes $\beta_k^T \beta_j = 0$ pour $j \leq k-1$ et $\beta_k^T \beta_k = 1$. Ce sont exactement les vecteurs propres de $\mathbf{K}$. Il suffit alors d'inverser pour obtenir les α_k. On trouve

$$\alpha_k = \frac{1}{\Delta_k}\beta_k.$$

Finalement, la projection d'un point x sur la k-ième direction, *i.e.* pour son image $\Phi(x)$ s'écrit :

$$\langle \Phi(x), f_k \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_{i,k} \langle \Phi(x), \Phi(x_i) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_{i,k} K(x, x_i).$$

On peut alors replacer les points dans la base $\{f_1, \ldots, f_d\}$.

Exemple

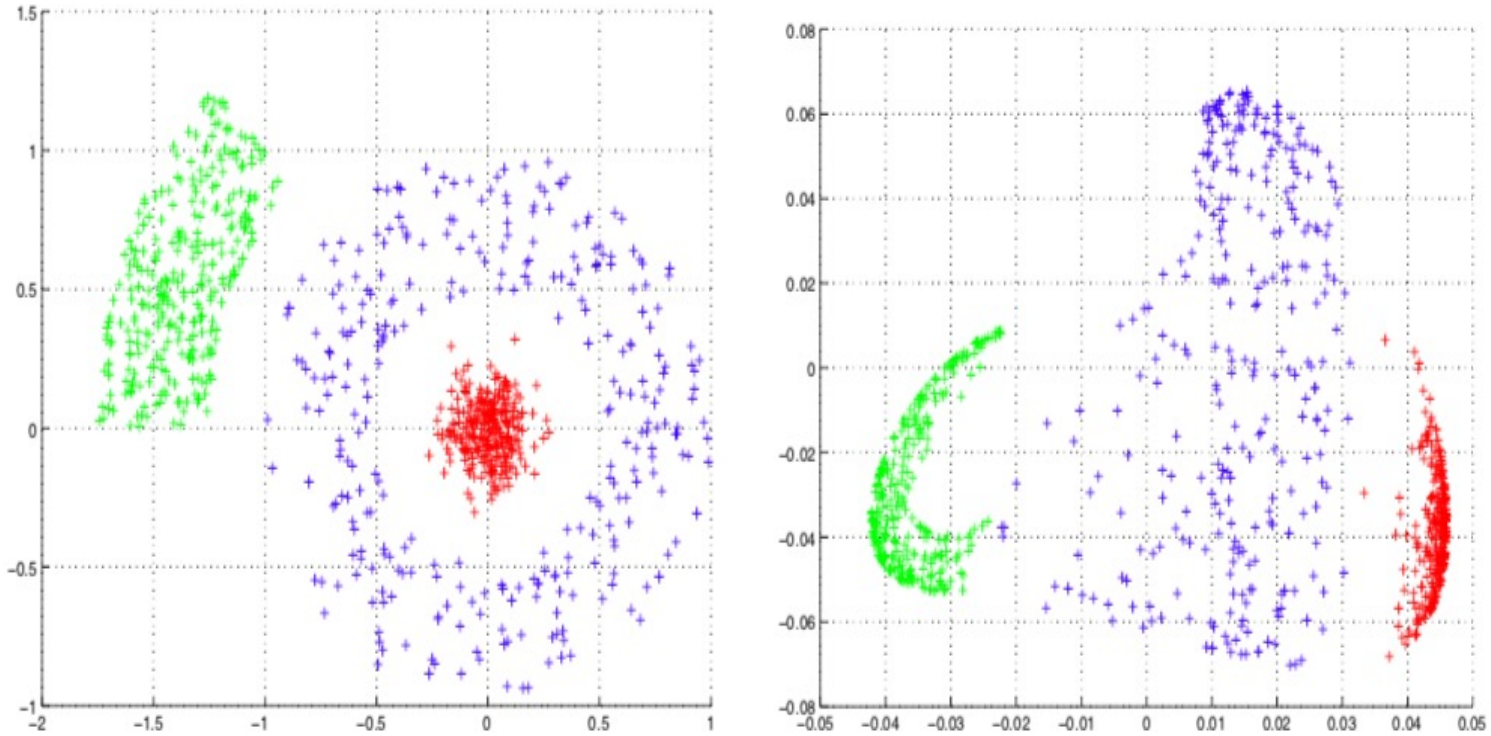

Figure: ACP à noyau (gaussien ?)