Classification de campagnes de publicité mobile : Modèle de mélange pour données longitudinales et non gaussiennes

Faustine Bousquet [1,2] & Christian Lavergne [2] & Sophie Lèbre [2]

[1] *TabMo Labs, Montpellier, France faustine.bousquet@tabmo.io*
[2] *IMAG Institut Montpelliérain Alexander Grothendieck, Université de Montpellier*

Résumé. De nombreux enjeux statistiques ont émergé du contexte de la publicité, il s'agit notamment d'afficher une publicité au bon endroit et au bon moment. Dans ce but, plusieurs métriques de performance de diffusion d'une campagne sont mesurées au cours du temps, comme le nombre ou le taux de clics sur une publicité. Un enjeu essentiel est de prédire le taux de clics. Mais, il est important de comprendre au préalable la structure des données volumineuses et hétérogènes dont nous disposons. Pour cela, nous proposons ici une méthode de classification des profils de campagnes publicitaires, i.e. de données longitudinales non gaussiennes, par un mélange de modèles linéaires généralisés (GLM).

Mots-clés. classification, modèle linéaire généralisé, modèle de mélange, données longitudinales, publicité

Abstract. Many statistical challenges arose in the advertisement field, such as displaying an advertisement at the right place and the right time. To achieve this, several performance metrics of airing of an ad were measured over time, for example the number or rate of clicks on an advertisement. A crucial aim is to be able to predict the clicks rate. Nevertheless, it is essential to understand beforehand the voluminous and heterogeneous data structure. Therefore, we introduce a clustering method of publicity campaign timseries, i.e. of non-Gaussian longitudinal data, with a mixture of generalized linear models (GLM).

Keywords. clustering, generalized linear model, mixture model, longitudinal data, advertising

1 Introduction

Le domaine de la publicité, et plus particulièrement la publicité en ligne, a été révolutionné par l'arrivée du RTB (Real-Time Bidding). Ce processus d'achat et de vente permet de mettre en relation des annonceurs et des éditeurs publicitaires en temps réel afin de cibler au mieux une audience. Plus précisément, le RTB permet à un éditeur de proposer un emplacement publicitaire en ligne dès qu'il est disponible à travers une enchère à un ensemble d'annonceurs. De plus, l'éditeur fournit un ensemble d'informations au sujet de l'emplacement publicitaire et de l'utilisateur afin que l'annonceur puisse décider s'il est intéressé par cette enchère. L'intérêt de mobinautes pour ces publicités est mesuré par certains métriques. La plus connue d'entre elles est le Click Through Rate (CTR) qui

correspond à la proportion de clics recensés par rapport au nombre de fois que la publicité est diffusée (appelé nombre d'impressions). Ainsi, de nombreux enjeux statistiques ont emergé du contexte de la publicité, notamment le dans le contexte des évènements rares (peu de clics) étudiés par Wang et al (2010). Afin de comprendre la structure de ces données volumineuses et hétérogènes, nous proposons ici une méthode de classification des profils de campagnes publicitaires. Notre objectif est d'obtenir une classification des individus (ici de campagnes) selon l'évolution d'une métrique observée au cours du temps.

2 Données

Nous avons à disposition un très gros volume de données issue de notre plateforme d'achat d'emplacement publicitaire [1] avec environ 1 million de propositions d'enchère arrivant sur la plateforme chaque seconde. Dans cette étude, nous avons extrait un jeu de données contenant **400** campagnes sur la période allant du **06/09/18** au **31/10/18**. Ces données ont été récoltées et pré-traitées afin de pouvoir étudier l'évolution du nombre d'impressions d'une publicité sur un téléphone portable, du nombre de clics et du CTR au cours de la durée de diffusion de chaque campagne publicitaire. Les valeurs du CTR sont très différentes d'une campagnes à l'autre tout comme la durée d'une campagne qui peut varier de quelques jours à plusieurs semaines. Le ratio existant entre le nombre de publicités cliquées et non cliquées est très déséquilibré avec une valeur médiane aux alentours de 1 clic pour 1000 impressions.

3 Modèles de Mélange de GLM

Les métriques d'intérêt sont le nombre et le taux de clics. Il s'agit de classer des données longitudinales non gaussiennes. Nous proposons un modèle de mélange de GLM.

3.1 Modèle linéaire généralisé

Le nombre d'impressions et de clics sont mesurés et agrégés toutes les heures au cours de la diffusion d'une campagne. Nous avons choisi de regrouper ces heures en $H = 5$ plages horaires sur une journée de 24h : 00h-7h, 7h-12h, 12-14h, 14h-18h, 18h-00h. Nous observons généralement plusieurs mesures de chaque métrique par plage horaire. Notons Y_{cjht} la métrique Y mesurée lors de la répétition t, de la plage horaire h, du jour j de la campagne c. Nous considérons que chaque variable Y_{cjht} suit une loi appartenant à la

[1] http://tabmo.io/

famille exponentielle dont la fonction de densité est définie (McCullagh, Nelder (1989)) :

$$\forall c = 1, \ldots, C, \forall j = 1, \ldots, J_c, \forall h = 1, \ldots, H, \forall t = 1, \ldots, T_{jh},$$
$$f_{Y_{cjht}}(y_{cjht}, \theta_{cjht}, \psi) = \exp\left(\frac{y_{cjht}\theta_{cjht} - b(\theta_{cjht})}{a_{cjht}(\psi)} + d(y_{cjht}, \psi)\right) \quad (1)$$

où θ_{cjht} est le paramètre canonique, ψ le paramètre de dispersion, b et d des fonctions spécifiques à chaque distribution, a_{cjht} une fonction telle que $a_{cjht}(\psi) = \frac{\psi}{\omega_{cjht}}$ avec ω_{cjht} correspondant au poids de chaque observation. Chaque campagne c a une durée de diffusion (J_c jours) qui lui est propre et il peut y avoir des données manquantes au cours de celle-ci.

Modèle binomial pour le CTR Le CTR représente un pourcentage. Nous faisons donc l'hypothèse d'une distribution binomiale de paramètre de paramètre $(n_{cjht}, p_{hs(c,j)})$ avec n_{cjht} le nombre d'impressions observé et $p_{hs(c,j)}$ la probabilité de clics dans la plage horaire h, du jour de la semaine $s(c, j)$ du j-ième jour de la campagne c. La métrique Y_{cjht} représente le nombre de clics. Pour la suite, nous étudions Y_{cjht}/n_{cjht} qui représente le taux de clics. Pour cette loi, le paramètre canonique θ_{cjht} correspond à $\log\left(\frac{p_{hs(c,j)}}{1-p_{hs(c,j)}}\right)$. Les fonctions a, b et d sont quant à elles définies comme suit :

$$a_{cjht}(\psi) = \frac{1}{n_{cjht}},\ b(\theta_{cjht}) = \log\left(1 + \exp\theta_{cjht}\right) \text{ et } d(y_{cjht}, \psi_{cjht}) = \log\binom{n_{cjht}}{y_{cjht}} \quad (2)$$

Nous introduisons une fonction de lien canonique de type logit pour lier le vecteur des paramètres β et l'espérance du taux de clics $\frac{Y_{cjht}}{n_{cjht}}$. Ainsi, le modèle linéaire généralisé pour la loi binomiale mis en place pour notre problématique est décrit par l'équation :

$$\log\left(\frac{E(Y_{cjht}/n_{cjht})}{1 - E(Y_{cjht}/n_{cjht})}\right) = \beta_0 + \beta_h^H + \beta_{s(c,j)}^S \quad (3)$$

Les variables explicatives du modèle représentent un effet temps. Il y a d'une part la plage horaire β_h^H (5 modalités) et d'autre part, le jour de la semaine $\beta_{s(c,j)}^S$ (7 modalités). β_0 correspond à la constante du modèle.

3.2 Inférence du modèle de mélange

L'objectif est d'obtenir un modèle de mélange associé à ce modèle linéaire généralisé. Nous supposons que les C campagnes sont issues de K populations différentes. On met en place un algorithme EM (Dempster (1977)) pour estimer les paramètres du modèle. Notons Z la variable aléatoire telle que $Z_{kc} = 1$ si la campagne c appartient au cluster k (0 sinon) et ϕ l'ensemble des paramètres (β, λ) du modèle à estimer où λ correspond aux proportions du mélange.

Etape E : Afin de calculer $Q(\phi|\phi^{(m)}) = E(\log Ln(Y, Z; \beta, \lambda)|Y = y, \beta^{(m)}, \lambda^{(m)})$ l'espérance de la log-vraisemblance associée aux données complètes conditionnellement aux données observées y, nous mettons à jour la probabilité π_{kc} qu'une campagne c appartienne à la population k sachant y_c, le vecteur des observations d'une campagne c.

Etape M : Notre objectif est de maximiser $Q(\phi|\phi^{(m)})$. Or, comme nous sommes dans le cas d'un modèle linéaire généralisé, l'estimation des β_k pour l'étape M ne possède pas de solution explicite. On utilise donc l'algorithme des scores de Fisher (McCullagh et Nelder (1989)) pour l'estimation de ce paramètre :

$$\beta_k^{(m+1)} = \beta_k^{(m)} - \left(E\left[\frac{\partial^2 Q(\phi|\phi^{(m)})}{\partial\beta_k \partial\beta_{k'}} \right] \right)^{-1} \frac{\partial Q(\phi|\phi^{(m)})}{\partial\beta_k} \tag{4}$$

Cet algorithme se base sur celui de Newton-Raphson, dans lequel la direction de recherche de la nouvelle valeur $\left(-\frac{\partial^2 B}{\partial^2 \beta_k}\right)$ est remplacée par son espérance. On retrouve donc la formule de l'Information de Fisher. Pour chaque étape M, on répète quelques itérations de l'algorithme des scores de fisher.

Modèle de mélange de loi Binomiale pour le CTR : Afin de répondre à notre volonté de regrouper des campagnes publicitaires ayant des profils de CTR (Y_c/n_c) similaires, nous considérons le modèle linéaire généralisé défini par les équations (1,2,3). On obtient une écriture matricielle de l'actualisation des paramètres β_k :

$$\beta_k^{(m+1)} = \left(\sum_{c=1}^{C} \pi_{kc} M_c^t W^{-1}_{c\beta_k^{(m)}} M_c \right)^{-1} \sum_{c=1}^{C} \pi_{kc} M_c^t W^{-1}_{c\beta_k^{(m)}} \left[M_c \beta_k^{(m)} + \frac{\partial \eta_{kc}}{\partial \mu_k} \left(\frac{Y_c}{n_c} - \mu_k \right) \right] \tag{5}$$

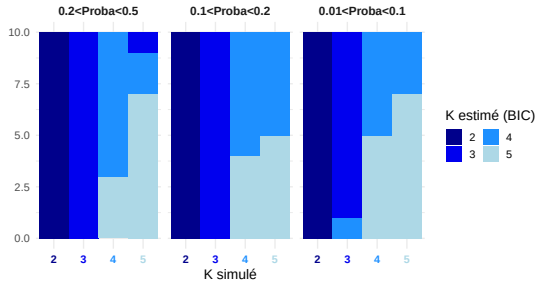
avec M_c la matrice de design des données de la campagne c, μ_k l'espérance du taux de clics dans le cluster k et les matrices diagonales suivantes : $W_{c\beta_k} = diag\left(\frac{1}{n_{cjht}} \frac{(1+\exp M_{cjht}\beta_k)^2}{\exp M_{cjht}\beta_k} \right)_{cjht}$ et $\frac{\partial \eta_k}{\partial \mu_k} = diag\left(\frac{(1+\exp M_{cjht}\beta_k)^2}{\exp M_{cjht}\beta_k} \right)_{cjht}$.

4 Simulations

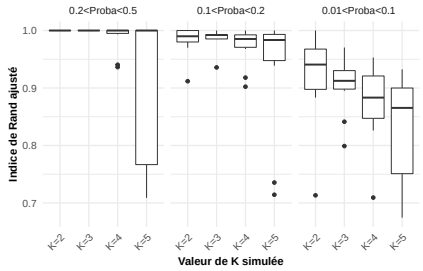
Nous effectuons une étude de simulation en deux étapes : dans la première, nous cherchons à retrouver la bonne partition lorsqu'on connait le modèle. Dans la seconde étape, l'objectif est de retrouver à la fois le bon modèle et la bonne partition.

Nous évaluons tout d'abord la capacité de notre approche à retrouver la bonne partition dans le cas où le modèle linéaire généralisé est connu. Nous avons simulé des taux de clics pour 400 campagnes de publicité, réparties uniformément en K=2 à 5 clusters. Au sein de chaque cluster, le taux de clics est simulé selon un modèle linéaire généralisé pour la loi binomiale avec 2 effets : jour et plage horaire (Equation 3). L'espérance du taux de clics est choisi dans 3 intervalles différents ($[0.2, 0.5], [0.1, 0.2]$, et $[0.01, 0.1]$) de façon à évaluer l'impact d'un faible taux de clics. 10 simulations ont été réalisées dans chaque cas. Le modèle retenu est celui minimisant le critère BIC (Schwarz, 1978). Les résultats sont présentés en Figure (1a). Le nombre de clusters est correctement retrouvé pour 2 et 3 clusters simulés, quelle que soit l'espérance du taux de clics. A partir de 4 clusters simulés, le nombre de clusters n'est pas toujours estimé correctement et ce d'autant plus que l'espérance du taux de clics est faible. C'est un comportement attendu du modèle

puisqu'il y a moins de campagnes participant à l'estimation des paramètres dans chaque classe. L'indice de Rand ajusté (ou ARI, Rand (1971)), calculé sur la Figure (1b), permet lorsque le nombre de clusters est correctement estimé, d'évaluer la similarité entre partition estimée et partition simulée. Un indice de 1 indique que les deux partitions sont identiques. Jusqu'à 4 clusters simulés, pour une espérance du taux de clics supérieure à 0.1, les partitions estimées sont très proches des partitions simulées. En revanche, la qualité d'estimation de la partition se dégrade pour une espérance du taux de clics inférieure à 0.1, même pour un petit nombre de clusters. D'après ces premières simulations, un jeu de données de 400 campagnes permet d'identifier correctement jusqu'à 3 à 4 clusters, pour une espérance supérieure à 0.1 dans le cas d'un modèle binomial défini par 11 paramètres libres (Equation (3)).



(a) Comparaison du nombre de cluster simulé et estimé par BIC

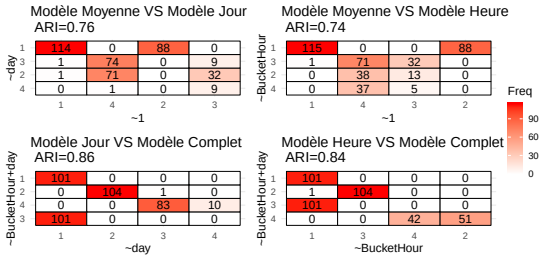


(b) Boxplot des indices de Rand ajustés

Figure 1: Sélection de la partition lorsque les variables du modèle GLM sont connues.

	2	3	4	5
Effet Jour	0	0	0	0
Effet Plage Horaire	0	0	0	0
Effet Constante	0	0	0	0
Effets Jour + Plage Horaire	0	0	7	1

(a) Nombre de clusters et modèles estimés pour 8 simulations (K=4, modèle à 2 effets Jour+Plage Horaire)



(b) Matrices de confusions associées

Figure 2: Sélection de la partition et des variables du modèle GLM.

Nous évaluons maintenant la capacité de notre approche à retrouver la bonne partition et le bon modèle simultanément. Nous avons simulé des taux de clics pour 400 campagnes de publicité, réparties uniformément en K=2 à 5 clusters. Le taux de clics est simulé selon un modèle linéaire généralisé pour la loi binomiale avec différents paramètres : 2 effets (jour et plage horaire (Equation 3)), un seul effet jour, un seul effet plage horaire, l'effet de la constante uniquement. L'espérance du taux de clics est fixée dans l'intervalle $[0.2, 0.5]$. Pour chacune des 8 simulations effectuées dans chaque cas, le modèle et la partition retenus sont la paire minimisant le critère BIC. Les résultats pour le cas (K=4

et modèle à 2 effets) sont présentés en Figure (2a). La bonne partition et le bon modèle sont retrouvés dans 7 cas sur 8, avec une erreur sur la partition pour le dernier cas. Les résultats sont comparables dans les autres cas. A titre illustratif, les matrices de confusion représentées Figure (2b) permettent de comparer les partitions estimées à partir de différents modèles pour une simulation K=4 et modèle à 2 effets. D'après les matrices du haut, l'ajout des variables Jour et Plage Horaire impacte effectivement la partition. En regardant les matrices en bas de la Figure (2b), les 2 variables prises séparémment versus le modèle complet indiquent que l'ajout de la 2ème variable impacte également la classification.. Avec ces simulations, nous pouvons vérifier que le modèle retrouve la bonne partition, le bon modèle mais aussi évaluer l'impact de l'ajout des variables sur le clustering.

5 Conclusion et perspectives

Ce modèle de mélange de GLM se base uniquement sur un aspect temporel avec les variables plage horaire et jour de la semaine. Mais il existe un grand nombre de variables disponibles comme le type d'OS (iOS, Android) ou le type de support (une application ou un site mobile). La suite de notre travail sera donc d'enrichir le modèle avec de nouvelles variables, les interactions entre ces variables, tout en controlant la complexité du vecteur des paramètres β. Nous avons initié une première approche avec l'étude des matrices de confusion pour mesurer l'impact de l'ajout d'une variable. L'état de l'art propose un certain nombre de méthodes pour faire de la sélection de variables dans le modèle utilisé pour la classification. La plus populaire est sans doute celle de selection de modèle de Raftery et Dean (2006), qui propose un algorithme glouton pour optimiser un facteur de Bayes approché. L'ajout de nouvelles variables va permettre d'affiner le modèle et mieux interpreter les groupes de campagnes publicitaires obtenues.

Bibliographie

McCullagh, P., et Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*

Wang, X., Li, W., Cui, Y., Zhang, R., and Mao, J. (2010). Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6(2)*, 461-464.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846-850.

Raftery, A. E., and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473), 168-178.