Université Paris Nanterre
Master 2 ISEFAR : Data-Mining Classification

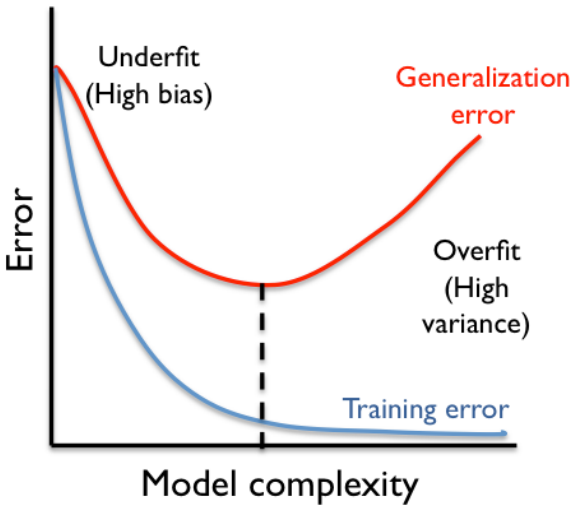
Examen du 11 décembre 2018 (durée de l'examen : 1h30)

Ne pas hésiter à sauter des questions plutôt que de rester coincer.
Bon courage à tous.

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 (sans oublier de parler de la fonction de perte 0/1 et du risque théorique).
2) Montrer que le meilleur classifieur possible (qu'on appelle classifieur de Bayes) est celui qui, pour tout x, affecte l'objet décrit par x à la classe dont la probabilité conditionnelle est la plus grande en ce point.
3) Pourquoi ne peut-on pas utiliser le classifieur de Bayes ? Parmi les classifieurs considérés dans le cours, donner l'expression d'un classifieur qui mime le classifieur de Bayes.
4) Commenter le graphique ci-dessous



5) Comment estimer l'erreur de généralisation à partir des données?

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 :

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

1) Pourquoi on considère terme $C \sum_{i=1}^{n} \max(1 - y_i(\langle w, \phi(x_i)\rangle + b), 0)$. Expliquer (sans oublier de parler du risque empirique et du problème pénalisé). Donner également une illustration graphique.

2) Quel est le rôle de la constante positive C ? Expliquer.

3) 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.

4) Quel est l'astuce du noyau ? Donner deux exemples de noyau, vus en cours, pour les SVM (sans oublier de donner leurs expressions)

Exercice 3

Nous avons utilisé un jeu de données téléchargé dans un concours du site datascience.net. L'objectif est expliquer une variable crédit Y en fonction de certaines caractéristiques des clients. Celle-ci a deux modalités DEFAULT pour un défaut de crédit et NO_DEFAULT sinon. Chaque client est décrit par 18 variables explicatives, comme par exemple BirthDate (date de naissance), Customer_Open_Date(date d'arrivée du client dans la filiale), Customer_Type (existant ou nouveau), Educational_Level (niveau d'éducation), Marital_Status (statut marital)...

Nous disposons d'un sous-ensemble des données d'entrainement (training) de 5380 clients et leurs caractéristiques et un sous-ensemble des données de test de 1345 clients et leurs caractéristiques.

Les effectifs des classes dans la base de données d'entrainement sont de

```
summary(CreditTraining$Y)
DEFAULT NO_DEFAULT
393     4987
```

Ceux pour la base de tests de

```
summary(CreditTesting$Y)
DEFAULT NO_DEFAULT
97      1248
```

Nous avons appliqué plusieurs méthodes vues en cours et, pour mesurer la performance de nos méthodes, nous avons estimé l'erreur de prédiction par validation croisé.

1 Avant de tester des méthodes plus compliquées, c'est toujours une bonne idée d'estimer les performances de la méthode prédisant toujours la classe majoritaire (NO_DEFAULT ici). Quelle est l'erreur de cette prédiction sur les données de tests?

2 La première méthode utilisée est une régression logistique. Le modèle obtenu a été sauvé sous le nom `CreditGlm`.

   i. On peut l'utiliser pour faire des prédictions:

   ```
   ProbGlm <- predict(CreditGlm, newdata = CreditTesting, type = "prob")
     DEFAULT     NO_DEFAULT
   1 0.01095451  0.9890455
   2 0.84764457  0.1523554
   ```

   À quelles classes sont attribués ces deux clients ? Justifier.

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

```
Matrice de Confusion (sur les données de CreditTesting)
Reference
prediction   DEFAULT NO_DEFAULT
DEFAULT         49         16
NO_DEFAULT      48       1232
```

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

iii. Nous avons estimé l'erreur de prédiction par la méthode de validation croisée k-fold et obtenus les erreurs suivantes pour chaque fold.

```
Error_Folds
Resample Error_Fold
1    Fold1 0.05762082
2    Fold2 0.04828227
3    Fold3 0.05209302
4    Fold4 0.05116279
5    Fold5 0.04828227
```

Rappeler très brièvement comment sont obtenus ces erreurs. Comment en déduire l'erreur de validation croisée (0.05148823)?

3 La seconde méthode proposée est la méthode des k plus proches voisins.

i. A quoi correspond la valeur de k?

ii. Que se passe-t-il lorsque k est trop petit? lorsque k est trop grand?

iii. Nous avons comparé 6 valeurs de k par validation croisée:

```
   k    ErrorCV
1  3 0.07825074
2  5 0.07583663
3  9 0.07025956
4 11 0.06895810
5 17 0.07118737
6 23 0.07081649
```

Laquelle choisir? Justifier la réponse.

4 Pour finir, nous avons testé un SVM. Le meilleur modèle obtenu correspond à cette sortie dans R

```
Support Vector Machine object of class "ksvm"
SV type: C-svc  (classification)
parameter : cost C = 1.25

Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  0.01
Number of Support Vectors : 784

Objective Function Value : -814.5187
Training error : 0.047026
Cross validation error : 0.056692
```

Expliquer **minutieusement** toutes les valeurs apparaissant dans cette sortie.

5 Quel modèle choisir ici parmi tous ceux vus dans l'exercice?