Projet Apprentissage supervisé

Master2 MLSD/AMSD
Année académique 2024/2025
Enseignant : Lazhar Labiod
Adresse : Centre Borelli – Université de Paris
Mail : lazhar.labiod@u-paris.fr

Objectif
L'objectif de ce travail est la mise en pratique concrète d'un certain nombre de techniques d'apprentissage supervisé (Bayesien Naïf, KNN, LDA, QDA, Linear SVM, Non Linear SVM, Régression logistique, CART et Random Forest, etc..), à travers l'étude de données réelles nécessitant l'utilisation de logiciels de traitement statistique de données R ou Python. Les applications visées concernent deux types de données réelles.

1. Données Crédits bancaires :

L'ensemble de données de crédits décrit les détails financiers et bancaires des clients et la tâche consiste à déterminer si le client est bon ou mauvais. L'hypothèse est que la tâche consiste à prédire si un client remboursera un prêt ou un crédit. L'ensemble de données comprend 1000 exemples et 20 variables, dont 7 numériques (entiers) et 13 catégorielles.

Noms des variables :
(Statut du compte courant existant, Durée en mois, Historique de crédit, Objectif, Montant du crédit, Compte d'épargne, Emploi actuel depuis, Taux de versement en, pourcentage du revenu disponible, Statut personnel et sexe, Autres débiteurs,Résidence actuelle depuis, Propriété, Age en années, Autres plans de versement, Logement, Nombre de crédits existants dans cette banque, Travail, Nombre de personnes à charge, Téléphone, Travailleur étranger).

Source des données Crédits :
https://raw.githubusercontent.com/jbrownlee/Datasets/master/german.csv

2. Données relationnelles :

Les données relationnelles représentent deux types d'information, une matrice des valeurs objets/caractéristiques et un graphe des liens entre objets, qui fournissent des informations utiles sous différents angles, mais ils ne sont pas toujours cohérents et doivent donc être soigneusement alignés pour obtenir les meilleurs résultats de classification. L'objectif de cette partie du projet est d'aborder ce problème, afin de mettre en lumières les différents challenges posés par ce type de données aux méthodes de classification.

Je vous encourage à faire preuve d'originalité : vous pouvez très bien utiliser des méthodes qui n'ont pas été présentés en cours, telles que Gradient Boosting, Xgboost, XtremTree, Adaboost

Description détaillée des données relationnelles

Cora : L'ensemble de données Cora comprend 2708 publications scientifiques classées dans l'une des sept classes. Le réseau de citations comprend 5429 liens. Chaque publication dans l'ensemble de données est décrite par un vecteur de mot de valeur 0/1 indiquant l'absence / la présence du mot correspondant dans le dictionnaire. Le dictionnaire se compose de 1433 mots uniques.

CiteSeer : CiteSeer comprend 3312 publications scientifiques classées dans l'une des six classes. Le réseau de citations se compose de 4732 liens. Chaque publication dans l'ensemble de données est décrite par un vecteur de mot de valeur 0/1 indiquant l'absence / la présence du mot correspondant dans le dictionnaire. Le dictionnaire se compose de 3703 mots uniques.

Pubmed : chaque publication de l'ensemble de données est décrite par un vecteur de mots pondéré TF / IDF du dictionnaire. Les relations de citation sont utilisées pour construire les structures du réseau.
La table suivante résume les caractéristiques de ces bases relationnelles (au format matlab*).

Dataset	**# individus**	**# liens**	**# variables**	**#classes**
Cora (fea, W, gnd)	2780	5429	1433	7
CiteSeer (fea, W, gnd)	3327	4732	3703	6
Pubmed(fea, W, gnd)	19717	44338	500	3

*Description des bases matlab : fea : la matrice X(n,d) où n est le nombre d'individus, d est le nombre de variable -- W: la matrice d'adjacence (des liens entre les individus) W(n,n) -- gnd : vecteur des labels (classes).

L'objectif de cette partie du projet est de mener une étude comparative des différentes méthodes de classification sur des données relationnelles en utilisant

1. Uniquement l'information contenue dans la matrice X
2. Une Combinaison des informations W et X ; M=D^(-1)*W*X, où D est une matrice diagonale, chaque valeur diagonale correspond à la somme des valeurs d'une ligne de W.
3. Discuter d'autres idées pour combiner et aligner les deux types d'information.

Travail à faire

1. Commencer par une étude exploratoire préliminaire
2. Utiliser les différentes techniques de classification supervisée vue en cours pour créer un modèle de scoring. Suivant les techniques utilisées (et les fonctions disponibles sous R ou python), vous pourrez utiliser l'ensemble des variables disponibles ou uniquement les variables quantitatives, et réaliser ou non une sélection de variables.
3. Comparer l'ensemble de ces techniques à l'aide des mesures telles que (Accuracy, NMI et la F-measure), évaluées soit par validation croisée soit sur échantillon test.

Rapport ou (Notebook Python, R)

Le rapport du projet doit présenter de façon claire et concise:
- l'objet de l'analyse
- la description des données (individus/variables utilisées, variables supplémentaires etc.)
- l'analyse proprement dite
- les commentaires sur les résultats obtenus.

Ce rapport ne devrait pas dépasser 20 pages (les codes sources des programmes utilisés peuvent être mis en annexe).

Le projet sera jugé selon les critères suivants:
- Adéquation des méthodes utilisées aux données et problèmes étudiés.
- Richesse des analyses proposées (au-delà du minimum requis).
- Justesse des commentaires sur les résultats.
- Qualité de la présentation du rapport.

Remise du rapport
Vous devez déposer votre rapport en format .pdf **au plus tard le 22/11 à 23:59** sur Moodle.

Important : le Projet est à réaliser par groupe de 3 étudiants maximum.

Aide1. Refaire le traitement proposé dans cet article de blog concernant les imbalanced data (partie avec le package caret) :
https://shiring.github.io/machine_learning/2017/04/02/unbalanced
Aide2. Imbalanced-learn ---- https://www.jmlr.org/papers/volume18/16-365/16-365.pdf
Aide3. Code python : comparaison des méthodes de classification