Classification non supervisée

Franck Picard*

*UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

Outline

Introduction

- Permet de synthétiser un tableau de données en groupes d'individus
- Objectif : réduire le nombre d'individus
- Différence avec l'analyse factorielle qui s'occupe plus des variables
- C'est une analyse **descriptive**

$$\underset{[n\times p]}{\mathbf{X}} = \begin{bmatrix} x_{1,1} & \cdots & \cdots & x_{1,p} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ x_{n,1} & \cdots & \cdots & x_{n,p} \end{bmatrix} \rightarrow \begin{bmatrix} g_{1,1} & \cdots & g_{1,p} \\ \vdots & \ddots & \vdots \\ g_{K,1} & \cdots & g_{K,p} \end{bmatrix}$$

Clustering

$$\searrow \begin{bmatrix} z_{1,1} & \cdots & z_{1,K} \\ \vdots & & \vdots \\ z_{n,1} & \cdots & z_{n,K} \end{bmatrix}$$

Analyse Factorielle

Hypothèses sous jacentes

- Etant donné un groupe d'individus, existe-t-il des individus qui se ressemblent ?
- Comment définir leur ressemblance ? Une classification n'a de sens qu'au regard du critère qu'on utilise pour la construire
- Mais avant tout : l'hypothèse est-elle justifiée ? Y a-t-il vraiment plusieurs groupes ?
- La question du nombre de groupe est donc centrale : c'est une question de choix de modèle
- En général on y répond à la fin !

Supervisé ou non ?

- La différence entre les deux situations est la connaissance des classes
- dans le cas non supervisé, on fait une recherche à l'aveugle
- alors en qu'en supervisé on a un échantillon d'apprentissage
- En supervisé on connait un indicateur de performance : le taux de mal classé, mais en non supervisé?
- Difficile de valider les résultats

Un problème de combinatoire

- On cherche bien sur la "meilleure" partition possible au vu d'un certain critère: Peut-on explorer toutes les partitions et choisir la meilleure ?
- Soit E un ensemble de n individus que l'on souhaite partitionner en K classes
  - Le nombre de partitions de E à K groupes (nombre de Stirling de première espèce)

$$g(n,k) \sim K^n/K!$$

  - Le nombre total de partitions de E (nombre de Bell)

$$B_n = \sum_{k=1}^{n} g(n,k) = \frac{1}{e}\sum_{k\geq 1}\frac{k^n}{k!}$$

Conclusion:

On ne peut pas explorer toutes les partitions possibles !

Stratégies itératives

- Les stratégies développées sont itératives et visent à explorer un sous ensemble de partitions dans lequel on espère que se trouve la partition optimale

Agglomérative	Classif. Hierarchique
Partitionnement	K-means
Probabiliste	modèles de mélange

- Mais comment faire des groupes d'individus "bizarres" comme des textes ? des réseaux, des courbes ?
- A chaque nature de données il faut savoir développer une méthode de classification car les logiciels courants ne sont adaptés qu'à des situations standards

Outline

Notations & notion de dissimilarité

- On travaille sur un tableau de données $\mathbf{X}$ de taille $n \times p$
- La ligne i $\mathbf{x}_i$ renseigne sur l'individu i, la colonne $\mathbf{x}^j$ sur la variable j
- Définition d'une dissimilarité d

$$\begin{aligned} d : E \times E &\rightarrow \mathbb{R}^+ \\ (i, i') &\rightarrow d(i, i') \end{aligned}$$

- Propriétés: $d(i, i') = d(i', i), \forall i', \ d(i, i) \leq d(i, i'), d(i, i) = 0$
- Distance : possède en plus l'inégalité triangulaire

$$d(i, i') \leq d(i, i') + \leq d(i', i'')$$

Cas de plusieurs variables

- En général on définit une dissimilarité $d(x_{ij}, x_{i'j})$ entre les individus i et i' pour la variable j

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

- la distance la plus utilisée est la distance euclidienne

$$d^2(i, i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

- Mais est-elle toujours la plus appropriée ?

Nuages de points

- On travaille à partir de $\mathbf{X}_{[n\times p]}$ sur le nuage des individus :

$$(\mathbf{x}_1, \ldots \mathbf{x}_n) \in \mathbb{R}^p$$

- Le centre de gravité de $\mathbf{X}$ (moyenne générale):

$$g = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} = \bar{x}^j = (g^1, \ldots, g^p)$$

- On utilise la moyenne de chaque variable:

$$g^j = \sum_{i=1}^{n} x_{ij} = \bar{x}^j$$

- On travaille en général sur le tableau centré: $\mathbf{X} - (g^1, \ldots, g^p)$

Notion d'inertie d'un nuage

- L'inertie totale d'un nuage de point mesure la variabilité totale de la position des points dans l'espace

$$\mathbf{I}_T = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - g)^2$$

- C'est la variance totale du nuage de point
- La contribution de **chaque variable** à la dispersion totale

$$v^j = \sum_{i=1}^{n}(x_{ij} - g^j)^2$$

- On préfère normaliser les données pour éviter les problèmes d'échelle

$$\mathbf{I}_T = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{ij} - g)^2}{v^j}$$

Pourquoi la distance euclidienne ?

- C'est la distance la plus utilisée
- Elle s'interprète de manière géométrique
- Elle produit des formes linéaires dans les problèmes d'optimisation
- Elle a un lieu avec le cadre gaussien $((x-\mu)^2/\sigma^2)$
- Comment faire quand les données ne peuvent pas être modélisées comme variables gaussiennes ?
- Réponse : avec les modèles probabilistes (mélange)

Autres distances

- La distance en valeur absolue (L^1) est plus robuste aux valeurs extrêmes

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

- Quand les données sont des comptages (variables qualitatives)

$$d^2(i, i') = \sum_{j=1}^{p} \left(\frac{x_{ij}}{x_{+j}} - \frac{x_{i'j}}{x_{+j}} \right)^2$$

Warnings!

Il est bien plus important de choisir la distance entre individu que l'algorithme de classification, c'est l'étape cruciale

Outline

Intuitions et principes

- A partir d'une matrice de dissimilarité, permet de former des groupes "de proche en proche" :
  - en divisant deux groupes (classification descendante)
  - en agglomérant des groupes (classification ascendante)
- Créent des hierarchies entre groupes (meme si l'emboitement n'a pas de sens du point de vue de l'interprétation)
- Chaque niveau de la hierarchie représente une partition particulière des données en groupes disjoints
- La hierarchie peut etre représentée sous forme d'arbre ou **dendrogramme**

Classification hierarchique indicée

- On s'intéresse aux hierarchies dites indicées
- Cet indice permet de passer de la matrice de distance/dissimilarité au dendrogramme,
- il permet de construire / représenter l'arbre
- A chaque partition on peut associer une valeur numérique représentant le niveau auquel ont lieu les regroupements

3 Ingrédients pour une classification hierarchique

1. dissimilarité entre individus : on peut donner directement la matrice en général, ce qui en fait un outil flexible
2. une dissimilarité entre groupes : étant donnée que l'on aggrège (divise) les individus en groupes, puis les groupes en groupes de groupes
3. une règle de fusion (division)

Distance euclidienne et Décomposition de l'inertie

- On souhaiterait définir une distance **entre groupes** qui permet d'obtenir une faible variance intra-groupe (homogénéité) **et** une forte variance inter-groupe (séparabilité)
- On utilise la notion d'inertie (déjà réduite)

$$\mathbf{I}_T = \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - g)^2 = \sum_{i=1}^{n} d^2(i, g)$$

- L'inertie se fonde sur une distance euclidienne pour la distance **entre individus**

Variance inter, Variance intra

- Si on considère une partition en K groupes ayant pour centres de gravité $(g_1, \ldots, g_K)$, on peut décomposer $\mathbf{I}_T$ telle que

$$\mathbf{I}_T = \sum_{k=1}^{K} \sum_{i \in k} d^2(i, g_k) + \sum_{k=1}^{K} N_k d^2(g_k, g) = \mathbf{I}_W + \mathbf{I}_B$$

- $\mathbf{I}_T$ représente la variabilité totale du nuage de point : **elle est constante pour un jeu de données fixé**.
- $\mathbf{I}_W$ (within variance) représente la dispersion des points autour de leur centre.
- $\mathbf{I}_B$ (between variance) représente la séparabilité des groupes : à **maximiser**

Stratégie d'optimisation

- On cherche à trouver une partition "optimale".
- Une partition optimale sera définie par une $\mathbf{I}_W$ minimale et une $\mathbf{I}_B$ maximale
- Etant donné que $\mathbf{I}_T = \mathbf{I}_W + \mathbf{I}_B$, si on maximise $\mathbf{I}_B$ on minimise $\mathbf{I}_W$
- Dans ce cadre les algorithmes de classification sont des algorithmes d'optimisation
- L'ensemble des solutions n'étant pas "visitable", on adopte des stratégies itératives d'optimisation

Distance euclidienne et Méthode de Ward

- Motivation: la fusion de deux groupes s'accompagne toujours d'une augmentation de la variabilité, mais on veut que cette augmentation soit la plus petite possible
- A chaque étape de fusion des groupes, on veut minimiser l'augmentation de la variance intra-groupes

$$\begin{aligned} \mathbf{I}_W(A,B) &= \mathbf{I}_W(A) + \mathbf{I}_W(B) = \sum_{i \in A} d^2(i, g_A) + \sum_{i \in B} d^2(i, g_B) \\ \mathbf{I}_W(A \cup B) &= \sum_{i \in A \cup B} d^2(i, g_{AB}) \end{aligned}$$

- Quelle est la distance $d^2(A,B)$ entre les groupes A et B qui permette d'atteindre l'objectif ?

$$\Delta = \mathbf{I}_W(A,B) - \mathbf{I}_W(A \cup B)$$

Propriétés de la distance de Ward

- On peut montrer que

$$\Delta = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B)$$

- Si on choisit la distance entre groupes,

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

  alors on minimise l'augmentation de l'inertie intra groupes à chaque étape de la hierarchie
- C'est une distance qui prend en compte les effectifs des groupes et qui créera des groupes équilibrés

Lien minimum, maximum, moyen ?

- Quelle distance/dissimilarité inter-groupes si on ne dispose pas de la distance euclidienne ? pour les dissimilarités ?
- On considère 2 groupes A et B et on définit la dissimilarité des deux groupes $d(A, B)$
- Lien minimum (simple, plus proches voisins)

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

- Lien maximum : $d(A, B) = \max_{i \in A, i' \in B} d(i, i')$
- Lien moyen

$$d(A, B) = \frac{1}{n_A \times n_B} \sum_{i \in A, i' \in B} d(i, i')$$

Dernière étape : Dendrogramme et représentation

- On a construit un shéma de classification à partir d'un raisonnement sur les distances (dissimilarités) des individus
- 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 entre objets regroupés

Exemple de construction d'arbre

- On considère les données suivantes:

	a	b	c	d
X	0	0	3	3
Y	0	1	0	1

- Calculer la matrice de distance euclidienne D et construire le dendrogramme avec la méthode du lien simple et la méthode du lien maximum
- Autre exemple avec la matrice de distance:

	a	b	c	d	e
a	0	3	7	3	4
b		0	4	4	1
c			0	2	6
d				0	0.5
e					0

Propriétés des liens

- Si la structure en groupe est très forte, les résultats seront peu différents
- Lien minimum: ne prend en compte qu'une seule observation par groupe (la plus proche): peut créer des "paquets" (grande variabilité intra-groupe)
- Lien maximum: deux groupes sont proches si toutes les observations dans la réunion sont relativement proches. Crée des petits groupes homogènes (grande variabilité entre-groupes)
- Lien moyen représente un compromis entre les deux

Jeux de données à analyser

- http://lib.stat.cmu.edu/DASL/Stories/EconomicsofCities.html
- http://lib.stat.cmu.edu/DASL/Stories/ClusteringCars.html
- http://lib.stat.cmu.edu/DASL/Stories/EuropeanJobs.html
- http://lib.stat.cmu.edu/DASL/Stories/ProteinConsumptioninEurope.html

Outline

Une idée simple

- Algorithme des plus utilisés: très rapide (par rapport à la classification hierarchique) et facile à appréhender
- Implémenté pour la distance euclidienne pour des variables quantitatives dans les logiciels courants
- Il s'appuie sur la décomposition de l'inertie du nuage, comme la méthode de Ward
- Peut être généralisé grâce aux modèles de mélanges (pour d'autres distributions, et pour des variables quantitatives)

Le groupe le plus proche

- On introduit la notation Z_{ik} qui vaut 1 si l'individu i est dans le groupe k

$$N_k = \sum_{i=1}^{n} Z_{ik}$$

- Les inerties s'écrivent:

$$\begin{aligned} \mathbf{I}_W &= \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} d^2(i, g_k) \\ \mathbf{I}_B &= n \times \sum_{k=1}^{K} d^2(g, g_k) \end{aligned}$$

- Pour déterminer le centre le plus proche pour chaque individu:

$$\widehat{Z}_i = \arg\min_k \left\{ d^2(i, g_k) \right\}$$

Un algorithme itératif

- On souhaite minimiser la variance intra-groupes avec un algorithme itératif

$$\min_{\mathbf{Z},g_1,\ldots,g_K} \{\mathbf{I}_W\} \simeq \min_{\mathbf{Z}} \left\{ \min_{g_1,\ldots,g_K} \{\mathbf{I}_W\} \right\} \simeq \min_{g_1,\ldots,g_K} \left\{ \min_{\mathbf{Z}} \{\mathbf{I}_W\} \right\}$$

- Etape $[i]$: on trouve les centres $g_1, \ldots, g_K$
- Etape $[i+1]$: on trouve les labels $\mathbf{Z}$
- On répète ces opérations, en s'assurant qu'à chaque étape $\mathbf{I}_W$ diminue

Description des deux étapes

- Etape $[i]$: on trouve les centres $g_1, \ldots, g_K$ (moyenne des variables pour les individus de chaque groupe)

$$g_k = \frac{1}{N_k} \sum_{i=1}^{n} \widehat{Z}_{ik} \mathbf{x}_i$$

- Etape $[i+1]$: on trouve les labels $\mathbf{Z}$:

$$\widehat{Z}_i = \arg \min_{1,\ldots,K} \left\{ d^2(i, g_k) \right\}$$

Décroissance de l'inertie

- On note $[h]$ l'indice d'itération et $\mathbf{Z}^{[h]}, \mathbf{g}^{[h]}$
- L'inertie dépend donc de deux indices $\mathbf{I}_W(Z^{[h]}, \mathbf{g}^{[h]})$
- Quand on met à jour les centres :

$$\mathbf{I}_W(Z^{[h]}, \mathbf{g}^{[h+1]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik}^{[h]} d^2(i, g_k^{[h+1]})$$

- Quand on met à jour les labels :

$$\mathbf{I}_W(Z^{[h+1]}, \mathbf{g}^{[h+1]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik}^{[h+1]} d^2(i, g_k^{[h+1]})$$

Convergence de l'algorithme

- A chaque étape l'inertie diminue:

$$\mathbf{I}_W(Z^{[h]}, \mathbf{g}^{[h]}) \geq \mathbf{I}_W(Z^{[h]}, \mathbf{g}^{[h+1]}) \geq \mathbf{I}_W(Z^{[h+1]}, \mathbf{g}^{[h+1]})$$

- L'inertie est une suite bornée, donc l'algorithme converge en un nombre fini d'étapes
- Mais la solution n'est qu'un minimiseur local : il dépend du point de départ
- Les algorithmes de classification itératifs sont sensibles à l'optimisation: il faut essayer plusieurs points de départ

Stratégies d'utilisation

- Le résultat dépend du point de départ: on essaie plusieurs points initiaux et on regarde la stabilité de la solution
- Utilisation combinée avec la CAH:
  - la CAH est couteuse en temps et mémoire ($\mathcal{O}(n^2)$)
  - Dans une hierarchie, les groupes sont emboités dès les premières itérations (sensibilité au point de départ également).
  - On peut combiner CAH + k-means: on fait les k-means avec $k = n \times f$ (une fraction importante des données), et ensuite on fait une CAH sur le résultat des k-means.

Outline

Introduction

- permettent de développer des méthodes de classification dans un cadre probabiliste
- Généralise l'algorithme des k-means et l'enrichit
- Peut etre généralisé à de nombreuses formes de distributions
- Donne un cadre statistique à la sélection du nombre de groupes

Idées de base

- A partir de l'observation de la distribution des données, on suppose qu'il existe plusieurs groupes
- On fait souvent l'hypothèse qu'à l'intérieur des groupes la distribution est la même (mais pas obligé)
- Ce sont aussi des méthodes utilisées pour estimer des distributions complexes
- Les modèles de mélange permettent aussi d'effectuer une classification probabiliste

Classification probabiliste

- Les labels Z_{ik} peuvent être modélisés comme des variables aléatoires
- Si on note $(\pi_1, \ldots, \pi_K)$ le vecteur inconnu des tailles des groupes

$$\sum_k \pi_k = 1$$

- La probabilité de $\{Z_{ik} = 1\}$ correspond à la probabilité pour un tirage de tomber dans la catégorie k (généralise la distribution binomiale à plus de 2 groupes)
- Z_{ik} est une variable de distributrion multinomiale:

$$Z_{ik} \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$$

- Les moments de Z_{ik} sont: $\mathbb{E}(Z_{ik}) = \pi_k$, $\mathbb{V}(Z_{ik}) = \pi_k(1 - \pi_k)$

Probabilité *a priori* / *a posteriori*

- Z_{ik} renseigne sur le classement, mais les observations apportent de l'information pour mieux classer
- Avant d'observer x_i, la probabilité a priori de classer dans le groupe k, c'est la taille du groupe

$$\Pr\{Z_{ik} = 1\} = \pi_k$$

- Après avoir observé les données, on a plus d'information pour classer: la probabilité a posteriori de classement : $\tau_{ik} = \Pr\{Z_{ik} = 1 | X_i = x_i\}$
- On calcule cette probabilité à l'aide des probabilités conditionnelles (Bayes) :

$$\Pr\{A|B\} = \frac{\Pr\{A\}\Pr\{B|A\}}{\Pr\{B\}}$$

Calcul des probabilités a posteriori

- En utilisant la formule de Bayes on obtient:

$$\tau_{ik}(x_i) = \frac{\pi_k \Pr\{X_i = x_i | Z_{ik} = 1\}}{\sum_{\ell=1}^{K} \pi_\ell \Pr\{X_i = x_i | Z_{i\ell} = 1\}}$$

- Classification probabiliste (floue): chaque individu a une probabilité d'appartenir a un groupe (plutôt qu'une affectation déterministe)

$$\sum_{k=1}^{K} \tau_{ik}(x_i) = 1$$

- On peut retrouver les labels:

$$\widehat{Z}_i = \arg \max_{k=1,K} \{\tau_{ik}(x_i)\}$$

Illustration avec R

```
K = 2; n1 = 30; n2 = 50; n = n1+n2; prop = c(n1/n,n2/n)
mu1 = 1; mu2 = -1; s = 0.2
x = c(rnorm(n1,mu1,s),rnorm(n2,mu2,s)); x = sort(x)
f1 = dnorm(x,mu1,s); f2 = dnorm(x,mu2,s)

tau1 = prop[1]*f1/(prop[1]*f1+prop[2]*f2)
tau2 = prop[2]*f2/(prop[1]*f1+prop[2]*f2)

par(mfrow=c(2,1))
plot(x,tau1,type="l"); lines(x,tau2,type="l")
plot(x,prop[1]*f1+prop[2]*f2,type="l")
lines(x,prop[1]*f1,type="l",col="red")
lines(x,prop[2]*f2,type="l",col="blue")
```

Loi conditionnelle des observations

- La probabilité $\Pr\{X_i = x_i | Z_{ik} = 1\}$ désigne la loi des observations lorsque l'on connait l'appartenance au groupe
- C'est la loi conditionnelle des observations $X_i | \{Z_{ik} = 1\}$
- Il faut donc modéliser la distribution des observations à l'intérieur des classes
- Exemples:

$$
\begin{aligned}
X_i | \{Z_{ik} = 1\} &\sim \mathcal{N}(\mu_k, \sigma^2) \\
X_i | \{Z_{ik} = 1\} &\sim \mathcal{N}(\mu_k, \sigma_k^2) \\
X_i | \{Z_{ik} = 1\} &\sim \mathcal{P}(\lambda_k) \\
X_i | \{Z_{ik} = 1\} &\sim \mathcal{E}(\lambda_k)
\end{aligned}
$$

Loi marginale des observations

- On modélise d'abord la loi conditionnelle des observations
- On peut se demander quelle est la loi marginale des observations

$$\begin{aligned}
\Pr\{X_i = x_i\} &= \sum_{k=1}^{K} \Pr\{X_i = x_i, Z_{ik} = 1\} \\
&= \sum_{k=1}^{K} \Pr\{Z_{ik} = 1\} \times \Pr\{X_i = x_i | Z_{ik} = 1\} \\
&= \sum_{k=1}^{K} \pi_k \times \Pr\{X_i = x_i | Z_{ik} = 1\}
\end{aligned}$$

- La loi de X_i est une combinaison linéaire de lois, ou mélange de distributions

Caractérisation des mélanges de distributions

- Les formes des distributions (en général on considère des mélanges de même famille)

$$\begin{aligned} f(x) &= \pi\mathcal{N}(\mu, \sigma^2) + (1-\pi)\mathcal{E}(\lambda) \\ f(x) &= \pi\mathcal{N}(\mu_1, \sigma_1^2) + (1-\pi)\mathcal{N}(\mu_2, \sigma_2^2) \end{aligned}$$

- Les paramètres qui dépendent des groupes, et leur nombre

$$\begin{aligned} \mathcal{N}(\mu_k, \sigma^2) &: K+1 \\ \mathcal{N}(\mu_k, \sigma_k^2) &: 2K \\ \mathcal{P}(\lambda_k) &: K \end{aligned}$$

- Dans le décompte des paramètres il ne faut pas oublier les proportions $\pi_1, \ldots, \pi_K$ qui comptent pour $K-1$ paramètres étant donné que $\sum_k = \pi_k$

Vraisemblances conditionelles et complètes

- On suppose que le mélange dépend d'un paramètre $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$
- La vraisemblance des observations connaissant les groupes :

$$\mathcal{L}_K(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{k=1}^{K} f(x_i|Z_{ik}=1)^{\mathbb{I}\{Z_{ik}=1\}} = \prod_{i=1}^{n}\prod_{k=1}^{K} f(x_i;\theta_k)^{\mathbb{I}\{Z_{ik}=1\}}$$

- C'est la loi jointe de toutes les observations $(x_i)_{i=1}^n$ quand tous les labels sont connus (Z_{ik}).
- La vraisemblance des labels

$$\mathcal{L}_K(\mathbf{Z};\boldsymbol{\pi}) = \prod_{i=1}^{n}\prod_{k=1}^{K} \Pr\{Z_{ik}=1\}^{\mathbb{I}\{Z_{ik}=1\}}$$

Exemple dans le cas Gaussien

$$
\begin{aligned}
\mathcal{L}_K(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta}) &= \prod_{i=1}^{n}\prod_{k=1}^{K} f(x_i;\theta_k)^{\mathbb{I}\{Z_{ik}=1\}} \\
&= \prod_{i=1}^{n}\prod_{k=1}^{K} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\left(\frac{x_i-\mu_k}{\sigma}\right)^2\right\}^{\mathbb{I}\{Z_{ik}=1\}} \\
-2\log\mathcal{L}_K(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta}) &= n\log(\sigma^2\sqrt{2\pi}) + \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left(\frac{x_i-\mu_k}{\sigma}\right)^2
\end{aligned}
$$

Exemple dans le cas Exponentiel

$$
\begin{aligned}
\mathcal{L}_K(\mathbf{X}|\mathbf{Z};\theta) &= \prod_{i=1}^{n}\prod_{k=1}^{K} f(x_i;\theta_k)^{\mathbb{I}\{Z_{ik}=1\}} \\
&= \prod_{i=1}^{n}\prod_{k=1}^{K} \frac{1}{\lambda_k}\exp\left\{-\frac{x_i}{\lambda_k}\right\}^{\mathbb{I}\{Z_{ik}=1\}} \\
-\log\mathcal{L}_K(\mathbf{X}|\mathbf{Z};\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left[\log(\lambda_k)+\left(\frac{x_i}{\lambda_k}\right)\right]
\end{aligned}
$$

Vraisemblance marginale des observations

- C'est la loi jointe de l'échantillon quand les labels sont inconnus

$$\mathcal{L}_K(\mathbf{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\theta)$$

- Or on sait que la loi de x_i se décompose telle que:

$$f(x_i;\theta) = \sum_{k=1}^{K} \Pr\{Z_{ik}=1\} f(x_i|Z_{ik}=1) = \sum_{k=1}^{K} \pi_k f(x_i;\theta_k)$$

- La vraisemblance marginale des observations est donc

$$\mathcal{L}_K(\mathbf{X};\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f(x_i;\theta_k) \right)$$

Cadre général de l'estimation par maximum de vraisemblance pour les mélanges

- La vraisemblance marginale des observations est une fonction difficile à calculer / optimiser

$$\frac{\partial \mathcal{L}_K(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Pour estimer les paramètres d'un mélange par maximum de vraisemblance, on utilise une vraisemblance annexe, la vraisemblance des données complètes (comme si on connaissait les labels)
- la vraisemblance des données complètes est souvent plus facile à maximiser

$$\frac{\partial \mathcal{L}_K(\mathbf{X},\mathbf{Z};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Sous certaines conditions on peut montrer qu'optimiser l'une revient à optimiser l'autre

Estimation (simplifiée) des paramètres par l'algorithme EM

- Dans ce contexte l'algorithme EM s'écrit comme un algorithme des k-means (2 étapes itérées)
- Etape E: update des *posteriors*

$$\tau_{ik}^{[h+1]}(x_i) = \frac{\pi_k f(x_i; \theta_k^{[h]})}{\sum_{\ell=1}^{K} \pi_\ell f(x_i; \theta_\ell^{[h]})}$$

- Etape M: update des paramètres intra-groupes (centres dans le cas gaussien)

$$\mu_k^{[h+1]} = \frac{\sum_{i=1}^{n} \tau_{ik}^{[h+1]}(x_i) \times x_i}{\sum_{i=1}^{n} \tau_{ik}^{[h+1]}(x_i)}$$

Paramétrisation des variances dans le modèle Gaussien

- Dans le cas univarié on a le choix entre un modèle homoscedastique et hétéroscedastique:

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} \tau_{ik} \times (x_i - \mu_k)^2}{\sum_{i=1}^{n} \tau_{ik}}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \times (x_i - \mu_k)^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}}$$

- modèle hétéroscedastique a deux fois plus de paramètres ($K + 1$ vs $2K$)
- On cherche plutôt des modèles parcimonieux (avec un nombre raisonnable de paramètres), mais sans trop perdre en qualité d'ajustement aux données

Modèle Gaussien Multivarié

- Si x est un vecteur Gaussien avec p coordonnées alors sa densité s'écrit

$$f(x, \boldsymbol{\theta}) = \frac{1}{(2\pi)^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})'\Sigma^{-1}(x-\boldsymbol{\mu})\right)$$

- $\boldsymbol{\mu}$ est un vecteur de dimension p et Σ est une matrice de dimension $p \times p$
- Un modèle de mélange Gaussien multivarié avec K composantes aurait donc $K(p + p(p-1)/2)$ paramètres
- On cherche donc à réduire le nombre de paramètres en reparamétrisant le modèle.

Décomposition de la matrice de variance

- Une idée proche de l'ACP: on décompose Σ_k à l'aide de vecteur propres: $\Sigma_k = \lambda_k D_k A_k D_k'$
- Chaque paramètre s'interprète de manière géométrique

λ_k	Volume du groupe
D_k	orientation du groupe (vecteurs propres)
A_k	forme (ellipse $\pm$ allongée), matrice diagonale

- On peut considérer différents modèles $\pm$ parcimonieux

Modèle	# param.	
$\lambda \mathbf{I}$	1	kmeans
$\Sigma_k = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$	$K \times p$	pas de corr., une variance/dim
$\Sigma_k = \Sigma$	$p + p(p-1)/2$	non spécifiée mais commune
Σ_k	$K(p + p(p-1)/2)$	non spécifiée différentes

Modèles parcimonieux et sélection de modèles

- On cherche à établir un compromis entre
  - un modèle qui s'ajuste bien aux données
  - un modèle qui n'a pas trop de paramètres
- La reparamétrisation des mélanges gaussiens multidimensionnels permet une certaine parcimonie
- Le nombre d'observations étant constant pour un échantillon donné,
  - la vraisemblance augmente avec le nombre de paramètres
  - les erreurs d'estimation augmentent avec le nombre de paramètres
- Un modèle est sur-paramétré lorsqu'il a "trop" de paramètres
- Les modèles sur-paramétrés auront un très faible pouvoir prédictif

Comment définir un modèle statistique ?

- On utilise un modèle statistique pour obtenir une approximation de ce que l'on observe
- En général tous les modèles sont faux, mais certains permettent de bien synthétiser le phénomène observé
- Un modèle statistique est constitué d'une famille de lois de probabilités sur un même espace
- En général ces lois de probabilités dépendent d'un paramètre θ qui appartient à un ensemble Θ
- On note alors:

$$\mathcal{M}_\theta = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$$

Exemples de modèles statistiques

- le modèle binomial de paramètre θ:

$$\mathcal{M}_\theta = \{\theta \in \Theta = [0,1],\ \mathbb{P}\{X = 1\} = \theta\}$$

- le modèle gaussien de paramètres (μ, σ)

$$\begin{aligned}\mathcal{M}_\theta &= \left\{\theta = (\mu,\sigma) \in \Theta = \mathbb{R} \times \mathbb{R}^+,\right. \\ &\qquad \left. f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\}\end{aligned}$$

- le modèle de mélange gaussien de paramètres $(\pi_k, \mu_k, \sigma_k)_k$

$$\begin{aligned}\mathcal{M}_\theta &= \left\{\theta = (\pi_k,\mu_k,\sigma_k)_k \in \Theta = [0,1]^K \times \mathbb{R}^K \times \mathbb{R}^{+,K},\right. \\ &\qquad \left. \sum_k \pi_k = 1,\ f(x) = \sum_k \frac{\pi_k}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)\right\}\end{aligned}$$

Taille de modèle et qualité d'ajustement

- On définit la taille d'un modèle $|\mathcal{M}_\theta|$ par le nombre de paramètres (libres) qui le caractérisent
- Modèle de Bernoulli : $|\mathcal{M}_\theta| = 1$, modèle Gaussien: $|\mathcal{M}_\theta| = 2$.
- Plus le modèle sera "riche", plus il décrira les observations de manière précise
- Le nombre d'observations étant limité, un modèle riche aura comparativement peu d'observations pour estimer tous ses paramètres comparé à un modèle "plus simple"
- Il faudra prendre en compte cet élément quand on voudra comparer des modèles

Principe de la selection de modèles

- Pour comparer des modèles, on cherche à les “noter”
- La vraisemblance d’un modèle permet de quantifier la qualité d’ajustement d’un modèle aux données
- Mais cette qualité dépend du nombre de paramètres du modèle ! La vraisemblance augmente avec la dimension du modèle
- Pour une comparaison “équitable” il faut comparer des modèles en “pénalisant” leur dimension
- On utilise des vraisemblances pénalisées:

$$\log \widetilde{\mathcal{L}}(\mathcal{M}_\theta) = \log \mathcal{L}(\mathcal{M}_\theta) - \beta \text{pen}\left(|\mathcal{M}_\theta|\right)$$

Exemples de deux critères couramment utilisés

- Les critères diffèrent dans leurs objectifs et dans leur définition de la pénalité
- Les plus utilisés sont le critères d'Akaike (Akaike Information Criterion) et le BIC (Bayesian Information Criterion)

$$\begin{aligned} \text{AIC} &= \log \mathcal{L}(\mathcal{M}_\theta) - |\mathcal{M}_\theta|/2 \\ \text{BIC} &= \log \mathcal{L}(\mathcal{M}_\theta) - \frac{|\mathcal{M}_\theta|}{2} \times \log(n) \end{aligned}$$

- Dans le contexte des mélanges on utilise plutôt le BIC

Exemples d'utilisation du BIC

- On sélection le modèle qui maximise le BIC

$$\widehat{\mathcal{M}}_{\widehat{\theta}} = \arg\max \{BIC(\mathcal{M}_\theta)\}$$

- Exemple: on fixe K, on veut trouver la meilleure paramétrisation du mélange gaussien multivarié. On met en concurrence les différents modèles et on prend le meilleur du point de vue du BIC

$$\text{BIC}_K(\lambda \mathbf{I}),\ \text{BIC}_K(\text{diag}(\sigma_1^2, \ldots, \sigma_p^2)),\ \text{BIC}_K(\Sigma), \text{BIC}_K(\Sigma_k)$$

- Pour un modèle donné (exemple avec $\Sigma_k = \lambda \mathbf{I}$) on met en concurrence des modèles avec de plus en plus de groupes

$$\text{BIC}_1(\lambda \mathbf{I}),\ \text{BIC}_2(\lambda \mathbf{I}), \ldots \text{BIC}_{K-1}(\lambda \mathbf{I}),\ \text{BIC}_K(\lambda \mathbf{I})$$

Outline

Contexte

- Dans le contexte supervisé on connait les labels (y_i) et les caractéristiques (x_i) pour chaque individu
- On cherche à apprendre une règle de classification, on parle d'apprentissage statistique (statistical learning)
- On cherche à contruire un modèle prédictif pour prédire la réponse d'une nouvelle observation: $\widehat{Y}_i = \widehat{f}(x_i)$.
- Vocabulaire: on parle de régression quand on souhaite prédire des réponses continues, et de classification quand on souhaite prédire des réponses catégorielles (0/1)
- Méthodes de classification les plus utilisées: Analyse Factorielle Discriminante, Analyse discriminante probabiliste, régression logistique, Support Vecteur Machines (SVM)

Restriction aux méthodes linéaires

- On dispose d'informations sur les observations, avec une matrice $\mathbf{X}_{[n \times p]}$ dont les colonnes (features) renseignent sur la réponse $\mathbf{Y} \in \{0, 1\}$ (labels)
- On s'intéresse dans un premier temps aux règles de décision qui sont linéaires en $\mathbf{X}$.

$$f_k(\mathbf{x}_i) = \beta_{k0} + \beta_k' \mathbf{x}_i$$

- La séparation entre deux groupes se fera quand $f_k(\mathbf{x}_i) = f_\ell(\mathbf{x}_i)$, *ie* quand

$$(\beta_{k0} - \beta_{\ell 0}) + (\beta_k - \beta_\ell)' \mathbf{x}_i = 0$$

- L'ensemble des variables explicatives est divisé en régions de classement constant avec des frontières de décision linéaires

Outline

Classement par maximum a posteriori

- Comme dans le cadre non supervisé, on utilise une règle de Bayes pour classer une nouvelle observation sur la base des probabilités a posteriori

$$\Pr\{Y_i = k|\mathbf{x}_i\} = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{x}_i)}$$

- La règle de classification la plus simple est de classer l'individu i dans le groupe k quand

$$\begin{aligned} \Pr\{Y_i = k|\mathbf{x}_i\} &> \Pr\{Y_i = \ell|\mathbf{x}_i\} \\ \log \frac{\Pr\{Y_i = k|\mathbf{x}_i\}}{\Pr\{Y_i = \ell|\mathbf{x}_i\}} &> 0 \end{aligned}$$

- Il faut modéliser la loi des observations dans le groupe

Exemple dans le cas gaussien

- Rappel: $\Pr\{Y_i = k|\mathbf{x}_i\} = \Pr\{\mathbf{x}_i|Y_i = k\} \times \Pr\{Y_i = k\} = \pi_k f_k(\mathbf{x}_i)$
- On suppose que les observations suivent une loi Normale à l'intérieur des groupes (avec une matrice de variance constante)
- Rappel: $(\mathbf{x}_i - \mu_k)'\Sigma^{-1}(\mathbf{x}_i - \mu_k) = \mathbf{x}_i'\Sigma^{-1}\mathbf{x}_i - 2\mu_k'\Sigma^{-1}\mathbf{x}_i + \mu_k'\Sigma^{-1}\mu_k$
- Règle de classement:

$$\pi_k f_k(\mathbf{x}_i) > \pi_\ell f_\ell(\mathbf{x}_i)$$

Fonction de Score

- Si on calcule $\log(\pi_k f_k(\mathbf{x}_i)/\pi_\ell f_\ell(\mathbf{x}_i)) > 0$

$$\mu_k' \Sigma^{-1} \mathbf{x}_i - \frac{1}{2} \mu_k' \Sigma^{-1} \mu_k + \log \pi_k > \mu_\ell' \Sigma^{-1} \mathbf{x}_i - \frac{1}{2} \mu_\ell' \Sigma^{-1} \mu_\ell + \log \pi_\ell$$

$$(\mu_k - \mu_\ell)' \Sigma^{-1} \mathbf{x}_i + \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2} (\mu_k - \mu_\ell)' \Sigma^{-1} (\mu_k + \mu_\ell) > 0$$

- On définit la fonction de score $S(\mathbf{x}_i)$:

$$S(\mathbf{x}_i) = (\mu_k - \mu_\ell)' \Sigma^{-1} \mathbf{x}_i + \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2} (\mu_k - \mu_\ell)' \Sigma^{-1} (\mu_k + \mu_\ell)$$

- Règle de classement: si $S(\mathbf{x}_i) > 0$ on classera l'individu i dans le groupe k

Fonction de Score et relation logistique

- On se place dans le cas de 2 groupes

$$\begin{aligned} S(\mathbf{x}_i) &= (\mu_1 - \mu_0)'\Sigma^{-1}\mathbf{x}_i \\ &+ \log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 + \mu_0) \\ \log\frac{\Pr\{Y_i = 1|\mathbf{x}_i\}}{\Pr\{Y_i = 0|\mathbf{x}_i\}} &= \beta_1\mathbf{x}_i + \beta_0 \end{aligned}$$

- Fonction logistique: $\text{logit}(p(x)) = \log(p(x)/(1-p(x)))$
- Dans l'analyse discriminante probabiliste, l'hypothèse de normalité conduit à une relation logistique entre la probabilité a posteriori d'appartenir à un groupe et la fonction de score
- Il faut estimer les paramètres $\widehat{\mu}, \widehat{\Sigma}, \widehat{\pi}$.

Outline

Principe de l'analyse discriminante

- On souhaite projeer les indivus sur des axes qui les séparent avec une petite variance intra-classe et une grande variance inter-classes
- On cherche les axes $\mathbf{u}$ tels que $\min\{\mathbf{u}'\mathbf{I}_W\mathbf{u}\}$, $\max\{\mathbf{u}'\mathbf{I}_B\mathbf{u}\}$
- On ne peut pas trouver de solution simultanément aux deux problèmes
- Mais étant donné que $\mathbf{I}_T = \mathbf{I}_W + \mathbf{I}_B$, on considère

$$\max\left\{\frac{\mathbf{u}'\mathbf{I}_B\mathbf{u}}{\mathbf{u}'\mathbf{I}_T\mathbf{u}}\right\}$$

- On souhaite effectuer une projection des individus sur les axes en prenant en compte les correlations entre les variables

Introduction de la distance de Malahanobis

- Si on connait les classes on connait les matrices de variance Intra Σ (souvent notée W pour within).
- D'un point de vue probabiliste $\Sigma_k = \mathbb{V}(\mathbf{x}_i | Y_i = k)$
- On crée une nouvelle distance entre les centres des groupes qui prend en compte ces correlations:

$$\begin{aligned} D^2_{1,2} &= (\widehat{\mu}_1 - \widehat{\mu}_2)' \widehat{\Sigma}^{-1} (\widehat{\mu}_1 - \widehat{\mu}_2) \\ &= (\widehat{\mu}_1 - \widehat{\mu}_2)' \widehat{\Sigma}^{-1/2} \left\{ \widehat{\Sigma}^{-1/2} (\widehat{\mu}_1 - \widehat{\mu}_2) \right\} \end{aligned}$$

- Si A est un vecteur d'observations à composantes corrélées, $\Sigma^{-1/2} A$ le transforme en un nouveau vecteur décorrélé et de variance unité
- Dans le cas unidimensionnel ($p = 1$) on obtient:

$$\frac{n_1 n_2}{n_1 + n_2} \left(\frac{\bar{x}_1 - \bar{x}_2}{\widehat{\sigma}} \right)^2 = \frac{n_1 n_2}{n_1 + n_2} D^2_{1,2}$$

Propriétés probabilistes de la distance de Malahanobis

- $D^2_{1,2}$ est la version empirique de la distance de Malahanobis théorique

$$\Delta^2_{1,2} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- Si $\Delta^2_{1,2} = 0$ alors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$
- $D^2_{1,2}$ est un estimateur biasé de $\Delta^2_{1,2}$ tel que

$$\mathbb{E}\left(D^2_{1,2}\right) = \frac{n-2}{n-p-1}\left(\Delta^2_{1,2} + \frac{pn}{n_1 n_2}\right)$$

Classification des nouveaux individus

- On dispose d'une nouvelle observation $\mathbf{x}_i$, et on calcule sa distance au groupe

$$\begin{aligned} D^2(\mathbf{x}_i, \boldsymbol{\mu}_k) &= (\mathbf{x}_i - \widehat{\mu}_k)' \widehat{\Sigma}^{-1} (\mathbf{x}_i - \widehat{\mu}_k) \\ &= \mathbf{x}_i' \widehat{\Sigma}^{-1} \mathbf{x}_i - 2\widehat{\mu}_k' \widehat{\Sigma}^{-1} \mathbf{x}_i + \widehat{\mu}_k' \widehat{\Sigma}^{-1} \widehat{\mu}_k \end{aligned}$$

- On classera dans le groupe le plus proche au sens de cette nouvelle distance
- Pas d'hypothèse de normalité (par rapport Ã la discriminante probabiliste)
- Lien avec la fonction de score vue dans le cas gaussien. Si $\pi_1 = \pi_2$:

$$S_{12}(\mathbf{x}_i) = \frac{1}{2} \left(\Delta^2(\mathbf{x}_i, \boldsymbol{\mu}_1) - \Delta^2(\mathbf{x}_i, \boldsymbol{\mu}_2) \right)$$