Notice d'utilisation

Elisa Korn, Imane Salihi et Alexis Lignoux

Produite le 11 novembre 2019

Démonstrateur SVM

Le but de ce démonstrateur est d'expliquer, dans un premier temps, le principe des Machines à Vecteurs de Support (SVM) et dans un deuxième temps, d'appliquer cette méthode sur des données dont la problématique est la détection de la fraude. Enfin, nous comparerons enfin cette méthode à plusieurs benchmarks.

Plan de notre démonstrateur :

Nous commençons, tout d'abord, par une brève introduction sur les problèmes de classification illustrés de graphiques.

Les deux onglets suivants expliquent le principe de la méthode SVM dans deux cas de figure :

- Lorsque l'échantillon est linéairement séparable.
- Lorsqu'il n'est pas linéairement séparable.

Afin de vous présenter la base de données que nous allons utiliser, nous vous présentons :

- Dans l'onglet "Les données", une description et un extrait de la table.
- Dans l'onglet "Traitement des données", la création de nos ensembles d'apprentissage et de test ainsi qu'un rééchantillonnage dû à la faible fréquence d'occurrence de l'événement (évènement rare).
- Dans l'onglet "Visualisation de notre échantillon", vous pouvez représenter des données dans un espace engendré par 2 ou 3 variables. Celles-ci appartiennent à l'échantillon d'apprentissage après rééchantillonnage.

L'onglet "Estimation" vous permet de pouvoir créer vous-même les échantillons test et apprentissage en fixant la proporition d'invidus dans chaque ensemble, puis de procéder au rééchantillonnage, d'estimer votre SVM, pour enfin évaluer les capacités prédictives de votre modèle sur l'échantillon test. Pour cela vous pourrez faire varier les paramètres suivants et ainsi juger de leurs impacts sur les prédictions et la qualité du modèle :

- La proportion d'individus dans l'échantillon d'apprentissage
- La proportion d'individus fraudeurs dans l'échantillon d'apprentissage après rééchantillonnage
- Le type de kernel utilisé (par exemple linéaire ou polynomial)
- Le paramètre de coût (le coût des erreurs de classification)
- Le paramètre d'ajustement γ (paramètre inutile avec un kernel linéaire)
- L'importance d'un individu fraudeur par rapport à un non fraudeur (poids relatif par rapport à un non fraudeur dont le poids vaut 1), ce qui permet de pouvoir contrôler la sensibilité et la spécificité. Augmenter le poids des fraudeurs conduit à augmenter la sensibilité, mais à réduire la spécificité.)

Dans l'onglet suivant, nous avons utilisé la validation croisée avec comme critère le taux d'erreur, afin de déterminer les valeurs optimales des hyper-paramètres et ainsi obtenir le modèle le plus performant.

L'onglet "Comparaison" regroupe l'utilisation de diverses méthodes de Machine Learning sur notre base de données afin d'évaluer et de comparer leurs différents pouvoirs prédictifs avec les SVM. Nous avons fait appel à la Régression Logistique, la méthode KNN, le Boosting et le RandomForest.

Enfin nous présentons dans le dernier onglet, une brève conclusion sur ces comparaisons, avec les forces et les limites de cette méthode.