Université Paris-sud
M2 AMI2B **Apprentissage statistique** Marie-Anne.Poursat@math.u-psud.fr

Modalités du Projet 2016

Généralités

Le projet est individuel. Vous pouvez travailler en groupe sur l'implémentation des méthodes (c'est même conseillé) MAIS votre code, la rédaction de votre démarche et de vos résultats ainsi que vos conclusions doivent être STRICTEMENT PERSONNELS. Il y a plusieurs démarches possible, l'important est la **cohérence** et la **fiabilité** de votre étude.

Vous me rendrez deux fichiers : un fichier pdf (*rapport_prénomNOM.pdf*) comportant votre compte-rendu de travail et un fichier de code (*code_prénomNOM.R*) contenant votre code "opérationnel" (c'est à dire que lorsque je le lancerai il marchera sans intervention de ma part et fournira les résultats. Votre code sera annoté de façon à être lisible facilement.

Les deux fichiers doivent me parvenir avant **9h00 le LUNDI 12 DECEMBRE 2016**.

Vous allez travailler sur un jeu de données publiques, *Breast Cancer Wisconsin (Diagnostic) Dataset*, disponible sur le *UC Irvine Machine Learning Repository* et sur votre espace de cours Dokeos. Vous trouverez deux fichiers, *wdbc.data* contenant les données et *wdbc.txt* contenant leur descriptif.

J'ai également déposé sur Dokeos deux documents qui peuvent vous aider à travailler le projet : un document de WikiStat fait par des collègues de Toulouse (attention, ce ne sont pas les mêmes données !) et un article qui présente une étude possible des données du projet. Vous pouvez également vous appuyer sur les références du cours pour développer vos analyses.

Travail demandé

L'étude se compose de deux parties :

1. Mise-en-œuvre avec R des méthodes et modèles développés en cours pour étudier le jeu de données choisi : trois modèles/méthodes au minimum doivent être implémentées. Le travail se termine par une comparaison des performances de chaque méthode. Ce travail se traduit par un code R commenté.
2. Rédaction d'un rapport, qui détaille et explique l'ensemble du travail effectué et contient les résultats numériques *commentés* et les illustrations graphiques.

De 5 à 8 pages maximum (figures éventuellement en plus), le rapport doit être clair, bien écrit et bien organisé.

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

Ce que l'on doit trouver dans le rapport

1. une étude descriptive concise des données.
2. un paragraphe par méthode d'estimation/prédiction étudiée :
   — donner une rapide motivation du modèle estimé, expliquer comment ont été choisis les *tuning parameters* et/ou les variables sélectionnées.
   — préciser quels diagnostics permettent de valider le modèle ;
   — expliquer comment est calculée la réponse prédite par le modèle pour un nouveau vecteur de covariables, donner un exemple.
   — évaluer la capacité de prévision ou classification du modèle.
3. une explication précise de la méthodologie d'évaluation et de comparaison des performances que vous avez mise en oeuvre.
4. une table des matières qui permet d'un coup d'œil d'apprécier le travail effectué et l'organisation du compte-rendu.

Remarque importante : les résultats doivent être inclus, commentés et interprétés dans le texte.

Notation

La note de projet se répartit selon les proportions suivantes :
— 1/2 pour la qualité de l'analyse statistique effectuée ;
— 1/6 pour la pertinence et/ou l'originalité de l'approche proposée ;
— 1/3 pour la qualité du rapport : explications, recul dans l'analyse, pertinence du choix des résultats montrés, analyse critique des résultats.

Note de l'UE apprentissage statistique = 50% note projet + 50% note examen