Université Paris-Nanterre
UFR SEGMI

Statistiques pour Big Data

Documents et calculatrice interdits. La plus grande importance sera accordée lors de la correction à la justification des réponses. Les exercices sont indépendants. Durée 2h.

Questions de cours

1. Présentez le principe de cross-validation et les différentes méthodes.

2. Décomposer l'erreur de prédiction en fonction du biais et de la variance.

Exercice 1

1. ($1/2$ point) La régression Ridge (ou Lasso) est en général utilisée lorsque l'hypothèse ci-dessous n'est pas satisfaite :

   1. H1 concernant le rang de X
   2. H2 concernant l'espérance et la variance des résidus
   3. H3 concernant la normalité des résidus.

2. ($1/2$ point) La régression pénalisée peut être vue comme une régression avec comme critère d'estimation la somme du carré des résidus et une contrainte sur :

   1. le plan de (X) X
   2. les paramètres
   3. Il n'y a pas de lien

Exercice 2

Toutes les variables sont centrées et réduites. Dans la régression multiple sur p variable explicatives, le nombre de coefficients inconnus $\{\beta_j\}$ est p, c'est-à-dire $tr(P_X)$ où P_X est l'application qui à Y fait correspondre $\hat{Y}$. La trace de cette application donne le nombre effectif de paramètres. Cette notion peut être étendue à la régression ridge.

1. (2 points) Dans le cas de la régression ridge, donnez l'expression de $\hat{\beta}_{ridge}$

2. (2 points) Donnez l'expression de $\hat{Y}$ (ou encore P_x).

3. (2 points) En utilisant la décomposition en valeurs singulière de X : $X = UDV'$ avec U et V matrices orthogonales et $D = \text{diag}(d_1, ..., d_p)$, $\hat{Y} = (UD(D^2 + \lambda I)^{-1}DU')$.

4. (1 point) En déduire que le nombre effectif de paramètres de la régression ridge est $\sum_{i=1}^{p} \frac{d_i^2}{d_i^2+\lambda}$

Exercice 3

Soit un modèle de régression $Y = X\beta + \varepsilon$ pour lequel nous nous intéressons à la régression ridge. Les variables sont déjà centrées-réduites. Nous allons considérer que λ est fixé et $\varepsilon \sim N(0, \sigma^2 I_n)$. De plus, $X\beta_{ridge} \neq P_X Y$ et la régression ridge est utile.

1. Dans le cadre de la régression par MCO pour $Y = X\beta + \varepsilon$, rappeler la loi de $\hat{\beta}$.
2. Rappeler l'expression de l'estimateur $\hat{\beta}_{\text{ridge}}$. et trouver sa loi.
3. D'après l'énoncé, pourquoi $\hat{Y}^{MCO} = P_X Y \neq \hat{Y}^{\text{ridge}}$? Comparer alors $Y - \hat{Y}^{\text{ridge}}$ et $Y - \hat{Y}^{MCO}$. Sont-ils colinéaires ? Conclure sur l'orthogonalité entre $\hat{Y}^{MCO}$ et $Y - \hat{Y}^{\text{ridge}}$.
4. Soit l'estimateur de σ^2 issu de la régression par MCO : $\hat{\sigma}^2 = \frac{||Y - \hat{Y}||^2}{n-p} = \frac{\hat{\varepsilon}^2}{n-p}$. Montrez que $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants. Pour cela exprimer $\hat{\beta}$ en fonction de la matrix de projection orthogonal P_X avec $P_X Y = X\beta$ et $\hat{\varepsilon}$ en fonction de $I_n - P_X$.
5. Soit l'estimateur de σ^2 issu de la régression ridge : $\hat{\sigma}^2_{\text{ridge}} = \frac{||Y - \hat{Y}^{\text{ridge}}||^2}{n - Tr(X(X'X + \lambda I_n)^{-1} X')}$. Peut-on aussi montrer que $\hat{\sigma}^2_{\text{ridge}}$ et $\hat{\beta}^{\text{ridge}}$ sont indépendants ? Pour cela, vous montrerez que $\hat{\beta}^{\text{ridge}}$ est fonction de $P_X Y$ et vous ferez le lien avec la question 4.
6. Quelle conséquence à votre réponse précédente sur les intervalles de confiance de l'estimateur ridge ?

Exercice 4

On considère ici un problème de classification binaire $Y = \{-1, +1\}$ de données dans un espace de description $X \in \mathbb{R}^d$. On note $\{(x_i, y_i) \in (X, Y)\}$, $i \in \{1, ..., n\}$ l'ensemble d'apprentissage considéré. La fonction de décision du classifieur considéré est donnée par : $f(\beta, \beta_0) = \text{sign}(\beta' x + \beta_0)$.On considère dans un premier temps un ensemble de données linéairement séparable. Cet ensemble de données et la frontière de décision sont représentés (en sur la figure 1.

1. Sur cette figure, l'échantillon x_i et de label y_i est représenté par le point A. On s'intéresse à sa distance signée γ^i à la frontière de decision dont le point le plus proche est représenté par B. Sachant que $\frac{\beta}{||\beta||}$ est un vecteur unitaire othogonal à la frontière de décision, donner l'expression de ^{i} en fonction de x_i, y_i, β et β_0. Que cela implique-t-il si l'on souhaite éloigner au maximum les points de la frontière de décision ?
2. On considère alors le problème d'optimisation sous contraintes suivant :

   $$\min_{\beta_0, \beta} \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{N} \xi_i$$
   $$\text{sc } y_i(x_i'\beta + \beta_0) \geq 1, \, \forall i$$

   Poser le Lagrangien à considérer pour optimiser ce problème sous contraintes
3. Donner la solution analytique de la minimisation de ce Lagrangien par rapport à β et β_0.
4. En déduire une nouvelle formulation "duale" de notre problème d'optimisation sous contraintes

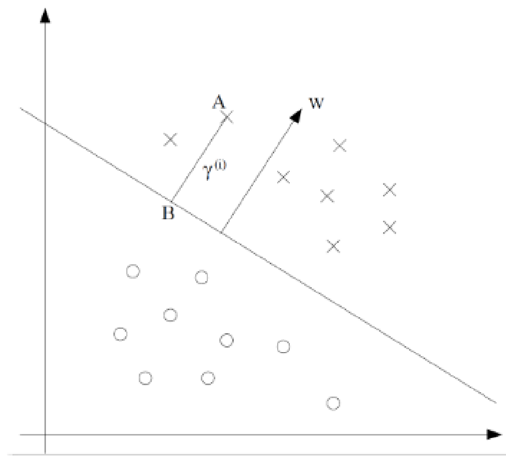

FIG. 1 : Ensemble de données

5. Quel est le problème du problème d'optimisation que l'on a considéré ? Proposer une nouvelle formulation qui corrige ce problème

Exercice 5

Soit des données provenant du Survey *British Social Attitudes*. Notre objectif est de prédire la variable *imm_brit*. Cette variable est comprise entre 0 et 100 et représente la proportion d'immigrant perçu au Royaume-Uni par le répondant :

imm_brit = Sur 100 personnes, selon vous combien sont issues de l'immigration hors pays occidentaux ? Nous utiliserons les inputs :

- resp_female : Est-ce que le répondant est une femme ?
- resp_age [RAge] : Age du répondant
- resp_household_size : De combien de personne est composé la foyer du répondant ?
- resp_party_cons : Est-ce que le répondant soutient le parti conservateur ?
- resp_party_lab : Est-ce que le répondant soutient le labor party ?
- resp_party_libdem : Est-ce que le répondant soutient le parti libéral démocrate ?
- resp_party_snp : Est-ce que le répondant soutient le Scottish National Party
- resp_party_green : Est-ce que le répondant soutient le Green Party
- resp_party_ukip : Est-ce que le répondant soutient le Respondent le UK Independence Party
- resp_party_bnp : Est-ce que le répondant soutient le British National Party
- resp_party_other : Est-ce que le répondant soutient un autre parti ou ne se pronnonce pas
- resp_newspaper : Le répondant lit les quotidiens

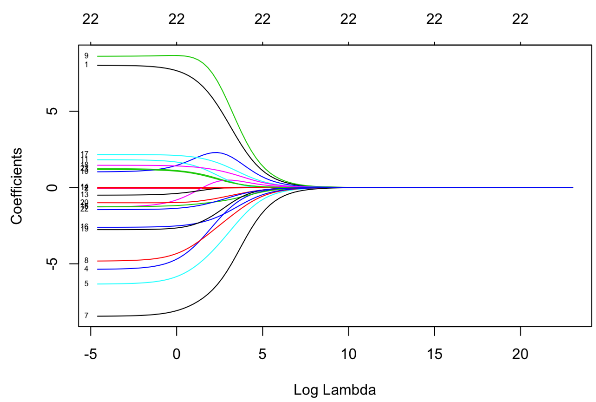

FIG. 2 : Estimation des coefficients d'une régression ridge pour plusieurs valeurs de λ

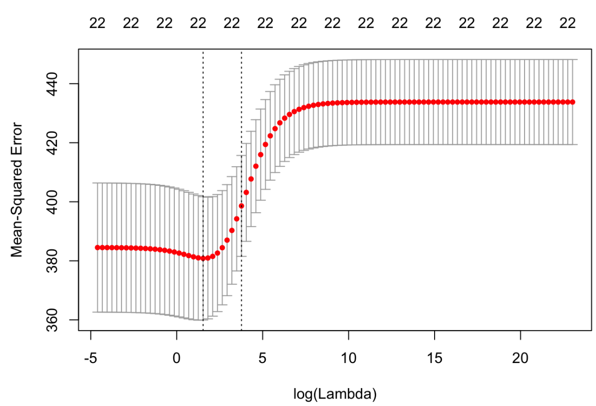

FIG. 3 : MSE pour plusieurs λ

- resp_internet_hrs : Le nombre d'heures passées sur internet par semaine
- resp_religious : Le répondant pratique une religion
- resp_time_current_employment : Mois d'ancienneté dans sont travail actuel resp_urban_area : Densité de la population
- resp_health : Etat de santé du répondant
- resp_household_income : Revenu sur foyer du répondant

1. Pourquoi la régression pénalisée vous semble adaptée dans ce problème ?
2. On estime le modèle ridge pour plusieurs valeurs de l'hyperparamètre λ et on obtient le graphique suivant (figure 2) : Est-ce que la régression Ridge vous semble utile ?
3. La figure 3 représente la MSE du modèle pour les différents λ. Quelle valeur (environ) est optimale ?
4. On obtient les résultats suivants pour les modèles ridge et lasso, que pouvez vous commenter ?

	Ridge	Lasso
cste	37.03	36.71
resp_female	5.94	5.34
resp_age	-0.04	-0.1
resp_household_size	1.15	0.93
resp_party_lab	-2.58	-0.58
resp_party_libdem	-3.90	-1.49
resp_party_snp	3.82	0.00
resp_party_green	-3.34	0.00
resp_party_ukip	-4.15	-0.00
resp_party_bnp	8.82	5.82
resp_party_other	2.73	2.56
resp_newspaper	1.61	0.01
resp_internet_hrs	-0.03	0.00
resp_religious	0.51	0.00
resp_time_current_employment	-0.01	0.00
resp_urban_area_rural	-1.22	0.00
resp_urban_area_rather_rural	-0.86	0.00
resp_urban_area_rather_urban	0.29	0.00
resp_urban_area_urban	1.66	0.15
resp_healthfair	-1.21	0.00
resp_healthfairly good	-0.24	0.00
resp_healthgood	-0.10	0.00
resp_household_income	-1.20	-1.28
MSE	386.4369	386.9932

Tab. 1 : Résultats de l'estimation ridge et lasso