Université d'Angers — Année 2022-2023
Département de Mathématiques — M2 Data Science

Statistique en Grande Dimension et Apprentissage
TP-Projet

Ce travail est individuel et compte pour 50% de la note finale. Le but est d'illustrer et de tester les méthodes d'apprentissage statistique vues tout au long du semestre. Un rapport de vos travaux (dont une partie peut être réalisée en Notebook) est à rendre pour le jeudi 12 janvier 2023 (la date est lointaine mais je vous conseille de commencer assez rapidement). Une très courte soutenance sera organisée la semaine du 23 janvier 2023 (semaine après les examens).

Exercice 1 (Programmation).

1. Ecrire "à la main" un algorithme de gradient boosting pour la classification multi-classes avec fonction de perte "déviance" prenant en entrée une classe de prédicteurs (pas forcément un arbre), un pas (learning rate), une taille d'échantillon bootstrap. On pourra mettre en option la possibilité de séparer les points qui servent à l'apprentissage et ceux qui servent aux échantillons bootstrap. Fabriquer également une fonction qui génère un graphe de l'erreur d'entraînement et de l'erreur test (et éventuellement de l'erreur de validation croisée) (Pour l'erreur test, on prendra bien entendu un échantillon indépendant des échantillons qui ont servi à la construction de l'algorithme de gradient boosting).
2. On propose dans cette seconde partie de mettre en application, de proposer des variantes, de tester différentes mises en oeuvre. Pour la mise en application, on pourra se servir de la base MNIST (pas forcément de toute la base) ou on pourra fabriquer des données simulées si on le souhaite. A partir de là, on pourra imaginer quelques développements : ci-après, quelques propositions parmi d'autres. On pourra en choisir une ou plusieurs. Comment fonctionne le gradient boosting avec les k-plus proches voisins, proposer des solutions pour gérer les données désequilibrées dans le gradient boosting, dans le cas spécifique des arbres, comment gérer la complexité de l'arbre (il n'est pas demandé de reprogrammer les arbres ou les k-ppv), comment régulariser (cf xgboost), peut-on remplacer la descente de gradient par une descente de Newton... ?

N.B. On pourra facilement trouver des jeux de données sur `Kaggle` ou sur `https://archive.ics.uci.edu/ml/datasets.php`.

Exercice 2. Dans ce second exercice, on s'intéresse à une base de données liée au cancer sein (Attention, les données sous un format $p \times n$, *i.e.* variable $\times$ individu). Les bases de données sont téléchargeables au lien suivant : `https://plmbox.math.cnrs.fr/f/fedcac32b2a949198dce/` Dans cette base de données, l'objectif de prédire la réaction au traitement. La variable à prédire est "treatment_response" à partir de données génétiques et d'autres caractéristiques (âge/ethnie/Stade de la tumeur T/N). On pourra éventuellement se contenter des données génétiques afin de simplifier le problème et probablement d'éviter de mettre sur un même plan des variables "cliniques" probablement très corrélées avec la réponse et des variables génétiques dont la complexité nécessite des méthodes plus complexes.

On propose ici de comparer plusieurs méthodes : Régression sur composantes principales (non recommandé a priori mais ça peut se tester !), Régression PLS, LASSO, Sparse PCA+

Régression et éventuellement un algorithme de type complètement différent pour terminer (probablement plus performant). Une approche "active" sera bien sûr appréciée (par "active", on sous-entend que vous pouvez prendre des initiatives pour améliorer les méthodes ci-dessus ou sur la manière de mesurer les résultats, pas uniquement en "accuracy" par exemple).

Exercice 3 (Data Challenge)**.** Ce projet sera proposé en décembre. Il s'agit d'un Data Challenge proposé par `https://challengedata.ens.fr/`. Le choix est en cours.