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.

$F_j\left(X_n; X_k\right)||^2$. En réalisant cette mesure de sensibilité pour la sortie p mais quelque soit la variable d'entrée[3] j on possède une distribution des sensibilités.

On définit alors l'importance de la variable V_j à l'instance X_n, $I(V_j|F, X_n, p)$, comme étant le rang, o, de la sensibilité du modèle $S(V_j|F, X_n, p)$ parmi l'ensemble des sensibilités $S(V_j|F, X_i, p)\ \forall i, j$. Cette mesure fournit l'importance d'une variable d'entrée pour l'instance X_n relativement à toutes les autres instances et toutes les autres variables. Cette mesure relative permet de se concentrer sur les seules informations pertinentes pour chaque instance. Cette mesure a été testée avec succès pour des problèmes de classification dans (Lemaire et Clérot, 2004) elle est notamment reliée aux travaux de (Breiman, 2001; Féraud et Clérot, 2002)

3 Influence à l'instance d'une variable d'entrée

Une variable peut "tirer vers le haut" (valeur forte) ou "tirer vers le bas" (valeur faible) la sortie du modèle. Pour l'exemple X_n la valeur "naturelle" de la sortie p ' du modèle est par définition $F(X_n)$. La valeur "perturbée" de la sortie du modèle pour l'exemple et en perturbant la variable d'entrée V_j est $F_j(X_n, X_k)$. La distribution des $F_j(X_n, X_k)$ représente ce qu'aurait pu être la valeur de la sortie du modèle pour l'instance X_n si sa variable V_j avait été différente. La position de sa sortie "naturelle" au sein de cette distribution renseigne sur la nature de la valeur de sa variable V_j. On définit alors l'influence de la variable V_j à l'instance X_n, $I_v(V_j|F, X_n, p)$, comme étant le rang, r, de la sortie "naturelle" parmi l'ensemble de ses sorties potentielles. Cette mesure fournit l'influence d'une variable d'entrée pour une instance relativement à toutes les autres valeurs "potentielles" de la variable.

4 Exemple d'utilisation pour un problème de classification

Dans le cas d'un problème de classification à deux classes (-1 ;$+1$) un rang important de I_v dénotera une influence positive par rapport à la classe $+1$ et négative par rapport à la classe -1 (et réciproquement pour un très faible rang de I_v). On obtiendra alors une interprétation de la forme (l'interprétation sera réalisée variable explicative, j, par variable explicative en entrée du modèle) : "Pour l'instance X_n la variable j qui est I importante indique qu'elle est I_v fortement de la classe +1".

Références

Breiman, L. (2001). Random forest. *Machine Learning 45*(1), 5–32.

Féraud, R. et F. Clérot (2002). A methodology to explain neural network classification. *Neural Networks 15*(2), 237–246.

Lemaire, V. et F. Clérot (2004). An input variable importance definition based on empirical data probability and its use in variable selection. In *International Joint Conference on Neural Networks IJCNN*, Volume 2, pp. 1375–1380.

Summary

This paper presents a method allowing to interpret the output of a predictive model of classification or regression. The intepretation is based on the importance of the value of the variable and the importance of the variable. This approach allows to intrepret the output model for every instances.

[3] L'importance n'est pas intrinsèque à une variable mais relativement à l'ensemble des variables. La distribution est donc établie quelle que soit la variable d'entrée et sur l'ensemble des instances que l'on possède.