SUJET DE « PROJET D'INTEGRATION » AU RESEARCH LAB VEKIA

RANKING HIERARCHIQUE

PRESENTATION DU SUJET

Dans le contexte d'un apprentissage supervisé, une technique possible pour numériser une ou plusieurs variable(s) catégorielle(s) est de remplacer chacune des instances de cette(ces) variable(s) par la moyenne de la variable cible à prédire, calculées sur cette instance. Nous appelons ce procédé « **ranking** ».

Par exemple, si l'on souhaite numériser un identifiant de produit (item_id) on pourra chercher à estimer la moyenne de la variable cible « group by item_id », ou « group by store_id, item_id » par exemple.

Cependant des problèmes apparaissent lorsque le nombre d'instances dans un groupe est très faible, voire nul (cas d'un nouvel article apparaissant uniquement dans le test set) : la moyenne empirique est alors un estimateur avec une grande variance. Il peut être tentant d'utiliser des données à l'extérieur du groupe en question afin d'enrichir l'estimation : on gagnera de la variance en introduisant du biais. Par exemple pour estimer la moyenne des ventes d'un certain article dans un certain magasin, on peut tenter d'utiliser les données de vente de ce même article mais dans un magasin différent. On dira qu'on a calculé la moyenne à la maille « store_id, item_id », avec un **backup** à la maille « store_id ».

D'une manière plus générale, on voudrait pouvoir estimer une moyenne à une maille fine (par exemple a,b,c) et utiliser éventuellement plusieurs backups hiérarchiques (par exemple (a,b), (a,c), (a), (b), (c)), tirant ainsi parti de la stratification des données pour rendre l'estimation meilleure et plus robuste.

Dans la littérature statistique, ces estimateurs sont connus sous le nom de « **shrinkage estimators** » (https://en.wikipedia.org/wiki/Shrinkage_estimator), dont l'estimateur de « James-Stein » est historiquement l'un des premiers exemples. De nombreux estimateurs de ce genre ont été proposés depuis, mais aucun à notre connaissance ne traite en toute généralité du problème posé ici :

1. Comment utiliser plus d'un niveau de backup ?
2. Pour des backups de même niveau, par exemple (a,b) et (a,c), comment choisir automatiquement le meilleur ou tirer parti des deux ?

TRAVAIL ET RESULTATS ATTENDUS

Après un travail de bibliographie et une première discussion avec les équipes VEKIA, permettant de rentrer dans le sujet, l'étudiant proposera une ou plusieurs solutions au problème de ranking hiérarchique.

Il proposera une implémentation de la (des) méthode(s) en R ou Python/Pandas. Il disposera de données de ventes réelles afin de les mettre en œuvre et procéder à une comparaison de leurs performances.

CONTACT ENCADRANT

- Alexandre Gerussi
- agerussi@vekia.fr