MASTER2- MLDS Année 2017-2018



**TP.2 Classification supervisée**

**Régression linéaire - K plus proches voisins – Bayésien naïf**

Rendu : TP à faire en binôme et à envoyer par mail à l’adresse suivante : [l.labiod@gmail.com](mailto:l.labiod@gmail.com) au plus tard le 30 .11.2011

**2 Données réelles (prostate)**

**2.1 Description des données prostate**

Les données prostate examinent la corrélation entre le niveau de l'antigène spécifique de la prostate et un certain nombre de mesures cliniques chez les hommes qui étaient sur le point de recevoir une prostatectomie radicale. La variable lpsa est la réponse au traitement qui sera discrétiser en deux catégories ; high si lpsa > median(lpsa) et 0 sinon. On veut construire un score de détection de la réponse applicable aux patients. Pour chaque patient on a mesuré une batterie de critères et finalement *p* = 3 critères ( lcavol, lweight et age) ont étés retenus pour construire le score.

**A. Charger le jeu de données dans R, décrire le jeu de données prostate**

Les donnees prostate ont 97 individus avec 10 variables suivantes :

lcavol :log cancer volume

lweight :log prostate weight

age :in years

lbph : log of the amount of benign prostatic hyperplasia

svi : seminal vesicle invasion

lcp : log of capsular penetration

gleason : a numeric vector

pgg45 : percent of Gleason score 4 or 5

lpsa : response

train : a logical vector

utiliser les 3 variables explicatives lcavol, lweight and age pour modéliser les classes g et tracer le nuage des points de ces trois variables en indiquant les classes

*prostate\_scatter\_plot.jpeg*

**2.2 Régression linéaire**

**B. Pourquoi la régression linéaire n'est pas adaptée!?**

If you are doing classification, you want to optimize something related to misclassifications. You only care about predicting the right class. When you are doing regression, you want to minimize some measure of distortion between the prediction and the actual value (mean squared error).

- coefficients de régression

(Intercept) lcavol lweight age

-0.943044969 0.200877054 0.388765006 -0.003983511

- tracer le modèle estimé lcavol et age pour lweight moyen

*2\_linear\_regression\_prostate.jpeg*

- nombre d'exemples mal classés

sum(lm.ghat != g)

sum(lm.ghat != g)

[1] 19

- erreur de classification

mean(lm.ghat != g)

[1] 0.1958763

- matrice de confusion

|  |
| --- |
| > table(lm.ghat, g)  g  lm.ghat high low  high 39 11  low 8 39 |
| |  | | --- | | > | |
|  |

**2.3 KNN**

***C.* On s’intéresse d’abord à la méthodologie du choix de *k***

(1) Créer un jeu de données de données d’apprentissage (75% des données) et un jeu de données test (25% des données) avec le code suivant.

(2) Calculer les taux d’erreur sur les données test pour *k* variant de 1 à 100. Avec la fonction plot, représenter ce taux d’erreur test en fonction de *k* (contrôler que l’abscisse du graphique partde 0). Avec la fonction which.min, trouver le nombre de voisins qui donne la plus petite erreur test.

1st test :

> which.min(err\_test)

[1] 4

*KNN\_seed\_30\_train\_0.75\_test\_0.25.jpeg*

(3) Recommencer avec un autre découpage aléatoire apprentissage/test et représenter la courbe d’évolution du taux d’erreur test sur le même graphique qu’à la question précédente.

2nd test:

> which.min(err\_test)

[1] 4

*KNN\_seed\_10\_train\_90\_test\_7.jpeg*

→ interpreter: more training example, better generalization

(4) Exécuter le code suivant et faire un choix pour *k*.

> which.min(mean\_err\_test)

[1] 9

*KNN\_train\_90\_test\_7\_multi.jpeg*

(5) Choisir maintenant le nombre *k* de voisin en utilisant par validation croisée (cross validation) leave-one-out (LOO) avec la fonction knn.cv.

> which.min(err\_test)

[1] 10

*KNN LOO.jpeg*

(6) Faire un petit bilan méthodologique concernant le choix du paramètre *k*.

Choose K what give the minimal error

**D. On veut maintenant non seulement choisir *k* mais également avoir une idée de l’erreur de prédiction de ce classifieur. Pour cela, il faut utiliser des données n’ayant jamais été utilisées. Les données doivent donc être découpées en trois parties : apprentissage/validation/test .**

(2) Utiliser la première approche pour choisir *k* sur l’ensemble "apprentissage-validation" :

i. Choisir *k* en découpant les 97 données de l’ensemble "apprentissage-validation" en deux

parties : une partie "apprentissage" (50% des données) et une partie "validation"

(25 % des données). Choisir *k* qui minimise le taux d’erreur moyen sur les

ensembles de validations de *B* = 25 découpages.

Apprentissage: 49

Validation: 24

Test: 24

*KNN\_train\_valid\_test\_mean\_err\_valid.jpeg*

which.min(mean\_err\_valid)

[1] 5

ii. Constuire le classifieur avec ce nombre de voisins sur l’ensemble "apprentissage-validation" et calculer le taux d’erreur des données test.

pred <- **knn**(Xtrainval[,-1],Xtest[,-1],Xtrainval[,1],k=**which.min**(mean\_err\_valid))

**sum**(pred!=Xtest[,1])/**length**(Xtest[,1])

> sum(pred!=Xtest[,1])/length(Xtest[,1])

[1] 0.4166667

(3) Utiliser la seconde approche pour choisir *k* par validation croisée LOO sur l’ensemble "apprentissage validation". Calculer ensuite le taux d’erreur des données test.

> which.min(err\_valid)

[1] 5

> pred <- knn(Xtrainval[,-1],Xtest[,-1],Xtrainval[,1],k=which.min(err\_valid))

> sum(pred!=Xtest[,1])/length(Xtest[,1])

[1] 0.4166667

**E. Pour les courageux, on pourrait recommencer avec plusieurs découpages des données en deux parties "apprentissage-validation" et "test". Cela permettrait d’avoir une erreur test moyenne, et une idée de sa variabilié. C’est assez rapide à faire avec la méthode de LOO pour le choix de *k.***

KNN LOO train val test.jpeg

**2.4 Bayésien naïf**

1. Appliquer le classifieur byésien naïf sur la table X.

> table(predict(m, prostate.d), g)

g

high low

high 31 6

low 16 44

**3. Données Spam**

Le jeu de données SPAM est une base de données e-mail, avec 4601 observations et 58 variables descriptives.

1. Obtenir une description de la table spam (utiliser le package{kernlab} pour récupérer le jeu de données)

2. Effectuer une première analyse statistique univariée et bivariée de la table spam

3. Réaliser une étude comparative des méthodes de classification suivantes : régression linéaire, k plus proches voisins et le classifieur bayésien naïf sur le jeu de données spam.

4. Refaire la même étude sur le jeu de donnée spam après normalisation (utiliser la normalisation suivante : diviser chaque case nij de la table spam par la racine carrée du produit des sommes marginale ni. Et n.j ) (sqrt(somme(ligne i) \* somme(colonne j))

5. Faire un petit bilan concernant les méthodes utilisées et l’impact de la normalisation sur les performances de ces méthodes.

1. Description des donnees SPAM

A data set collected at Hewlett-Packard Labs, that classifies 4601 e-mails as spam or non-spam. In addition to this class label there are 57 variables indicating the frequency of certain words and characters in the e-mail.

2. Analyse statistique univariee/bivariee

J’ai decide’ de laisser les colonnes 55, 56, 57

3. Réaliser une étude comparative des méthodes de classification suivantes : régression linéaire, k plus proches voisins et le classifieur bayésien naïf sur le jeu de données spam/spam without last 3 columns/spam normalized

|  |  |  |  |
| --- | --- | --- | --- |
|  | Spam | Spam normalized | spam(without last 3 columns) |
| Regression lineaire | erreur de classification  0.1119322 | erreur de classification | erreur de classification  0.1141056 |
| KNN  seed(30)  75 % train, 25 % test  test with K from 1→ 100 | Minimum error test  0.1902693  K optimal = 1 |  | Minimum error test  0.0799305  K = 5 |
| KNN  seed(10)  75 % train, 25 % test  test with K from 1→ 100 | Minimum error test  0.1720243  K optimal = 1 |  | Minimum error test  0.09383145  K = 6 |
| KNN  20 decoupages  75 % train, 25 % test  test with K from 1→ 100 | Minimum mean error test  0.1849696  K optimal = 1 |  | Minimum mean error test  0.08913988  K = 1 |
| KNN CV LOO  utilise 100 % donnees  K from 1→ 100 | Minimum error test  0.1690937  K optimal = 1 |  | Minimum error test  0.08913988  K = 1 |
| KNN  seed(30)  50 % train, 25 % val, 25 % test  25 decoupages  K from 1→ 50 | Minimum mean error validation  0.2021913  K optimal = 1  Test error  0.1902693 |  | Minimum mean error validation  0.1036174  K optimal = 3  Test error  0.08427454 |
| KNN CV LOO  utilise 75 % donnees  10 decoupage  K from 1→ 50 | mean(err\_test)  [1] 0.1874891 |  | Mean error test  0.09391833 |
| Classifieur Bayesien Naif | 0.2864595 |  | 0.2936318 |