Rapport de stage d'observation

Mathieu Comparetto

12 décembre 2018

Présentation de la structure d'acceuil

I.N.S.E.R.M veut dire : institut nationale de la santé et de la recherche médicale. On l'appellera désormais Inserm. Actuellement je suis dans le bâtiment 15-16

Présentation des métiers de chercheur et enseignant-chercheur

L'équipe est fondée d'enseignant chercheur et de chercheur a temps plein.Pour devenir chercheur a temps plein ou enseignant chercheur il faut déjà avoir un doctorat puis passer un concours.La gestion de recrutement est confié a la structure d'accueil des candidats et les postes ouverts sont diffusés. Les constitue les dossier incluant un résumé succin de leur parcoure avec un projet de recherche détaillé et un descriptif des travaux effectuées en thèse ou en années post-doctorales. Un comité de sélection étudie les dossiers et sélectionnent une dizaine de candidats admissibles à une audition. Le comité de sélection auditionnent les candidats et établie un classement. Les effectifs des différentes équipes de recherche du CESP sont composées d'hommes et de femmes de différentes nationalités. La majorité des chercheurs post-doctoraux sont de nationalités étrangères. La mobilité et le séjour post-doctoral à l'étranger est devenue un critère très important dans le recrutement d'un chercheur.

le service d'enseignement est de 192 heures par années repartie sur l'années selon l'emploie du temps de la formation dans laquelle il enseigne. Certains enseignants peuvent avoir un service d'enseignement étales sur l'années alors que d'autre ont un service d'enseignement étale sur une courte période de l'année. L'enseignant chercheur passe aussi son temps a préparé des cours et participe à des rénions autours de projet de recherche à demandé ou déjà financé.

Le chercheur à plein temps peut devenir chargé de recherche puis directeur de recherche et l'enseignant-chercheur peut devenir maitre de conférences puis professeur.

Déroulement des journées

Première journée

Résumé de la première journée:

Présentation avec le tuteur et son équipe d'enseignat-chercheur qui travaillent sur pharmacovigilance et biostatistique.

Discussion sur le métier de chercheur et d'enseignant chercheur

Présentation de l'outil Rstudio utilisé pour la rédaction des des développement d'enseignant et développement logiciel

Découvrte de la programmation HTML à l'aide de l'outil Rmarkdown.

Seconde journée

Résumé deuxième journée:

Mise en place du squelette de la page web du compte rendue du stage

choix des rubriques de la page web, écriture de la page web

hébergement de la page web

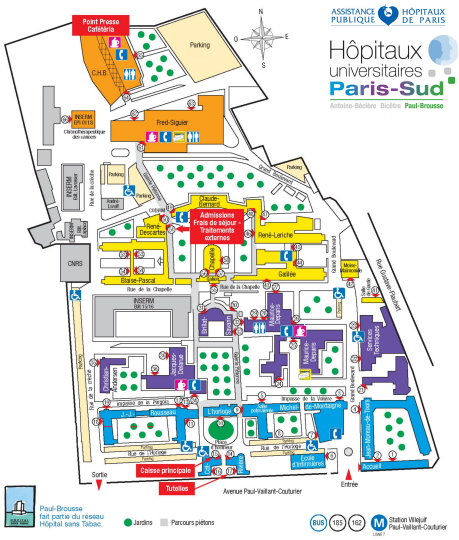


Figure 1:

Troisème journée

Résumé troisième journée:

Exemple de classification d'impact de balle

Découverte de la programmation avec le logiciel libre R

Rédaction de rapport à l'aide du language Rmakerdown

Introduction à la problèmatique de detection du nombre de groupes homogènes dans un nuage de points

détection des dimensions discriminates dans une image de points

Quatrième journée

Dernière journée

Introduction au problème de classification

Dans cette partie, nous simulons des impacts de balles provenant de deux tireurs différents. Nous présentons une procédure de détection des impacts de chaque tireurs. Cela revient à mettre une couleur noir ou rouge sur chaque impact où chaque couleur caractérise un tireur. Le bloc de code suivant propose de simuler 200 impacts sur un mur. Ces lignes de code ont été réalisées sous le logiciel libre R. Les principaux éléments de ce codes correspondent à

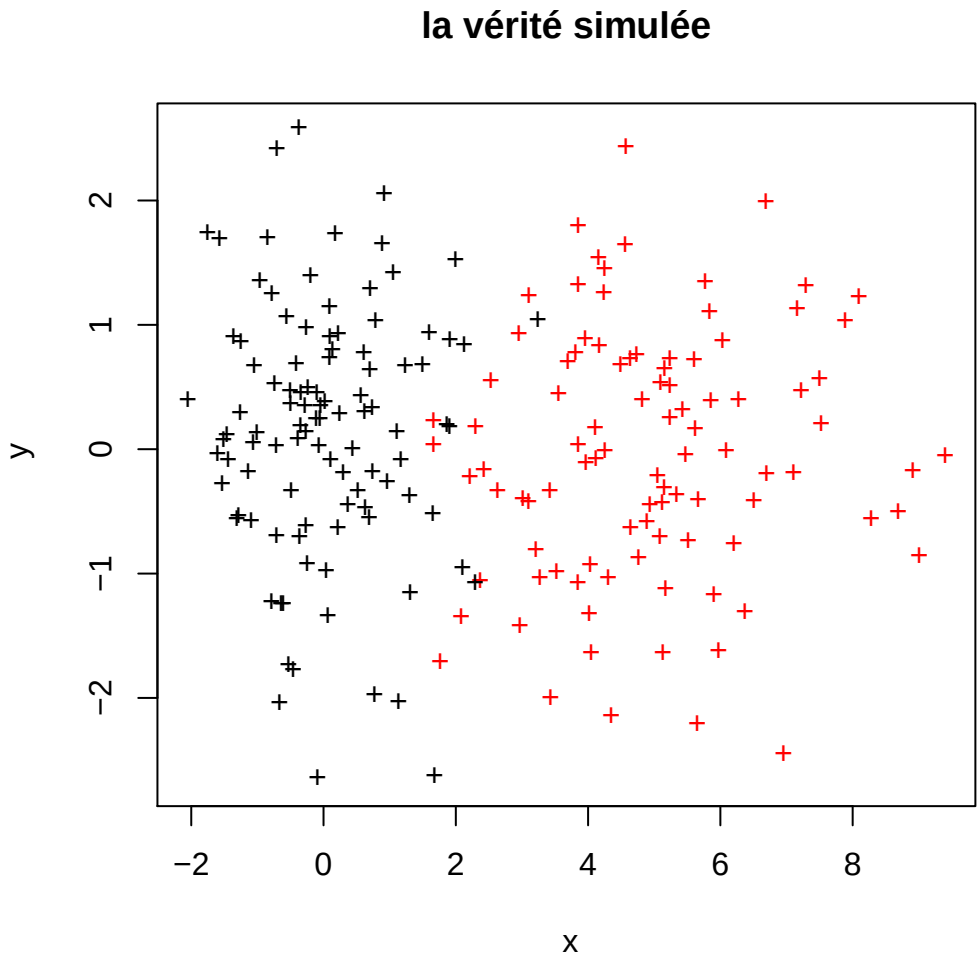
- l'instruction `set.seed(123)` permet de rendre l'expérience de simulation reproductible
- l'objet intitulé `rmixture` qui simule les impacts de balles où
  - `n` est l'agrument qui contrôle le nombre d'impacts à simuler
  - `delta` contrôle la différence entre les impacts des deux tireurs
  - `d.disc` est le nombe de dimensions discriminantes (une seule dans notre exemple)
  - `d.nondisc` est le nombe de dimensions non-discriminantes (une seule dans notre exemple)

Dans la suite, nous allons utiliser le terme *étiquette* pour désigner les couleurs rouge et noir qui permettent de distinguer les impacts d'un tireur de ceux de l'autre. On propose de cacher la vérité qui correspond

aux étiquettes rouges et noirs des impacts et nous allons faire appel à une procédure de détection appelée **algorithme des k-moyennes**.

```
set.seed(123)
# Fonction de generation des donnees
rmixture <- function(n, delta, d.disc, d.nondisc){
  z <- sample(1:2, n, replace = TRUE)
  x <- matrix(rnorm(n*(d.disc+d.nondisc)), n)
  x[which(z==2), 1:d.disc] <- x[which(z==2), 1:d.disc]*2 + delta
  list(x=as.data.frame(x), z=z)
}


ech <- rmixture(n=200, delta=5, d.disc = 1, d.nondisc = 1)
plot(ech$x[,1], ech$x[,2], pch="+", col=ech$z, xlab=~x, ylab=~y, main="la vérité simulée")
```



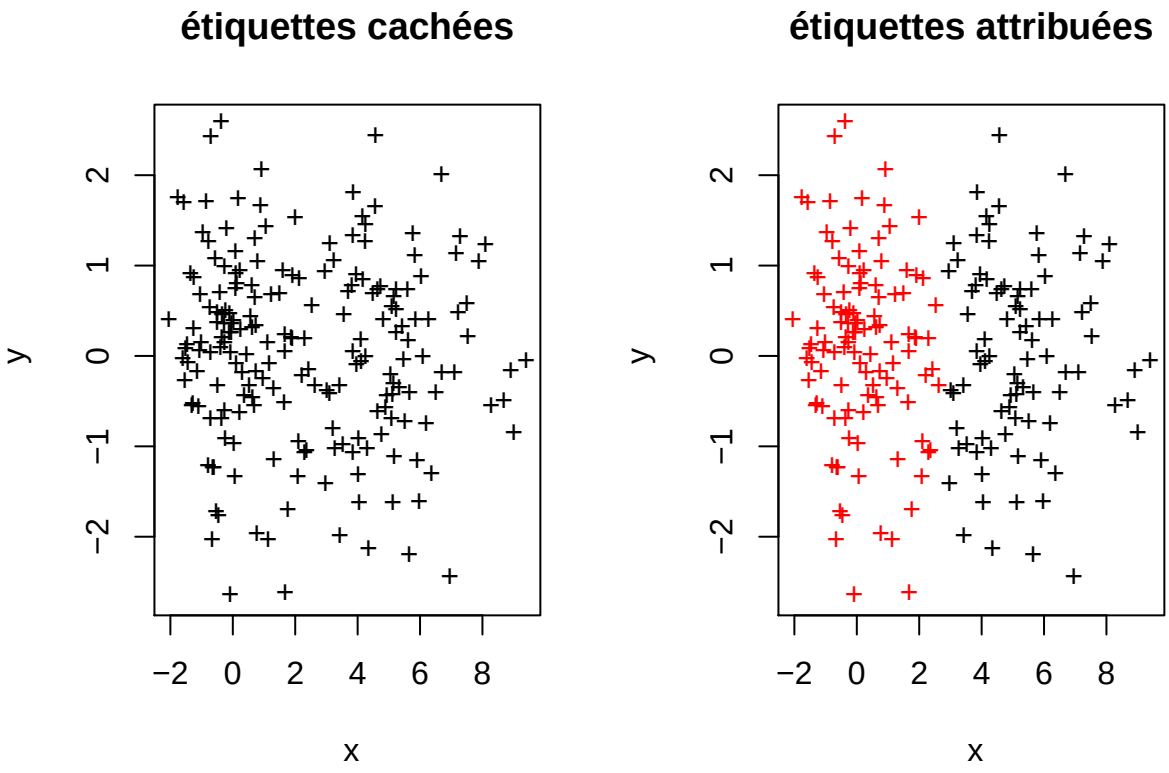
L'algorithme des k-moyennes peut être résumé par les 4 étapes suivantes :

1. On choisit deux points moyens au hasard, un rouge et un noir.
2. On affecte une étiquette rouges aux points les plus proches du point rouge et une étiquette noir aux points les plus proches du point noir;

3. On calcule les coordonnées moyennes des points rouges et les coordonnées moyennes des points noirs qui vont jouer le rôle des deux points moyens rouge et noir.
4. On répete l'etape 2 et 3 jusqu'à ce que les deux points moyens rouge et noir se stabilisent et ne bougent plus.

Le bloc de code suivant permet de mettre en place la procédure des kmoyennes sur les 200 impacts simulés précédemment

```
par(mfrow=c(1,2))
plot(ech$x[,1], ech$x[,2], pch="+", xlab=~x, ylab=~y, main="étiquettes cachées")
detect <- kmeans(ech$x, 2, nstart = 100)
plot(ech$x[,1], ech$x[,2], pch="+", col = detect$cluster,
     xlab=~x, ylab=~y, main="étiquettes attribuées")
```




```
min(mean(ech$z!=detect$cluster), 1-mean(ech$z!=detect$cluster))
```

```
## [1] 0.055
```

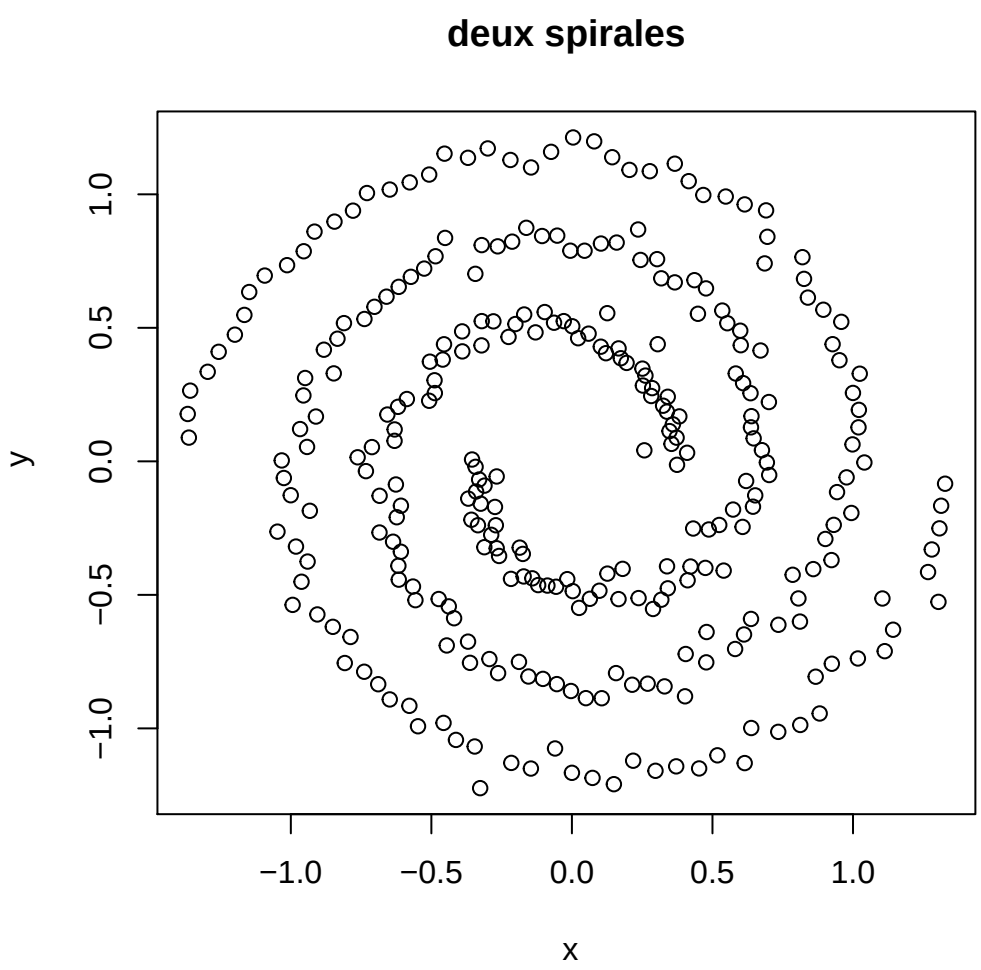
Comme on peut le constacter, l'affectation des couleurs des étiquettes se fait à une permutation près. L'algorithme des k-moyennes se trompes dans l'étiquetage de 5.5% des impacts.

On s'intéresse à un autre jeu de données sous forme de deux spirales qu'on cherche à détecter automatiquement.

```
require(kernlab)
```

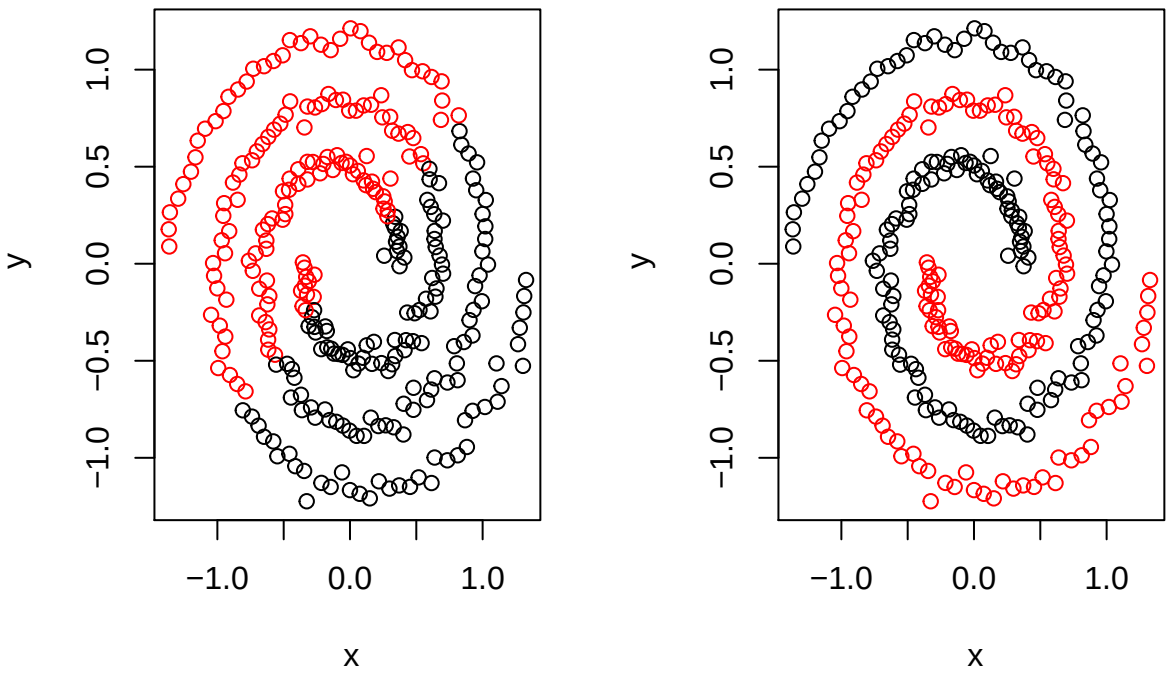
```
## Loading required package: kernlab
```

```
data(spirals)
plot(spirals, xlab=~x, ylab=~y, main="deux spirales")
```



Ce jeu de données est intéressant par la forme des deux groupes à détecter. Cette forme spirale des deux groupes met en échec l'algorithme des k-moyennes comme on peut le remarquer sur la figure à gauche ci-dessous. En raison de la forme non-convexe des spirales, l'agorithme des k-moyennes échoue dans la détection des deux groupes.

```
require(kernlab)
data(spirals)
par(mfrow=c(1,2))
plot(spirals, col = kmeans(spirals, 2)$cluster, xlab=~x, ylab=~y )
plot(spirals, col = specc(spirals, 2), xlab=~x, ylab=~y)
```



Revenons au premier exemple d'attribution d'étiquettes rouges et noirs aux impacts de tirs, nous avons donné en entrée de l'algorithme des k-moyennes, le nombre de groupes qui correspond au chiffre 2 dans l'instruction `kmeans(ech$x, 2, nstart = 100)`. Cette information est inconnu a priori dans une situation réelle. Il serait donc intéressant de mettre une place un mécanisme de choix du nombre de tireurs. Comme on peut le constater sur le même jeu de données, c'est uniquement l'abscisse d'un point qui permet de déterminer l'appartenance à une couleur ou une autre. La question de détection de dimensions discriminante est une question très importance quand il s'agît de classer des données génétiques où une dimension discriminante correspond à une position du génome qui permet de faire un diagnostic. La procédure implémentée dans la librairie VarSelLCM de R permet de répondre aux deux questions précédentes comme on peut le voir sur le bloc de code suivant. Ce libraire est le résultat des travaux des trois dernières années de recherche de mon tuteur en collaboration étroite avec son coauteur enseignant-chercheur à l'ENSAI à Rennes.

```
require(VarSelLCM)
```

```
## Loading required package: VarSelLCM
```

```
##
## Attaching package: 'VarSelLCM'
```

```
## The following object is masked from 'package:kernlab':
##
##     predict
```

```
## The following object is masked from 'package:stats':
##
##     predict
```

```
res_with <- VarSelCluster(ech$x, gvals=1:6, nbcores = 8, crit.varsel = "BIC")
print(res_with)
```

```
## Data set:
##    Number of individuals: 200
##    Number of continuous variables: 2
##
## Model:
##    Number of components: 2
##    Model selection has been performed according to the BIC  criterion
##    Variable selection has been performed, 1  ( 50 % ) of the variables are relevant for clustering
##
## Information Criteria:
##    loglike: -745.3824
##    AIC:     -752.3824
##    BIC:     -763.9265
##    ICL:     -781.7757
```

```
min(mean(ech$z!=res_with@partitions@zMAP), 1-mean(ech$z!=res_with@partitions@zMAP))
```

```
## [1] 0.065
```

On reprend l'exemple d'impacts de balles sur un mur avec un nouvel algorithme. L'argument `gvals=1:6` de la fonction `VarSelCluster` permet de tester les nombre de groupes (nombre de tireurs dans cet exemple) allant de 1 à 8. La commande `nbcores = 8` permet de faire appel aux 8 coeurs du processeur pour faire le calcul en un temps plus court. Il a detecté la présence de deux tireurs, il y'a une des deux dimensions qui est inutile pour détecter le nombre de tireurs et le taux d'erreur est de 6.5%.

Application des méthodes de classifications aux sciences du vivants

Classification de types de leucémies

Pour faciliter la comrpéhension du problème de classifaction et des deux questions du choix du nombre de groupes et la détection des dimensions discriminantes, nous nous sommes contentés d'exemples jouets de données (impacts de balles et spirales). En réalité les deux questions précédentes ont été soulevées des cadres variés comme le diagnostic à partir d'expression de gênes. Pour illustrer ce domaine d'application, nous allons faire appel à des données observées en très grande dimension. La table de données `golub` contient les valeurs d'expressions génétiques de 3051 gènes prélevés chez 38 patients atteints de leucémie. Vingt-sept patients ont reçu un diagnostic de leucémie lymphoblastique aiguë (ALL) et onze leucémie myéloïde aiguë (LMA). L'idée est réussir à classer les deux groupes de leucémies avec le taux d'erreur le plus bas et détecter les dimensions classifiantes qui serviront au diagnostic et à la compréhension de la maladie.

```
require(VarSelLCM)
data(golub, package = "multtest") # lecture des données
golub <- t(golub)
nb.CPU <- 8# utilisation de la capacité maximale de calcul de l'ordinateur
# détection des deux groupes
res.selec <- VarSelCluster(golub,
                           gvals = 2,
                           vbleSelec= TRUE,
                           crit.varsel = "MICL",
                           nbcores = nb.CPU)
print(res.selec)
```

```
## Data set:
##    Number of individuals: 38
##    Number of continuous variables: 3051
##
## Model:
##    Number of components: 2
```

```
##    Model selection has been performed according to the MICL  criterion
##    Variable selection has been performed, 553  ( 18.13 % ) of the variables are relevant for clusteri
##
## Information Criteria:
##    loglike: -77235.87
##    AIC:     -84444.87
##    BIC:     -90347.55
##    ICL:     -103858.8
##    MICL:    -103858.8
##    Best values has been found  8 times
```

```
print(golub.cl)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
## [36] 1 1 1
```

```
print(res.selec@partitions@zOPT-1)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 1 1 1 1 1 1
## [36] 1 1 1
```