Département MIDO
Module ML/DM
2015 / 2016

Travaux Dirigés n° 3 : Régression

Objectifs : revoir les modèles de régression vu en cours : linéaire simple en calculant un intervalle de confiance, multiple, ridge, pratique sous R de logistique+ridge et logistique+lasso.

1 Régression simple

1.1 Exercice 1

Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elles subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

1. Représenter le nuage de points. Déterminer la droite de régression. Calculer le coefficient de corrélation empirique. Commenter.
2. Deux stagiaires semblent se distinguer des autres. Les supprimer et déterminer la droite de régression sur les dix points restants. Calculer le coefficient de corrélation empirique. Commenter.

1.2 Exercice 2

On considère le modèle de régression linéaire simple $\mathbf{y} = \beta_1 + \beta_2 \mathbf{x} + \epsilon$. Soit un échantillon $(x_i, y_i)_{1 \leq i \leq 100}$ de statistiques résumées

$$\sum_{i=1}^{100} x_i = 0, \sum_{i=1}^{100} x_i^2 = 400, \sum_{i=1}^{100} x_i y_i = 100, \sum_{i=1}^{100} y_i = 100, \hat{\sigma}^2 = 1.$$

— Exprimer les intervalles de confiance à 95% pour β_1 et β_2 en vous servant de la table des quantiles de la loi de Student.

2 Regression ridge

2.1 Exercice 1

Trouver la solution donnée par la "régression ridge" aux données qui suivent, pour une valeur quelconque de λ et pour le modèle linéaire simple $\mathbf{y} = \beta_1 + \beta_2 \mathbf{x} + \epsilon$ (appliquer la régularisation ridge à la pente et non à l'intercept). Montrez que quand $\lambda = 4$, la solution ridge s'écrit : $\hat{\mathbf{y}} = 40 + 1.75\mathbf{x}$.
Données : $\mathbf{x}^T = (x_1, x_2, \cdots, x_8)^T = (-2, -1, -1, -1, 0, 1, 2, 2)^T$, et $\mathbf{y}^T = (y_1, y_2, \cdots, y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

n / q	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.9995
1	1	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.061	1.386	1.886	2.92	4.303	6.965	9.925	31.599
3	0.765	0.978	1.25	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.19	1.533	2.132	2.776	3.747	4.604	8.61
5	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.44	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.86	2.306	2.896	3.355	5.041
9	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	4.781
10	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.14
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.69	0.865	1.071	1.337	1.746	2.12	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.74	2.11	2.567	2.898	3.965
18	0.688	0.862	1.067	1.33	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.86	1.064	1.325	1.725	2.086	2.528	2.845	3.85
21	0.686	0.859	1.063	1.323	1.721	2.08	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.06	1.319	1.714	2.069	2.5	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.06	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.69
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.31	1.697	2.042	2.457	2.75	3.646
40	0.681	0.851	1.05	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	0.679	0.848	1.045	1.296	1.671	2	2.39	2.66	3.46
80	0.678	0.846	1.043	1.292	1.664	1.99	2.374	2.639	3.416
100	0.677	0.845	1.042	1.29	1.66	1.984	2.364	2.626	3.39
120	0.677	0.845	1.041	1.289	1.658	1.98	2.358	2.617	3.373

2.2 Exercice 2

Les coefficients β du modèle de régression linéaire, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, sont estimés par $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$. Nous avons donc les valeurs prédites sui s'écrivent $\hat{\mathbf{y}} = X\hat{\beta} = X(X^T X)^{-1} X^T \mathbf{y} = P\mathbf{y}$, où $P = X(X^T X)^{-1} X^T$

1. Montrez que P est une matrice de projection, i.e. $P^2 = P$.
2. Que pouvez vous déduire sur les vecteurs $\mathbf{y}, \hat{\mathbf{y}}, \hat{\varepsilon}$.

Soit maintenant, l'estimateur ridge des coefficients de régression : $\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$. Notons $\hat{\mathbf{y}}(\lambda) = X\hat{\beta}(\lambda)$ le vecteur associé aux valeurs prédites.

1. Montrez que la matrice $Q = X(X^T X + \lambda I)^{-1} X^T$, associé à la régression ridge n'est pas une matrice de projection (pour toute valeur de $\lambda > 0$).
2. Montrer que le "prédicteur ridge" $\hat{\mathbf{y}}(\lambda)$ n'est pas orthogonal au "résidu ridge" $\hat{\varepsilon}\lambda$ (pour toute valeur de $\lambda > 0$)..

3 Pratique sous R

3.1 Regression logistique + ridge

Télécharger le package `multtest` à partir de BioConductor :

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("multtest")
```

Activez la librairie et chargez les données `leukemia` à partir du package :

```
> library(multtest)
> data(golub)
```

Les objets `golub` et `golub.cl` sont désormais disponibles. L'objet matrice (matrix-object) `golub` contient les expressions des profils de 38 patients atteints de leucémie. Chaque profil comporte les niveaux d'expressions de 3051 gênes. L'objet numérique (numeric-object) `golub.cl` est variable indicatrice encodant le type de leucémie (AML ou ALL) de chaque patient.

1. Associer le sous-type de leucémie avec les niveaux d'expression des gênes en utilisant un modèle de régression logistique. Trouver la solution à ce modèle en employant une pénalité ridge avec comme paramètre $\lambda = 1$. Ceci est implémenté dans le package `penalized`. NB : Pensez à centrer les expressions des gênes à zéro.
2. Déterminer l'ajustement donné par le modèle de régression. Celui-ci est presque parfait. Cela est-il dû au phénomène de sur-apprentissage ? Ou alors, ceci est-il dû au fait que les informations biologiques sur les niveaux d'expressions de gênes déterminent de façon exacte le sous-type de leucémie ?
3. Afin de trancher entre les deux explications précédentes, mélanger de façons aléatoire les sous-types, et re-déterminer l'ajustement donné par la regression logistique. A la lumière de ce nouveau résultat, quelle est l'explication la plus plausible ?
4. Comparer les résultats du modèle logistique avec différents paramètres, de $\lambda = 1$ à $\lambda = 1000$. Comment λ influence la possibilité de sur-apprentissage ?
5. Que suggériez-vous pour éviter le sur-apprentissage ?

3.2 Régression logistique + lasso

Télécharger le package `breastCancerNKI` à partir de BioConductor :

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("breastCancerNKI")
```

Activez la librairie et chargez les données `nki` à partir du package :

```
> library(breastCancerNKI)
> data(nki)
```

L'objet (eset) `nki` est désormais disponible. Il contient les profils de 337 patientes atteintes du cancer du sein. Chaque profil comprend les niveaux d'expression de 24481 gênes. Commencer par extraire des données de l'objet `nki`, supprimer les gênes avec valeurs manquantes, centrer les expressions des gênes autour de zéro, et restreindre l'analyse aux premiers mille gênes. Cete réduction a pour seul but d'accélérer le temps de calcul.

```
 X <- exprs(nki)
X <- X[-which(rowSums(is.na(X)) > 0),]
X <- apply(X[1:1000,], 1, function(X) X - mean(X) )
```

Continuer en extrayant l'*estrogen receptor status* (ER status), un important indicateur de pronostic pour la cancer du sein.

```
Y <- pData(nki)[,8]
```

1. Associer l'ER status avec les niveaux d'expression des gênes par un modèle de régression logistique (modèle ridge). Déterminer d'abord, par validation croisée, la valeur optimale du paramètre de pénalisation λ. Utiliser la fonction `optL2` du paquet `penalized`.
2. Evaluer si la vraisemblance calculée par validation croisée atteint son maximum à la valeur optimale de λ. Utiliser la fonction `profL2` du paquet `penalized`.
3. Evaluer la sensitivité de la sélection du paramètre de pénalisation en fonction du choix du paramètre de la validation croisée (fold).
4. Est-ce que le λ optimal produit un ajustement raisonnable ? Comment se compare t-il à l'ajustement ridge ?