M1 MA Université Paris-Saclay
Marie-Anne.Poursat@math.u-psud.fr

STA212 : Méthodes de simulation statistiques

TP2 à envoyer avant le jeudi 9 mai 2019 16h

Modalités

L'évaluation de l'enseignement STA212 est un mini-projet composé de 2 TP notés à faire de préférence en binôme. Vous pouvez discuter ensemble de l'implémentation des méthodes (c'est même conseillé) MAIS votre code, la rédaction de votre démarche et de vos conclusions doivent être STRICTEMENT PERSONNELS (au binôme). Il y a plusieurs démarches possible, l'important est la **cohérence**, la **fiabilité** et la reproductibilité de votre étude.

Vous devez transmettre par mail pour chaque TP :

1. un fichier script à vos noms (Nom1-Nom2.R) comportant votre code "opérationnel" (c'est à dire que lorsque je le lancerai il marchera sans intervention de ma part). Votre code sera commenté de façon à être lisible facilement.
2. un fichier pdf (maximum 4 pages) comportant la rédaction de l'analyse statistique, les résultats numériques pertinents *commentés* et les illustrations graphiques. Le compte-rendu doit être clair, synthétique et bien écrit. **PAS de code R** dans cette rédaction.

   Si des ressources obtenues par une recherche bibliographique et/ou en ligne sont utilisées, elles doivent être impérativement citées.

TP2 : travail demandé

Nous étudions des données destinées à effectuer une reconnaissance automatique de caractères manuscrits [1]. Elles proviennent des codes postaux manuscrits lus sur les enveloppes de courriers américains. L'objectif est d'identifier dans une image noir et blanc de 16×16 pixels un chiffre de 0 à 9. Les images ont été normalisées pour avoir approximativement la même taille et la même orientation. Un échantillon est représenté dans la Figure 1. Chaque ligne du tableau de données représente une image : elle contient le chiffre représenté `y`, de 0 à 9, et les 256 niveaux de gris de chaque pixel. A chaque pixel d'une image est associé un nombre réel compris entre -1 (noir) et 1 (blanc). On a donc $Y \in \{0, \ldots, 9\}$ et $X \in [-1, 1]^{256}$.

1. Quelle est la loi de Y et le type de modèle statistique associé à ce problème ?

   Chargez les données d'apprentissage `train.txt`. Quels sont les effectifs de chaque modalité de Y ?
2. Construisez un arbre CART : expliquez soigneusement chaque étape et vos choix de paramètres. Décrivez l'arbre obtenu.

   Quelle est la prédiction de l'arbre CART pour une image toute noire ? une image toute blanche ? Justifiez.
3. Construisez une forêt aléatoire : expliquez comment vous choisissez ses paramètres.

   Comparez ses performances avec celles de la méthode bagging.
4. Chargez les données `test.txt` et comparez les trois modèles précédents sur ces données. Rédigez une conclusion de cette étude.

1. *The elements of Statistical Learning*, Hastie, Tibshirani and Friedman

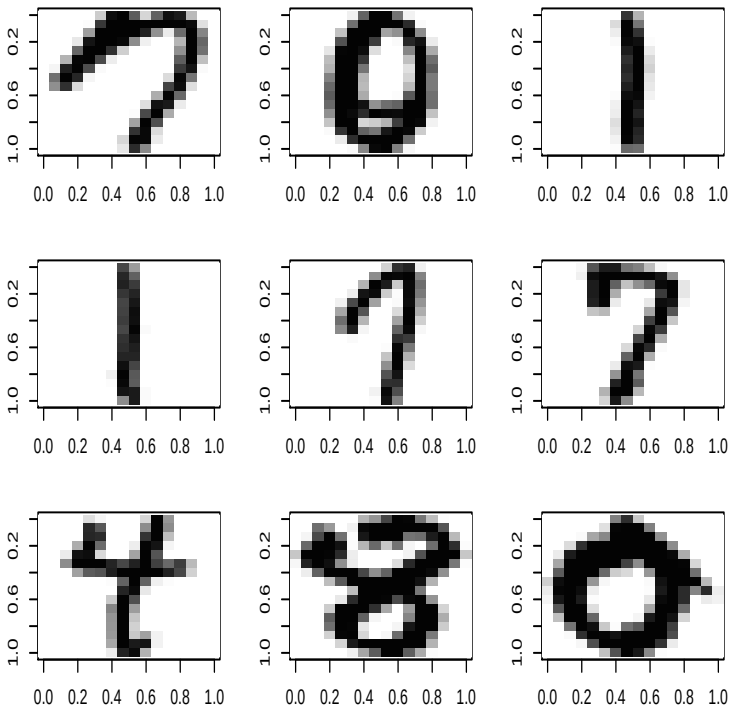

FIGURE 1 – Représentation graphique de 9 images.