Projet « Apprentissage supervisé »

Master2 MLSD

Année académique 2019/2020

Enseignant : Lazhar Labiod
Adresse : LIPADE - Université Paris Descartes
Mail : lazhar.labiod@parisdescartes.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 synthétiques et deux cas pratiques nécessitant l'utilisation de logiciels de traitement statistique de données R ou python. Les applications visées sont :

1. **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.
2. **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.

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 et Adaboost.

Etude sur données synthétiques

Cette partie concerne un travail sur données synthétiques. Il s'agit de mettre en œuvre des méthodes vues en cours sur l'ensemble des jeux de données synthétiques proposés. Le travail consiste à réaliser une étude comparative des ces différentes approches de classification supervisée.

Données synthétiques

Il s'agit de 3 bases de données synthétiques possédant des caractéristiques différentes, en termes de nombre de classes et de la structure des classes.

Tables	# d'observations	# de variables	# nombre de classes
Flame	240	2	2
Spiral	312	2	3
Aggregation	788	2	7

Etude de cas pratiques

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.

Données réelles

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.

Tables (réelles)	# d'observations	# de variables	# de classes
VISA	1073	47	2
Fraud-carte-crédit	284 807	31	2

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 de courbes ROC (AUC), évaluées soit par validation croisée soit sur échantillon test.

Rapport

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 6 Janvier 2020 avant minuit** à l'adresse suivante *l.labiod@gmail.com*

Aide. 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