Université de Paris | **Master 2 Mathématiques et Applications**
Parcours MMA | **Classification**

Examen (2h30)

Les deux premiers exercices sont à faire sur feuille, tandis que le troisième exercice est à faire sur machine. Vous me remettrez d'une part votre copie pour les exercices 1 et 2 et vous m'enverrez par email les réponses de l'exercice 3. L'exercice 3 peut être rédigé en Rmarkdown (pdf ou html) ou tout simplement en word avec les figures incluses dans le document (les codes peuvent également être inclus ou alors mis dans un fichier R à part). Si l'exercice 3 est rédigé en word, pensez à le convertir en pdf avant de me l'envoyer.

Exercice 1.
Pour un problème de classification, on considère les données d'apprentissage ($n = 9$) présentées dans la Figure 1. Chaque point représente un individu dont la covariable (appartenant à $\mathbb{R}^2$) est définie par les coordonnées du point et la couleur du point indique la valeur de l'étiquette Y correspondante (rouge pour 1 et noir pour 0). Par exemple, le point en bas à droite de la figure représente un individu ayant pour covariable $X = (6, 1)^t$ et pour étiquette $Y = 0$.

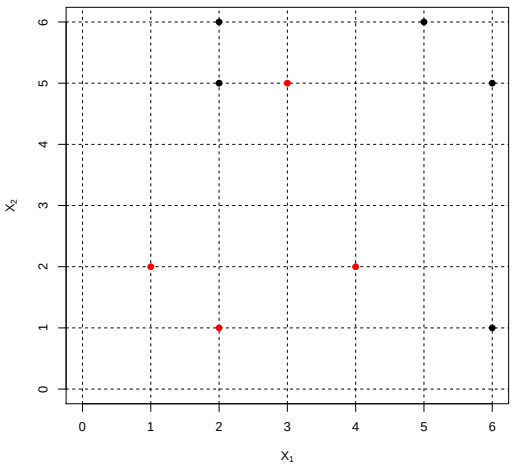

FIGURE 1 – Données pour l'exercice 1.

1. Rappeler le principe de l'algorithme des k plus proches voisins. Quel serait le taux de mauvaise classification de l'algorithme des 1 plus proches voisins sur l'échantillon d'apprentissage. En général, vous semble-t-il que l'algorithme des 1 plus proches voisins est un bon algorithme de classification ? Justifier.
2. Pour toute nouvelle observation, quelle classification donnerait l'algorithme des 9 plus proches voisins ? Quel serait son taux de mauvaise classification sur l'échantillon d'apprentissage (les données de la Figure 1) ?
3. On considère à présent l'algorithme des 3 plus proches voisins. Comment cet algorithme classifierait-il la nouvelle observation $X^* = (2, 4)^t$? Justifier.
4. À partir de l'algorithme CART, on a obtenu sous R l'arbre complet T^c, présenté dans la Figure 2. Remplir les cases blanches permettant de préciser l'étiquette prédite dans chaque feuille de l'arbre T^c. Que signifient les valeurs 0.00 et 33% dans la feuille terminale en bas à gauche de T^c ?
5. Expliquez pourquoi il n'est pas possible d'obtenir un arbre plus profond que celui de la Figure 2.
6. Pour le premier découpage de l'arbre (à la racine), expliquer pourquoi il est préférable de découper l'arbre selon le critère $X_1 \geq 5$ plutôt que selon le critère $X_2 \geq 4$?

7. On considère le sous-arbre $\tilde{T}$ de la Figure 2. Remplir les cases blanches permettant de préciser l'étiquette prédite dans chaque feuille de l'arbre T^c. Comparer le risque de classification de l'arbre complet T^c avec celui de $\tilde{T}$.
8. Pour un paramètre $\alpha > 0$ de pénalisation, un arbre T appartenant à l'ensemble des sous-arbres de T^c, on considère alors le critère pénalisé suivant

$$\mathrm{Crit}_\alpha(T) = R(T) + \alpha|T|,$$

où $R(T)$ représente le risque de classification de T et $|T|$ représente la complexité de T, mesurée par son nombre de nœuds terminaux. Pour quelles valeurs de α, le sous-arbre $\tilde{T}$ présente-il un risque pénalisé Crit_α inférieur à celui de l'arbre maximal T^c ?
9. Quelle classification obtient-t-on pour la nouvelle observation $X^* = (2, 4)^t$ en utilisant l'arbre complet T^c ? En utilisant le sous-arbre $\tilde{T}$?

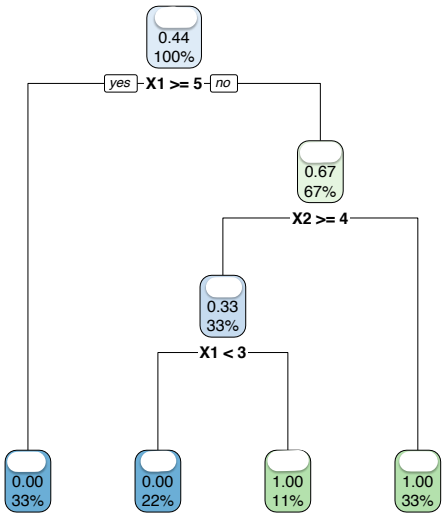

FIGURE 2 – Arbre de classification maximal T^c.

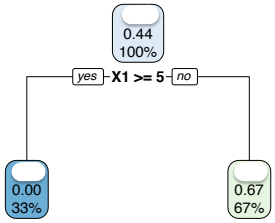

FIGURE 3 – Sous-arbre de classification $\tilde{T}$.

Exercice 2.
On considère un modèle de classification où le couple aléatoire (X, Y) suit la loi décrite par les deux relations suivantes :

$$\mathcal{L}(X \mid Y = 0) = \mathcal{U}([0, \theta])$$
$$\mathcal{L}(X \mid Y = 1) = \mathcal{U}([0, 1]),$$

où $Y \in \{0, 1\}$, $\mathcal{U}$ représente la loi uniforme continue et $\theta \in]0, 1[$. On note également $p = \mathbb{P}[Y = 1]$, avec $p \in]0, 1[$.

1. Calculer $\eta(x) = \mathbb{P}[Y = 1 \mid X = x]$ en fonction de p et θ, puis en déduire une expression du classifieur de Bayes en fonction de p et θ.
2. En particulier, que prédit le classifieur de Bayes quand $\theta < x < 1$? Montrer que si $p \geq 1/(1 + \theta)$ alors le classifieur est constant et prédit toujours 1 quelque soit la valeur de x.
3. Montrer que le risque de classification du classifieur de Bayes (noté g^*) vaut :

$$R(g^*) = (1 - p)\mathbb{1}_{p \geq 1/(1+\theta)} + \theta p \mathbb{1}_{p < 1/(1+\theta)}.$$

Exercice 3.
Dans cet exercice, vous allez analyser une base de données sur le cancer du sein. Chaque observation est une cellule tumorale contenant de nombreuses informations. Au total, 30 covariables telles que le rayon, la texture, le périmètre etc. sont accessibles et 569 observations ont été recueillies. Le but est de classifier si la tumeur est maligne (M) ou bégnine (B).

Charger le fichier "Breast.RData", accessible depuis ma pageweb. Depuis RStudio, taper `load("Breast.RData")`. Cette commande a pour effet de charger deux bases de données, `breast_train` (l'échantillon d'apprentissage) et `breast_test` (l'échantillon test). La première contient 455 observations tirées aléatoirement depuis la base d'origine et la seconde contient les 114 observations restantes. Dans chacune des bases de données, la première colonne correspond à l'étiquette et les 30 autres colonnes correspondent aux covariables.

1. En jouant avec les paramètres de la fonction `rpart`, construire un arbre de classification CART de profondeur maximale. Sélectionner un ou deux arbres élagués avec la ou les méthodes de votre choix en justifiant. Représenter graphiquement cet ou ces arbres élagués. Commenter.
2. Donner le taux de mauvaise classification, calculé sur l'échantillon test, à partir de l'arbre ou les arbres sélectionnés à la question précédente. Comparer avec le taux de mauvaise classification calculé sur l'arbre CART de profondeur maximale.
3. Construire un modèle de bagging basé sur 50 arbres de profondeurs maximales et calculer son taux de mauvaise classification.
4. Tester l'effet du nombre d'arbres en calculant le taux de mauvaise classification pour des valeurs du nombre d'arbres égales à $5, 10, 20, 30, 40, 50, 60, 80, 100$. On pourra faire ce calcul dans une boucle `for` et on pourra donner les valeurs numériques du taux de mauvaise classification ou faire une représentation graphique.
5. Procéder de la même manière, cette fois à partir des forêts aléatoires. On justifiera d'un choix optimal du nombre d'arbres (option `ntree`) et du nombre de variables candidates à chaque découpage de l'arbre (option `mtry`) en entraînant les arbres sur l'échantillon d'apprentissage et en évaluant le taux de mauvaise classification sur l'échantillon test.
6. Pour terminer, on va construire un modèle de svm linéaire pour prédire l'étiquette. Tout d'abord, sélectionner une valeur du paramètre `cost` par cross-validation. On pourra pour cela utiliser la fonction `tune` pour la méthode `svm` du package `e1071`.
7. Entraîner un svm avec le choix du paramètre `cost` sélectionné à la question précédente sur l'échantillon d'apprentissage et calculer le taux de mauvaise classification sur l'échantillon test.
8. Conclure sur l'ensemble des méthodes étudiées et sur leurs performances respectives.