Sujet d'impact 2017-2018

Copules et dépendance : vers de meilleurs algorithmes de classification

8 septembre 2017

Encadrants

Benjamin Guedj (Inria & Laboratoire Paul Painlevé), benjamin.guedj@inria.fr
John Klein (Lille1, CRIStAL), john.klein@univ-lille1.fr

Thèmes abordés

Machine learning, Apprentissage supervisé, Classification, Agrégation, Vote majoritaire, Copule.

Présentation du sujet

Dans le cadre d'un problème d'apprentissage supervisé, le *data scientist* a une très large palette d'outils à sa disposition : des approches non paramétriques (ex : plus proches voisins), des approches discriminatives (ex : régression logistique), des approches génératives (ex : naïf bayésien), des méthodes déterministes (ex : SVM), etc. Pour certains *datasets*, le choix est tout indiqué mais la plupart du temps, plusieurs approches fournissent des résultats comparables. On peut alors se demander pourquoi ne pas utiliser plusieurs approches puis combiner leurs prédictions par un vote. On obtient alors un comité de classifieurs (*classifier ensemble*) [1].

Dans cet impact, on se propose de construire un *classifier ensemble* reposant sur une méthode d'agrégation des prédictions prenant en compte les performances des classifieurs. Ces performances sont représentées à partir de distributions conditionnelles des classes prédites sachant la classe réelle. Si on note Y la variable aléatoire de la classe réelle d'un exemple $\mathbf{x}$ et f_i la fonction prédicatrice vers laquelle a convergé le classifieur, une telle distribution se note

$$P\left(f_i\left(\mathbf{x}\right)|Y=c\right).$$

Notre objectif est de déterminer la distribution $P\left(Y=c|f_1\left(\mathbf{x}\right),\ldots,f_\ell\left(\mathbf{x}\right)\right)$ qui évalue la probabilité que l'exemple $\mathbf{x}$ appartiennent à la classe c sachant les prédictions fournies par un comité de ℓ classifieurs.

La difficulté est que pour obtenir cette distribution, il est nécessaire d'avoir la distribution jointe

$$P\left(f_1\left(\mathbf{x}\right),\ldots,f_\ell\left(\mathbf{x}\right)|Y=c\right),$$

alors que nous disposons seulement des distributions marginales. Le travail principal de cet impact consistera à tester différents modèles permettant de remonter à la distribution jointe en s'appuyant sur des copules [2]. Les copules permettent de caractériser le niveau de dépendance entre les variables $f_i\left(\mathbf{x}\right)|Y=c$. L'étudiant(e) devra se familiariser avec la notion de copule et le théorème de Sklar. L'implémentation se fera en Python et les classifieurs à agréger seront par exemple choisis parmi ceux disponibles dans la librairie scikit-learn.

Références

[1] Zhang, C., and Ma, Y. (Eds.). (2012). Ensemble machine learning : methods and applications. Springer Science & Business Media.

[2] Genest, C., and Nelehová, J. (2007). A Primer on Copulas for Count Data. ASTIN Bulletin, 37(2), 475-515. doi :10.1017/S0515036100014963