Validation croisée et bootstrap
TP 2

B. Delyon, V. Monbet

Master 1 - 2013-2014

Exercice 1 - Krigeage - Températures en France

Dans le TP précédent, nous avons construit un modèle linéaire de prédiction de la température en France. Pour construire ce modèle de prédiction, nous avons supposé que les données étaient issues d'un vecteur gaussien tel que les dépendances entre deux composantes ne dépendent que de la distance géographique entre les composantes. Cette hypothèse est vraisemblablement un peu forte, et il est utile de faire une validation plus approfondie du modèle.

Validation croisée

Pour répondre aux questions ci-dessous, vous pouvez utiliser les programmes du fichier `Pour_validation_croisee.R`.

1. Proposer une méthode de validation croisée pour estimer l'erreur en moyenne quadratique des prédictions.
   Quand on constitue l'ensemble de validation, il faut éviter d'y mettre trop de points de la même région. En effet, on n'aurait plus de points prédicateur dans la région en question.
   Pour visualiser la position des points de mesure, on peut tracer la figure suivante.

   ```
   Coast <- read.table("coast.dat",header=TRUE)
   plot(D$lon,D$lat,pch=20,col="white")
   text(D$lon,D$lat,1:n)
   lines(Coast$x,Coast$y)
   ```

2. La validation croisée peut aussi être utilisée pour estimer le paramètre de la fonction de covariance. Mettre en oeuvre cette méthode.

3. Conclure.

Bootstrap

1. Quels sont les paramètres du modèle utilisés pour le krigeage?

2. Proposer une méthode de bootstrap paramétrique, pour estimer la distribution des estimateurs des paramètres. En déduire le biais et la variance des estimateurs.

3. Proposer une méthode de bootstrap paramétrique, pour estimer la distribution des erreurs de prédictions pour 10 points représentatifs en France. Comparer la variance de krigeage obtenue au TP1 avec cette distribution. Commenter le résultat.

4. Proposer une méthode permettant de simuler des températures en France conditionnellement aux températures observées. Comparer la variance de ces simulations avec la variance de krigeage obtenue au TP1. Commenter le résultat.

Exercice 2 - Classification par plus proches voisins

Il y a quelques années, le traitement du cancer de la prostate dépendait de son extension ou non au niveau des ganglions du système lymphatique. Afin d'éviter une intervention chirurgicale (laparotomie) pour vérifier la contamination, des études ont tenté de la prévoir à partir de l'observation de variables explicatives. Dans ce but, 5 variables ont été observées sur 53 patients atteints d'un cancer de la prostate et sur lesquels une laparotomie a été réalisée afin de s'assurer de l'implication ou non du système lymphatique. Ces données sont extraites de Collet (1991), elles sont disponibles sur la page web du cours. Les variables considérées sont les suivantes :

- `age` du patient
- `acid` niveau de "serum acid phosphatase",
- `radio` résultat d'une analyse radiographique (0 : négatif, 1 : positif),
- `taille` taille de la tumeur (0 : petite, 1 : grande),
- `gravite` résultat de la biopsie (0 : moins sérieux, 1 : sérieux).
- `lymph` indique l'implication (1) ou non (0) du système lymphatique.

L'objectif est donc prédictif (variable lymph). On va utiliser une méthode de plus proches voisins et plus particulièrement la fonction `knn` du package `class`. Sa syntaxe est la suivante

```
res = knn(train, test, cl, k = )
```

avec `train` un ensemble d'apprentissage, `test` un ensemble de test, `cl` la variable de classe, `k` le nombre de voisins. La distance utilisée est la distance euclidienne.

Proposer une méthode de validation croisée pour sélectionner le nombre de voisins qui conduit à la plus faible erreur de classement.