TP1 Data Mining M2 FST

Les 5 premiers TPs que vous allez faire vont vous permettre de revoir la plupart des notions vues en analyse des données en L3. Cette partie du cours correspond à la partie aussi appelée apprentissage statistique non supervisé.

Dans cette première partie du cours et des tps, vous utiliserez R. Attention, n'oubliez pas de changer de répertoire en début de TP. Je vous invite à reprendre vos documents de L3. Si vous avez oublié, j'ai remis les textes des Tps de L3 et la short refcard sur le site FST. Vous pouvez aussi consulter

`http://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf`

Enfin rappelez-vous que la commande help(nom d'une commande) vous ouvre une fenêtre d'information sur la commande et qu'il est toujours plus prudent d'écrire les commandes dans un script. En les marquant puis en faisant CTRL R, elles sont exécutées dans la fenêtre de commande.

Première analyse : Analyse en composantes principales des iris de Fisher

Le fichier iris contient la description de 150 iris , 50 Setosa, 50 Virginica et 50 Versicolor. Sur chaque iris, on a mesuré la longueur et la largeur des pétales, la longueur et la largeur des sépales.

Pour charger le fichier iris, il suffit de taper attach(iris). En tapant iris, vous avez le contenu et le nom des variables.Comme les noms des variables sont assez longs, vous pouvez les renommer. Par exemple Selo=Sepal.Length mais rappelez-vous que iris[,1] vous donnera la même information.

Tri à plat des variables

- Faites un histogramme de chacune des variables quantitatives.(hist(nom de la variable)) Commentez vos résultats.
- Pour chacune des variables, faites un histogramme par espèce et mettez pour une variable donnée l'ensemble sur un même graphique.

- Calculez les caractéristiques des variables globalement et par espèce.(summary(nom de la variable))
- En utilisant la commande boxplot, comparez les variables par espèce :
  Par exemple pour la variable longueur de sépale,
  ```
  boxplot(Selo ~ iris[,5])
  ```
  vous dessine les boîtes à moustache par espèce (vous voyez déjà l'intérêt de renommer les variables). Si vous voulez mettre des couleurs ou des commentaires, faites help(boxplot) pour avoir les différentes options.
  Le tilde indique une relation entre deux variables, la première étant la variable à expliquer et la seconde la variable explicative.
  Commentez les résultats. Si vous deviez discriminer entre espèces (on fera ça plus tard) , quelle(s) variable(s) suggérez-vous d'utiliser ?

Relations entre les variables

- Représentez les nuages de points obtenus en prenant les variables quantitatives deux par deux.On appelle aussi cela diagramme de dispersion.
  La commande
  ```
  plot(iris[,1:4])
  ```
  vous permettra d'obtenir ce résultat pour les variables quantitatives.
  Pour visualiser les espèces (ie les modalités d'une variable qualitative), vous utiliserez :
  ```
  plot(iris[,1:4], col=as.numeric(iris[,5]))
  ```
  Vous pouvez mettre un titre, nommer l'axe des ordonnées, celui des abscisses. Faites help(plot) pour avoir les différentes possibilités.

  Vous pouvez calculer la matrice de corrélations avec la commande cor
  ```
  cor(iris[,1:4])
  pour ne conserver que 2 chiffres décimaux,
              faites round(cor(iris[,1:4]),2)
  ```
- Que pensez-vous des corrélations entre les variables ?A-t-on besoin de 4 variables ? Peut-on représenter les iris dans un espace de dimension réduit ? Pensez-vous qu'une analyse en composantes principales pourrait être intéressante ? Pourquoi ?

Petite ACP

Nous espérons pouvoir représenter correctement les iris dans un espace de dimension strictement inférieure à 4.

Vous devez charger la librairie Factominer.Pour cela, tapez la commande

library(FactoMineR) (Attention à respecter la casse : majuscule, minuscule)

L'ACP est obtenue en précisant qu'ici la variable 5 est une variable qualitative supplémentaire.

J'appelle le résultat res.pca. Vous pouvez lui donner le nom que vous voulez, res tout court si vous préférez ou resiris ... !

res.pca=PCA(iris,quali.sup=5)

La commande précédente exécute l'ACP et fournit le graphe des variables (cercle de corrélations lorsqu'on fait une ACP sur matrice de corrélation, ce qui est la valeur par défaut) et le graphe des individus avec les modalités des variables qualitatives supplémentaires.

- Que constatez-vous en regardant le graphe des variables ? A quelles variables est corrélé l'axe 1 ? l'axe 2 ?
- Que constatez-vous en examinant la projection des individus sur le premier plan factoriel ?

Tapez res.pca pour afficher les objets qui le constituent.

Si vous tapez res.pca$eig pour obtenir les valeurs propres (en anglais eigenvalues) de la matrice de corrélation, vous allez obtenir beaucoup de décimales. je vous suggère de taper plutôt

round(res.pca$eig,2) pour ne conserver que 2 chiffres après la virgule.

Si vous souhaitez un "histogramme" des valeurs propres, faites la commande suivante `barplot(res.pca$eig[,1],main="valeurs propres")`

Si vous souhaitez avoir les numéros des valeurs propres sous les barres, rajoutez
`,names.arg=paste("dim",1:nrow(res.pca$eig))` dans les arguments de la commande.

- Combien de valeurs propres allez-vous conserver ? Pourquoi ? Quel est le pourcentage d'inertie expliqué par le premier axe, le second, le premier plan factoriel ?
- La commande `res.pca$var` vous fournit les coordonnées des variables sur les axes factoriels, leurs cosinus carrés et leurs contributions. Si vous voulez tronquer à 2 décimales, vous devez faire 3 commandes (si vous souhaitez garder uniquement les informations sur les 2 premiers axes, vous prenez 1 :2 au lieu de 1 :4).

  ```
  round(res.pca$var$coord[,1:4],2)
  ```

```
round(res.pca$var$cos2[,1:4],2)
round(res.pca$var$contrib[,1:4],2)
```

– La commande `res.pca$ind` vous fournit les coordonnées des individus sur les axes factoriels, leurs cosinus carrés et leurs contributions.Vous avez aussi pour chaque individu sa distance à l'origine. En utilisant le même procédé qu'auparavant, vous obtiendrez uniquement 2 décimales.

Etude du nuage des individus actifs et des variables

Pour visualiser les groupes d'iris(variable 5 définie dans habillage) sur le plan factoriel, vous devez taper `plot.PCA(res.pca,choix="ind",habillage=5)`.

Si vous examinez les contributions des individus à l'inertie des axes, quelques uns ont une contribution supérieure à deux fois la moyenne (ie ici $1/150 * 100$). Notez-les puis trouvez leurs coordonnées et vérifiez leurs positions sur le plan factoriel.

Que constatez-vous sur le premier plan factoriel ? Pouvez-vous aussi vérifier la qualité de représentation des iris ?

Quelles sont les coordonnées des variables dans le premier plan principal ? Comment interpréter le premier axe, le second ?