Université Paris Ouest Nanterre La Défense
Master 2 ISEFAR : Data-Mining Classification

Examen du 13 décembre (durée de l'examen : 1h30)

Exercice 1

Notons $X \in \mathbb{R}^d$ l'ensemble des variables explicatives et $Y \in \{c_1, \ldots, c_k\}$ la classe à prédire. La distribution jointe de (X, Y) est inconnue. Nous disposons d'un échantillon $\mathsf{D} = \{(x_i, y_i)\}_{i=1}^n$ de n copies indépendantes de (X, Y).

1) Expliquer brièvement le principe de la classification supervisée.
2) Soit f un classifieur et ℓ la perte 0/1, à quoi correspond la quantité $\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$.?
3) Est-ce un bon estimateur de la moyenne de l'erreur de prédiction pour un nouveau point x? Expliquer pourquoi?
4) Pourquoi la validation croisée est-elle une meilleure méthode? (Expliquer au passage la méthode de validation croisée)

Exercice 2

Considérons un ensemble de données d'apprentissage $D = \{(x_i, y_i)\}_{i=1}^n$ où $x \in \mathbb{R}^d$ et $y \in \{-1, +1\}$. Supposons que les données d'apprentissage ne sont pas séparable linéairement. Soit $g(x) = \langle w, \phi(x)\rangle + b$ avec ϕ une fonction non-linéaire. Nous nous intéressons à l'apprentissage du classifieur $f = sign(g(x))$ à l'aide de la méthode SVM.

Le classifieur SVM est obtenu par la résolution du problème d'optimisation :

$$\begin{cases} \text{Minimiser} & \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^n \xi_i \\ \text{sous la contrainte} & y_i(\langle w, \phi(x_i)\rangle + b) \geq 1 - \xi_i \quad \text{et} \quad \xi_i \geq 0 \quad i = 1, \ldots, n \end{cases}$$

Remarque : Le minimum en w est le même que celui du problème:

$$\text{Minimiser} \quad \frac{1}{2}\|w\|^2 + \mathrm{C}\sum_{i=1}^{n} \max(1 - y_i(\langle w, \phi(x_i)\rangle + b), 0)$$

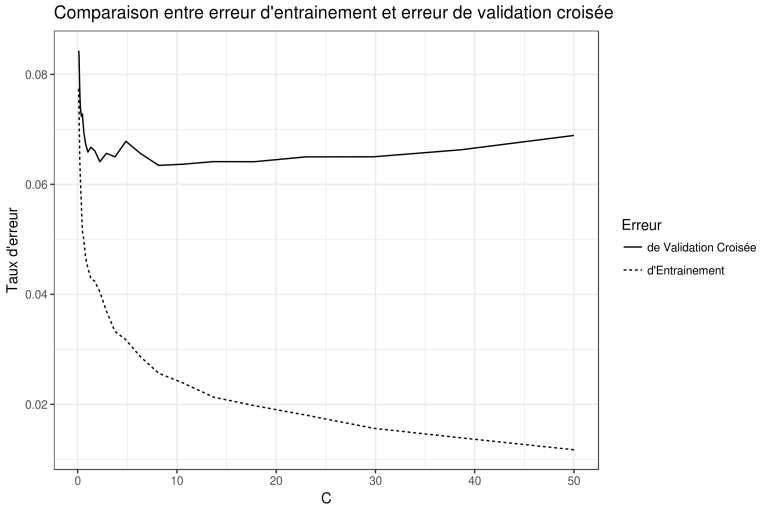
1) Quel est le rôle de la constante de régularisation C ?
2) On peut monter que le w optimal s'écrit $\sum \alpha_i y_i \phi(x_i)$ où les α_i des réels positifs qui s'estiment numériquement. Donner l'expression du classifieur en fonction des α_i. Comment appelle-t-on le produit scalaire $\langle \phi(x), \phi(x_i)\rangle$?
3) Nous avons considéré les données spam (traité dans le TP5 et TP6) de la librairie kernlab qui classe 4601 e-mails comme spam ou non-spam. Ici chaque mail est codé sous la forme d'un vecteur x de dimension 57 ($x \in \mathbb{R}^{57}$).

   i. Nous avons utiliser tous les données et nous avons calculer le classifieur des SVM. Expliquer **minutieusement** toutes les valeurs de sortie de R ksvm ci-dessous.

```
ksvm(as.matrix(spam[,1:57]),spam[[58]],kernel="rbfdot", kpar=list(sigma=0.03),
C=8.161508,cross=20)

parameter : cost C = 8.161508
Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  0.03
Training error : 0.025864
Cross validation error : 0.063685
```

ii. Expliquer **brièvement** le graphique ci-dessous



Exercice 3

1) Donner la formule de la probabilité conditionnelle dans le modèle logistique. Donner l'expression du classifieur du modèle logistique.

2) Nous traitons le problème de classification de e-mails (données spam). Nous avons découpé nos données en deux sous-échantillon : un échantillon d'entrainement (`Dtrain`) et un échantillon de test (`Dtest`). Nous avons ajusté un modèle logistique à l'aide du logiciel R (noté `modelFit_glm`) et utiliser une fonction `predict` qui donne les probabilités estimées pour les deux classes pour les emails (ici le 5ème et le 7ème de `Dtest`).

```
predict(modelFit_glm, newdata=Dtest[c(5,7),],type="prob")
   nonspam      spam
   0.0006390619 0.9993609
   0.5024974029 0.4975026
```

À quelles classes sont attribués ces emails? Justifier.

3) Nous avons calculé la matrice de confusion sur l'échantillon de training et sur l'échantillon de test (pour un seuil s=0.5 fixé).

```
Matrice de Confusion (sur les données de Dtrain)
          Reference
Prediction nonspam spam
nonspam    1872     121
spam         80    1149

Matrice de Confusion (sur les données de Dtest)
          Reference
Prediction nonspam spam
nonspam     795   49
spam         41  494
```

Calculer le taux d'erreur sur les données de training et de test.

4) Comme il est indiqué dans la question 4 de l'exercice 1, il vaut mieux estimer l'erreur de prédiction par validation croisée. Nous avons utilisé tous nos données et nous avons estimé l'erreur de prédiction par validation croisée. L'erreur obtenue est de 0.075. Quel méthode choisir sur ces données? la méthode SVM ou le modèle Logistique? Justifier.