TP4 (a): Régression Ridge

Yannig Goude: yannig.goude@edf.fr

M2 statML/MDA

Exercice 1

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

1. charger les données `longley` par la commande `data(longley)`.

2. séparer les données en un échantillon test comprenant les $k = 6$ dernières valeurs observées et un jeu d'apprentissage.

3. reprendre la démarche de l'exercice 1 et proposer un modèle de prévision.

4. mesurer les performances de votre prévision sur l'échantillon test (RMSE et MAPE).

5. refaire le même exercice avec $k = 9$, puis $k = 11$ que se passe t-il pour $k = 11$?