Université Paris Nanterre
Master2 ISEFAR Data Mining - Classification

TP5 : Régression logistique pour la classification

Exercice 1 *On considère la classification des emails selon une classe binaire (spam, non-spam). On dispose de 58 variables par email et l'étiquette vaut 0 ou 1. On note* `x` *la matrice des données de taille* $n \times d$ *avec* $n = 4601$ *et* $d = 57$ *et on note* `y` *le vecteur d'étiquettes de taille* $n = 4601$.

1. ***Données (échantillon noté*** `D`***) :***

   Télécharger le package `kernlab`. *Considérer et regarder les données spam de la libraire* `kernlab` *grâce aux commandes*

   ```
   library(kernlab)
   data(spam); str(spam); dim(spam)
   x <-  spam[ ,1:57] ; y <-  as.numeric(spam[,58])-1
   D <- cbind(x,y) ; n <- nrow(spam)
   ```

2. *Utiliser l'échantillon* `D` *et considérer 75% de points pour faire l'apprentissage et 25 % de points pour faire le test (i.e. donner* `Dtrain` *et* `Dtest`*).*

3. *Construire un modèle puis prédire ou classifier dans les points* `Dtest` *grâce aux commandes*

   ```
   fit.glm  <- glm(Dtrain[,58]~., data = Dtrain[,1:57], family=binomial)
   score.glm  <-  predict(fit.glm, Dtest[,1:57], type="response")
   class.glm <-  as.numeric( score.glm >= 0.5)
   ```

4. *Utiliser votre logiciel* `R` *pour calculer la matrice de confusion pour un seuil fixé à 0.5.*

5. *Calculer la proportion de vrais positifs (PVP) et de vrais négatives (PVN), respectivement, pour un seuil fixé à 0.5 grâce aux commandes*

   ```
   sum( score.glm >= 0.5 & Dtest[,58]==1)/sum(Dtest[,58]==1)
   sum( score.glm < 0.5 & Dtest[,58]==0)/sum(Dtest[,58]==0)
   ```

6. *Calculer la proportion de faux posififs (PFP) pour un seuil fixé à 0.5.*

7. *Tracer la courbe ROC associée à la méthode régression logistique.*

   On vous rappelle que la courbe ROC correspond à la proportion de vrais positifs (PVP) en fonction de la proportion de faux positifs (PFP). Elle est obtenue en faisant varier le seuil de décision.

   Remarque : N'hésite pas à regarder la doc. Vous pouvez voir par exemple sur wikipedia (ROC curve) ou dans une autre source. Attention la définition de la courbe ROC n'est pas unique. Dans le livre Elements of Statistical Learning de Hastie et al., elle correspond à la PVP en fonction de la proportion de vrais négatifs (PVN) (page 277-278). L'utilisation pour la comparaison de modèles reste cependant la même.

8. *Donner l'erreur empirique. Donner également l'erreur par validation croisée.*