meritis

Semaine 8: Détection d'anomalie

Détection de fraude par carte de crédit

On s'intéresse à développer un modèle intelligent capable de détecter des transactions frauduleuses par carte de crédit et prévenir les clients de ce risque. La détection d'anomalie sera faite par la distribution gaussienne (distribution normale).

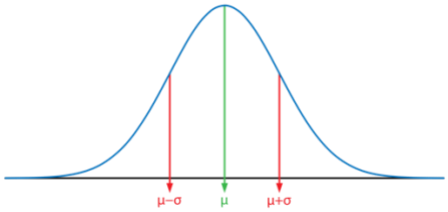
À notre disposition est un ensemble de données contenant 284,807 transactions effectuées en Europe en Septembre 2013 dont 0.172% sont frauduleuses (dont le label est positif).
Les caractéristiques V1, …, V28 ont été transformées en utilisant une technique de préparation des données intitulée analyse des composants principaux (ACP) pour réduire la dimensionnalité des caractéristiques et des raisons de sécurité.

La base de données se trouve ici : https://www.kaggle.com/mlg-ulb/creditcardfraud

I. Préparation de l'environnement

1. Quel est l'objectif de l'ACP et pourquoi réduire la dimension de vos données ?
2. Importer les données depuis le fichier csv

II. Feature Selection



En se basant sur la courbe en cloche gaussienne, il est important de **rejeter** les caractéristiques où la distribution normale des transactions frauduleuses correspond à celles qui ne le sont pas.
Donc **en divisant votre ensemble de données en normaux et anormaux**, il faudra :

- Visualiser la distribution des cas frauduleux et des cas normaux pour **chaque** caractéristique
- Rejeter celles où les 2 distributions correspondent.
    - *Utiliser les librairies Seaborn et Matplotlib.pyplot pour la visualisation*
- Rejeter également la caractéristique « Time » qui n'est pas nécessairement importante.

III. Apprentissage

1. Répartir votre ensemble de données de la façon suivante :

- Un sous-ensemble des données normales (60%) dédiées à l'apprentissage

- Un sous-ensemble des données normales (20%) **et** anormales (50%) dédiées à la validation croisée.
- Un sous-ensemble des données normales (20%) **et** anormales (50%) dédiées au test final.

2. Normaliser vos caractéristiques en utilisant la fonction fit_transform de la classe StandardScaler de Scikit-Learn et afficher les résultats (*Revoir TP 5/6 Exercice 2*).

3. En prenant l'ensemble d'apprentissage :
   a. Estimer les paramètres μ_j *et* ${\sigma^2}_j$ et pour chaque caractéristique

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} {x_j}^{(i)}$$

$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2$$

   b. Construire une fonction qui prend votre moyenne, variance et une donnée d'entrée et vous retourne la probabilité (Distribution normale) :

$$p(x) = \prod_{j=1}^{n} p\left(x_j; \mu_j; {\sigma^2}_j\right) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\left(\frac{(x_j-\mu_j)^2}{2{\sigma_j}^2}\right)}$$

IV. Évaluation

L'évaluation de notre modèle se fait sur l'ensemble de la validation croisée et de la base de test de la façon suivante :

$$y = \begin{cases} 1 \; si \; p(x) \leq \; \varepsilon \\ 0 \; si \; p(x) > \; \varepsilon \end{cases}$$

Où le label 1 veut désigner qu'une anomalie est détectée.
On s'intéresse à déterminer le meilleur seuil ε en appliquant la technique de grid-search sur la validation croisée.

1. Est-ce que la classification est une bonne technique pour évaluer la précision d'un tel modèle ? Pourquoi ?
2. Déterminer le meilleur seuil ε en appliquant un grid-search sur l'ensemble de la validation croisée et généraliser la performance de votre modèle sur la base test.
   a. Pour ε variant de 1e-15 à 1e-10 avec un pas de 5e-10.
   b. Afficher la précision (moyenne des valeurs correctes → +1 pour une prédiction correcte) pour chaque évaluation ;
   c. Choisir l'epsilon avec le score le plus élevé.
4. Déterminer la précision de votre modèle sur la base test.

meritis

Sources :

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project

Please cite the following works:

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon
Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE
Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)
Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier
Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing
Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019
Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019