TP A RENDRE AU PLUS TARD LE 21 JUIN

Armel YODE

2025-06-01

Contexte

Dans ce TP, vous travaillez sur un jeu de données simulé de très grande dimension : le nombre de variables (p) dépasse largement le nombre d'observations (n). Cette situation est fréquente en génomique, en neurosciences ou en finance.

Vous devrez appliquer plusieurs techniques de réduction de dimension, de sélection de variables, et de classification, puis analyser leur pertinence dans ce contexte.

Objectifs pédagogiques

Appliquer des méthodes adaptées aux données de grande dimension : ACP, Lasso, PCR, PLS Comparer des algorithmes de classification : régression logistique, SVM, forêts aléatoires, etc. Évaluer les performances par validation croisée Interpréter et justifier les résultats obtenus

Jeu de donnée

```
# Exemple : jeu de données simulé
set.seed(123)
n <- 72
p <- 1000
X <- matrix(rnorm(n * p), n, p)
colnames(X) <- paste0("V", 1:p)
y <- factor(sample(c("ALL", "AML"), n, replace = TRUE))
#
# Vérifier les dimensions
dim(X)
```

```
## [1]   72 1000
```

```
table(y)
```

```
## y
## ALL AML
##  34  38
```

Contexte

Vous travaillez sur un jeu de données simulé où le nombre de variables (p) dépasse largement le nombre d'observations (n), une situation fréquente en bioinformatique, neurosciences ou finance.

Votre mission est d'appliquer plusieurs techniques statistiques pour réduire la dimension, sélectionner les variables pertinentes, et classer les observations, tout en analysant les résultats.

Objectifs

- Appliquer des méthodes adaptées aux situations $p \gg n$: ACP, Lasso, PCR, PLS
- Mettre en œuvre des algorithmes de classification : régression logistique, SVM, forêts aléatoires, etc.
- Évaluer les performances par validation croisée
- Interpréter et justifier les résultats obtenus avec esprit critique

1. Chargement et exploration du jeu de données

Chaque groupe doit générer un jeu de données unique à partir d'un `set.seed()` fourni séparément.

À faire :

- Résumer les dimensions du jeu de données
- Identifier la répartition des classes
- Vérifier la normalité et l'échelle des variables

2. Analyse en Composantes Principales (ACP)

À faire :

- Réaliser une ACP sur les données centrées et réduites
- Représenter les deux premières composantes
- Identifier si les classes sont séparables visuellement

Attention : l'interprétation du graphique est obligatoire. Ne vous contentez pas de dire "les classes sont bien séparées".

3. Régression logistique Lasso

À faire :

- Effectuer une régression logistique pénalisée (Lasso)
- Identifier les variables sélectionnées
- Interpréter l'effet de la régularisation dans ce contexte

Piège pédagogique : la sélection dépend du choix de pénalisation. Justifiez vos choix et vos observations.

4. Régression sur composantes principales (PCR)

À faire :

- Réaliser une régression sur les composantes principales
- Comparer les résultats à ceux du Lasso
- Discuter la différence d'approche

Question guidée : que perd-on en passant des variables initiales aux composantes principales ?

5. Régression PLS

À faire :

- Appliquer une régression PLS discriminante
- Comparer ses performances avec PCR et Lasso
- Interpréter les résultats

6. Comparaison de méthodes de classification

À faire :

Tester au moins trois des méthodes suivantes :

- Régression logistique
- SVM (linéaire ou RBF)
- Forêt aléatoire
- Naive Bayes
- Gradient boosting

Pour chaque méthode :

- Décrire brièvement son principe
- Appliquer la méthode avec validation croisée
- Évaluer les performances (accuracy, sensibilité, etc.)

  Les résultats attendus doivent être numériques et justifiés à partir de **votre propre jeu de données**.

7. Synthèse comparative

À faire :

- Présenter un **tableau comparatif** des résultats
- Discuter la pertinence de chaque méthode dans le contexte $p \gg n$
- Identifier les limites rencontrées et proposer des pistes d'amélioration

Livrable attendu

**Un rapport structuré (.pdf ou .html) contenant :

- Vos résultats chiffrés
- Vos **interprétations personnelles**
- Vos comparaisons justifiées
- Une annexe avec le code complet et commenté**

à deposer à mon secretariat; la version electronique (.Rmd avec .pdf ou .htlm) est à envoyer à yafevrard@yahoo.fr

Dispositifs anti-plagiat et anti-IA

- Données personnalisées par groupe
- Interprétations spécifiques exigées
- Analyse obligatoire des résultats propres
- Code brut non commenté = pénalité