TP5: modèles GAM

Yannig Goude: yannig.goude@edf.fr

M2 "Statistique et Machine Learning"

Exercice 1: ridge regression

1. charger les données `data_ridge0.txt` et `data_ridge1.txt`, les stocker dans des `dataframes` que l'on appelera `data0` et `data1`. Effectuer une analyse descriptive des données. Quelle est la particularité de la matrice de design X -la matrice composée des vecteurs X_1, X_2,...,X_{30}-?

2. en utilisant les données `data0`, effectuer une régression linéaire de Y sur X. Utiliser la fonction `lm` de R. Effectuer la même régression linéaire à l'aide des formules matricielles vues en cours -utiliser les fonctions `solve` et `crossprod` de R-. Comparer les résultats.

3. effectuer la prévision de `data1$Y` et calculer le risque quadratique empirique associé. Représenter successivement:

   - les valeurs de `data1$Y` et les prévisions associés
   - les résidus de la prévision
   - les prévisions en fonction de `data1$Y`, comparer avec la première bissectrice

4. tirer au hasard 10% des individus de `data0` et effectuer la régression linéaire précédente. Itérer l'opération $K = 100$ fois. Comparer les résultats obtenus. Que dire?

5. à l'aide des formules matricielles vues en cours, programmer la régression ridge de Y sur X -on introduira pour cela un nouveau paramètre $\lambda > 0$-.

6. reprendre la question 4 précédente en remplaçant la régression linéaire par la régression ridge avec $\lambda = 20$. Comparer les résultats obtenus.

7. calculer, pour $\lambda = 0, 1, 2, ...100$, l'estimateur des degrés de liberté. Le représenter graphiquement en fonction de λ.

8. calculer, pour $\lambda = 0, 1, 2, ...100$, le vecteur de coefficient de la régression ridge β_{ridge}, représenter sa norme en fonction des degrés de liberté estimés.

9. en utilisant la validation croisée, choisir la valeur de λ optimale pour la prévision.

10. en utilisant la validation croisée généralisée, choisir la valeur de λ optimale pour la prévision. Comparer les temps de calcul de ces deux méthodes -utiliser la fonction `proc.time()` de R-.

11. comparer ces deux méthodes à l'erreur de prévision sur les données `data1`. Représenter sur un même graphique les erreur de VC, GCV et l'erreur de prévision en fonction des degrés de liberté estimés. Commenter.

Exercice 2: choix de la dimension de la base

Importer les données `data_conso_hebdo0.txt` déjà étudiées dans le TP1.

1. estimer le modèles $Y_i = f(T_i) + \varepsilon_i$ en utilisant la fonction gam du package mgcv, choisir des bases de spline cubiques

2. déterminer, en exploitant les sorties de votre modèle (edf, tests), la dimension optimale de la base

3. représenter, pour cette dimension, l'effet estimé. superposer les données observées. Etes vous satisfaits? (utiliser le package mgcViz pour les diagnostiques).

4. proposer un script permettant d'optimiser le choix de la dimension de la base en se basant sur la validation croisée par blocs

5. comparer, à l'aide de ce script, les performances d'ajustement en fonction de la dimension de la base pour plusieurs types de splines (par ex.: cr, ps, cs, ad)

6. représenter les bases de splines en fonction de T (choisir la dimension optimale au sens de l'erreur quadratique estimée par VC par bloc).

7. représenter sur un même graphique les effets estimés obtenus pour chacune des bases de splines (choisir la dimension optimale au sens de l'erreur quadratique estimée par VC par bloc).

8. refaire l'expérimentation en considérant le modèle suivant: $Y_i = f_1(T_i) + f_2(NumWeek_i) + f_3(Time_i) + \varepsilon_i$, choisir une base de thin plate splines et $k = 20$ pour f_2, $k = 3$ pour f_3.

Exercice 3: estimation d'un modèle GAM

Proposer un modèle GAM permettant d'obtenir la meilleur erreur de prévision sur l'échantillon test (data1).