TD noté

Etienne Birmelé

Ce TP est à rendre pour le **mercredi 8 janvier à minuit**. Il le sera sous la forme d'un fichier Markdown (en .Rmd et .pdf), ou tout autre solution R ou Python me permettant simultanément de faire tourner votre code et de lire aisément les résultats. Veillez à rendre un document lisible, ce qui veut dire ne pas forcément afficher toutes les sorties, seulement les parties qui méritent d'être lues.

On considère un jeu de données d'expression de gènes, disponible sur Kaggle à l'adresse

https://www.kaggle.com/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

On remarquera que faire défiler la fenêtre de description du jeu permet de voir à quoi correspondent les variables médicales (les trente premières) prises en compte. LE reste du jeu est composée de données d'expression de gènes et de mutations.

Préparation du jeu de données

Nous allons restreindre les données génomiques au 100 gènes les plus corrélées à un indice de gravité phénotypique (l'indice de Nottingham) utilisé pour décider d'opérer ou non, et supprimer les données de mutation.

```
library(tidyverse)

metabric.data <- read.csv2('./data/METABRIC_RNA_Mutation.csv',header=TRUE,sep=",",
                           row.names='patient_id')

# supprimer les données de mutation que nous n'allons pas utiliser
genes.data <- metabric.data[,1:519]


# Conversion des colonnes au bon type
genes.data <- type_convert(genes.data)

# Première diminution de dimension par selection des $100$ gènes
#les plus corrélées au score de gravité de Nottingham
corre <- c()
for (i in c(1:489)){
  corre <- c(corre,cor(genes.data[,21],genes.data[,i+30])**2)
}
selected <- order(corre,decreasing = TRUE)[1:100]
genes.data <- genes.data[,c(1:30,30+selected)]
```

ACP et k-means

1. Réaliser une ACP normée sur les données des 100 gènes considérées en gardant les variables médicales comme variables supplémentaires d'interprétation (les variables $1, 19, 20, 21, 23$ et 28 sont quantitatives, les autres sont qualitatives.

Vous justifierez le nombre de composantes gardées.

2. Appliquer l'algorithme des k-means sur les vecteurs projetés, pour un nombre de composantes allant de 1 à 30 et regarder l'évolution de la somme des carrés intra-classes. En utilisant un critère du coude, décider du nombre d'un nombre de classes à retenir. Appliquer l'algorithme pour ce nombre de classes et en déduire une classification des patientes.

3. Tracer le nuage de points dans le plan principal de l'ACP avec des couleurs correspondant aux classes de k-means. Que constatez-vous?

Interprétation des résultats

4. A l'aide de tests du chi-deux, trouver les variables médicales qualitatives auxquelles les classes trouvées sont les plus liées (ici, on comprendra lié comme éloigné de l'indépendance). Comparer les premières d'entre elles (par exemple les trois premières) avec la classification établie, et utilisant la fonction *table* qui permet d'établir des tables de contingence. Commenter les résultats en terme d'interprétation de la classification établie.

Prédiction de l'indice de Nottingham

5. Mettre en place une prédiction de l'indice de Nottingham à l'aide d'un modèle linéaire. Vous justifierez les choix faits.