Analyse Statistique Multidimensionnelle - TP

Flora Jay flora.jay@lri.fr

Les calculs, les graphiques et la programmation se feront avec le logiciel **R**.
La représentation graphique est très importante en statistique. Chaque graphe doit être accompagné au minimum d'un titre et de noms d'axes. Il doit toujours être expliqué par quelques phrases dans le texte (Qu'a-t-on tracé ? Qu'en déduire ?).

Génétique des populations et Classification supervisée

1. Chargez dans `dat` et `info` les objets se trouvant dans NAgenetic.txt et NAinfo.txt. Pensez à l'option header=T. Utilisez `names(..), colnames() rownames()` pour connaître le contenu des différents fichiers. Les individus dont nous connaissons les informations génétiques sont les mêmes que ceux du TP2 (voir la carte du TP2). Par contre les marqueurs génétiques ne sont pas exactement les mêmes.

2. **Analyse linéaire discriminante. LDA**
   Le facteur d'intérêt est ici la population d'appartenance des individus : `info$pop` . L'objectif est de prédire la population d'appartenance d'un individu à partir de ses informations génétiques.

   a) Réalisez une analyse linéaire discriminante en utilisant comme données d'apprentissage la moitié des individus. (Prendre 1 individu sur 2 par exemple)
   Utilisez la fonction `lda` et pensez à convertir la colonne `pop` en facteur.
   b) Prédisez maintenant les populations d'origine des individus qui n'ont **pas** servi à l'apprentissage. (fonction `predict`, attribut `class`)
   c) A l'aide de la fonction `table`, vous pouvez afficher la matrice de confusion pour ces individus.
   `table(vraies pop, pop prédites)`
   d) À partir de la matrice de confusion, calculez le **taux** d'individus mal classés (que l'on appellera **erreur de classification ou erreur de prédiction**).
   e) De manière analogue, calculez l'**erreur d'apprentissage**.

3. **Analyse linéaire discriminante précédée de l'analyse en composante principale. (PCA – LDA)**
   Comme dans le TP2, on aimerait utiliser moins de prédicteurs. On va donc réaliser une ACP puis utiliser certains axes de l'ACP comme prédicteurs pour l'analyse linéaire discriminante.

   a) Réalisez l'ACP de **toutes** les données génétiques (pensez à l'argument `scale=T`).
   b) Au lieu d'utiliser les marqueurs génétiques pour faire la LDA, utilisez les 2 premiers axes de l'ACP.
   Attention : tout comme dans la question 2., seuls les individus de `train` doivent servir à l'apprentissage.
   Calculez l'erreur d'apprentissage et l'erreur de classification.
   c) Pour un nombre d'axes `naxes` variant de 2 à 215 (de 5 en 5 par exemple), répétez l'étape 3b) en utilisant `naxes` pour la LDA, et calculez les erreurs d'apprentissage et de classification.
   Tracez les courbes d'erreur d'apprentissage et prédiction en fonction du nombre d'axes de l'ACP utilisés.
   d) Quelle est l'erreur de prédiction minimale ? Pour le modèle y correspondant, y a-t-il des populations plus difficiles à classifier que d'autres ?

4. Le facteur d'intérêt est maintenant la **région** des individus. On considère 3 régions « North America », « Central America » et « South America ».
   a) Construisez un vecteur indiquant la région de chaque individu.
   b) Refaites la **pca-lda** pour ce facteur. Comparez l'erreur de prédiction à celle obtenue dans le 3 pour la prédiction des populations. (On comparera les erreurs minimales).
   c) Conclure