Projet « Apprentissage supervisé »

Master2 MLSD

Année académique 2020/2021

Enseignant : Lazhar Labiod
Adresse : LIPADE – 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), à travers l'étude de données réélles nécessitant l'utilisation de logiciels de traitement statistique de données R ou Python. Les applications visées concernent deux types de données réelles ;

1. **Données bancaires :**
   a. **Carte visa** : le scoring d'une base de données de « carte visa ». En résumé, il s'agit de travailler dans ce projet sur une base de données décrivant les clients d'une banque et leurs comportements (mouvements, soldes des différents comptes). L'objectif est l'estimation d'un score d'appétence à la carte VISA Premier. C'est une carte de paiement haut de gamme qui cherche à renforcer le lien de proximité avec la banque en vue de fidéliser une clientèle aisée.
   b. **Fraude bancaire** : Détection de fraude dans des transactions bancaires : En résumé, il s'agit de travailler dans ce projet sur une base de données décrivant des transactions bancaires sur une période donnée, l'objectif est la détection des transactions frauduleuses.

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 basées 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

Application 1 : Données bancaires

Cette partie s'intéresse à deux cas pratiques (clients d'une banque, transactions bancaires), l'objectif est d'appliquer les différentes approches vues en cours, choisir pour chaque méthode le meilleur modèle et ensuite comparer ces modèles sur un ensemble de test qui n'a pas été utilisé dans les phases d'apprentissage et de validation des modèles en concurrence

Data1 : (Visa Premier) : Il s'agit d'une base de données décrivant les clients d'une banque et leurs comportements (mouvements, soldes des différents comptes). La variable à expliquer Y est la variable binaire « Possession de la carte Visa Premier ».

Voici le dictionnaire des variables de la table Visa

Identif.	Libellé
matricul	Matricule (identifiant client)
departem	Département de résidence
ptvente	Point de vente
sexe	Sexe (qualitatif)
age	Age en année
sitfamil	Situation familiale (Fmar : marié, Fcel : célibataire, Fdiv : divorcé, Fuli : union libre, Fsep : séparé de corps, Fveu : veuf)
anciente	Ancienneté de relation en mois
csp	Catégorie socio-professionnelle (code num)
codeqlt	Code « qualité » client évalué par la banque
nbimpaye	Nombre d'impayés en cours
mtrejet	Montant total des rejets en francs
nbopguic	Nombre d'opérations par guichet dans le mois
moycred3	Moyenne des mouvements nets créditeurs des 3 mois en kF
aveparmo	Total des avoirs épargne monétaire en francs
endette	Taux d'endettement
engagemt	Total des engagements en francs
engagemc	Total des engagements court terme en francs
engagemm	Total des engagements moyen terme en francs
nbcptvue	Nombre de comptes à vue
moysold3	Moyenne des soldes moyens sur 3 mois
moycredi	Moyenne des mouvements créditeurs en kF
agemvt	Age du dernier mouvement (en jours)
nbop	Nombre d'opérations à M-1

Identif.	Libellé
mtfactur	Montant facturé dans l'année en francs
engageml	Engagement long terme
nbvie	Nombre de produits contrats vie
mtvie	Montant des produits contrats vie en francs
nbeparmo	Nombre de produits épargne monétaire
mteparmo	Montant des produits d'épargne monétaire en francs
nbeparlo	Nombre de produits d'épargne logement
mteparlo	Montant des produits d'épargne logement en francs
nblivret	Nombre de comptes sur livret
mtlivret	Montant des comptes sur livret en francs
nbeparlt	Nombre de produits d'épargne long terme
mteparlt	Montant des produits d'épargne long terme en francs
nbeparte	Nombre de produits épargne à terme
mteparte	Montant des produits épargne à terme
nbbon	Nombre de produits bons et certificats
mtbon	Montant des produits bons et certificats en francs
nbpaiecb	Nombre de paiements par carte bancaire à M-1
nbcb	Nombre total de cartes
nbcbptar	Nombre de cartes point argent
avtscpte	Total des avoirs sur tous les comptes
aveparfi	Total des avoirs épargne financière en francs
cartevp	Possession de la carte Visa Premier
sexer	Sexe codé en 0/1
cartevpr	Possession de la carte Visa Premier codé en 0/1
nbjdebit	Nombre de jours de débit

Data2 : Credit card_ Fraud : (pour plus de détails, voir https://www.kaggle.com/dalpozz/creditcardfraud).

Le jeu de données contient les transactions effectuées par cartes de crédit en septembre 2013 par les titulaires de carte européens. Cet ensemble de données présente les transactions qui se sont produites en deux jours, où nous avons 492 fraudes sur 284 807 transactions. L'ensemble de données est très déséquilibré, les classes positives (fraudes) représentent 0,172% de toutes les transactions. Il contient uniquement des variables d'entrée numériques résultant d'une transformation PCA. Les caractéristiques V1, V2, ... V28 sont les composantes principales obtenues avec PCA, les seules caractéristiques qui n'ont pas été transformées avec PCA sont 'Time' et 'Amount'. La variable 'Time' contient les secondes écoulées entre chaque transaction et la première transaction de l'ensemble de données. La variable 'Amount' est le Montant de la transaction, cette caractéristique peut être utilisée pour l'apprentissage sensible aux coûts dépendant de l'exemple. La fonction 'Class' est la variable de réponse et prend la valeur 1 en cas de fraude et 0 sinon. Compte tenu du rapport de déséquilibre de classes, nous recommandons de mesurer la précision en utilisant l'aire sous la courbe de rappel de précision (AUPRC). La précision de la matrice de confusion n'est pas significative pour une classification non équilibrée.

dataset	# d'observations	# de variables	# de classes
VISA	1073	47	2
Fraud-carte-crédit	284 807	31	2

Application 2 : Données relationnelles

1. 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.
2. 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.
3. 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
CitSeer (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 (Question facultative)

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 problème 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 envoyer votre rapport en format *.pdf* au plus tard **le 8 Février 2021 avant minuit** à l'adresse suivante *l.labiod@gmail.com*

Important : le projet est à faire 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