Une méthode d'interprétation de scores

Vincent Lemaire *, Raphaël Féraud *

*France Telecom R&D - 2 avenue Pierre Marzin 22300 Lannion
vincent.lemaire@orange-ft.com

Résumé. Cet article présente une méthode permettant d'interpréter la sortie d'un modèle de classification ou de régression. L'interprétation se base sur l'importance de la variable et l'importance de la valeur de la variable. Cette approche permet d'interpréter la sortie du modèle pour chaque instance.

1 Introduction

Dans les applications de gestion de la relation clients, les scores permettent d'identifier les clients les plus susceptibles de réagir positivement à une campagne marketing. L'interprétation du score apporte alors une information supplémentaire pour améliorer l'efficacité des campagnes marketing. L'utilisation de la méthode présentée[1] ici doit se faire après une étape de sélection de variable qui aura supprimer les variables redondantes pour ne pas risquer de diluer l'interprétation. L'interprétation d'un score est constituée de l'association de l'importance à l'instance (I) d'une variable d'entrée et de l'influence à l'instance d'une variable d'entrée (I_v) présentées ci-dessous.

Notations - Soit V_j : la variable explicative j, X : un vecteur de dimension J, K : le nombre d'instances, X_n : le vecteur représentant l'instance n, X_{nj} : la composante j du vecteur n, F : le modèle, p : la sortie p du modèle, $F^p(X)$: la valeur de la sortie p du modèle pour le vecteur X et $F_j^p(X_n; X_k)$ désigne la sortie p du modèle étant donné le remplacement de la composante j de l'instance X_n par celle de l'instance X_k.

2 Importance à l'instance d'une variable d'entrée

Etant donné[2] le modèle F, l'instance considérée X_n , la variable explicative V_j du modèle et la variable à expliquer p du modèle, on définit la sensibilité du modèle $S(V_j/F, X_n, p)$ par : la moyenne des variations mesurées en sortie du modèle lorsqu'on perturbe l'instance considérée X_n en fonction de la distribution de probabilité de la variable V_j. La variation mesurée, pour l'instance X_n est la différence entre la "vraie sortie" du modèle $F_j(X_n)$ et la "sortie perturbée" du modèle $F_j(X_n, X_k)$.

La sensibilité du modèle pour l'exemple X_n à la variable V_j est alors la moyenne des $||F_j(X_n) - F_j(X_n, X_k)||^2$ sur la distribution de probabilité (distribution empirique observée sur K exemples) de la variable V_j. On a alors : $S(V_j|F, X_n, p) = \frac{1}{K} \sum_{k=1}^{K} ||F_j(X_n) -$

[1] Voir le rapport technique associé sur `perso.rd.francetelecom.fr/lemaire` pour plus de détails.

[2] On définit ici les notions "d'importance (I) d'une variable pour une instance" et "d'influence (I_v) d'une variable pour une instance" pour l'une des variables V_j en entrée du modèle sur l'une des variables de sortie p du modèle. Ces définitions sont rigoureusement les mêmes pour toutes les variables en entrée et en sortie du modèle. On simplifie donc les notations en remplaçant F_j^p par F_j.