1 Normaliser et standardiser les données

On dispose d'un échantillon $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N} \subset \mathbb{R}^{n+1}$.

1. Comment faire pour que les données soient centrées ? Proposez un algorithme pour réaliser cette normalisation.
2. Comment faire pour que les données correspondant à chaque attribut aient en plus, une variance égale à 1 ? Proposez un algorithme pour réaliser cette standardisation.
3. Comment ensuite utiliser sur de nouvelles données les fonctions induites sur les données transformées ?

2 Protocole pour déterminer l'hyperparamètre de la régression Ridge par validation croisée.

On dispose d'un échantillon $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N} \subset \mathbb{R}^{n+1}$ et l'on souhaite réalise une régression Ridge de façon à expliquer la dernière coordonnée en fonction des n premières. On recherche donc les paramètres $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ pour lesquels

$$\sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 + C||\boldsymbol{\alpha}||^2$$

est minimum. On envisage de rechercher C dans l'ensemble $\{i * 0.1 | 1 \leq i \leq 100\}$.

1. Expliquez le protocole algorithmique permettant de calculer la valeur optimale de C par validation croisée.
2. Supposez que l'on trouve un résultat optimal pour $C_{opt} = 9.8$. Que convient-il de faire ?

3 Résolution de la régression Ridge

On dispose d'un échantillon $S = \{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,N} \subset \mathbb{R}^{n+1}$ et l'on souhaite trouver les paramètres $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ pour lesquels

$$\sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 + C||\boldsymbol{\alpha}||^2$$

est minimum.

1. Montrez que cela revient à résoudre une régression linéaire multiple ordinaire après avoir rajouté n nouveaux points à l'échantillon S.
2. En déduire que la solution du problème est donnée par

$$\boldsymbol{\alpha} = (\mathbf{X}^\top \mathbf{X} + C\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

   où $\mathbf{I}$ est la matrice identité de dimension n.

4 Variation de $||\alpha_C^{ridge}||$ en fonction de C

On note α_C^{ridge} la solution de la régression Ridge pour la constante C.

1. Que vaut α_C^{ridge} lorsque $C = 0$? lorsque $C \to \infty$?
2. Montrez que si $C \leq C'$, alors $||\alpha_C^{ridge}|| \geq ||\alpha_{C'}^{ridge}||$.
3. Qu'en est-il pour le paramètre de la régression Lasso ?

5 Pénalisation vs régularisation

On peut présenter la régression Ridge sous deux formes équivalentes :

$$\alpha^{ridge} = \underset{\boldsymbol{\alpha}}{ArgMin} \left[\sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 + C||\boldsymbol{\alpha}||^2 \right]$$

ou

$$\alpha^{ridge} = \underset{\boldsymbol{\alpha}}{ArgMin} \sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 \text{ sous la contrainte } ||\boldsymbol{\alpha}||^2 \leq s.$$

Ces deux problèmes étant convexes, ils admettent une solution unique. Notons $\boldsymbol{\alpha}_{1,C}$ (resp. $\boldsymbol{\alpha}_{2,s}$) la solution du premier (resp. du second) problème d'optimisation.

1. Montrez que $\boldsymbol{\alpha}_{1,0} = \boldsymbol{\alpha}_{2,\infty}$ et que $\boldsymbol{\alpha}_{1,\infty} = \boldsymbol{\alpha}_{2,0} = 0$.
2. Pour $C \in [0, \infty[$, soit $s_C = ||\boldsymbol{\alpha}_{1,c}||^2$. Montrez que $\boldsymbol{\alpha}_{1,C} = \boldsymbol{\alpha}_{2,s_C}$.
3. Montrez que si $s \geq ||\boldsymbol{\alpha}_{1,0}||^2$, $\boldsymbol{\alpha}_{2,s} = \boldsymbol{\alpha}_{2,\infty}$.
4. Qu'en est-il pour la régression Lasso ?