UNIVERSITÉ DE SHERBROOKE
Faculté des sciences
2500, boul. de l'Université
Sherbrooke (Québec) J1K 2R1
CANADA
Département d'informatique
+1 (819) 821-7030
+1 (819) 821-8200
http://www.usherbrooke.ca/informatique/

An intelligent shopping list based on the application of partitioning and machine learning algorithms

Présentation générale du laboratoire

Le but du TP2 est de pratiquer l'exploration de données :

- Visualisation de données
- Analyse de corrélation entre attributs
- Réduction de dimension
- Choix d'une mesure de similarité entre objets

Ce devoir est à faire en équipe. Il devra être complété avant le vendredi 5 avril 2024 avant 23h59. Vous devez remettre, sur Moodle, un fichier Ipython notebook (nommé nomEquipe_tp2.ipynb et les données nettoyées – au format souhaité) contenant votre rapport et vos scripts Python pour ce devoir.

Description des tâches à réaliser :

On vous fournit un ensemble de données stockées dans 5 fichiers au format csv.

1. aisles.csv,
2. departments.csv,
3. order_products__prior_specials.csv,
4. orders_distance_stores_softmax.csv, et
5. products.csv.

Ces données proviennent originalement d'Instacart et modifié dans l'étude de Tahiri *et al.* (2018) intitule « An intelligent shopping list based on the application of partitioning and machine learning algorithms » [1], voir lien 1 et lien 2 pour plus de détails. Les données ont été modifié pour la composition de ce travail pratique.

La composition de l'ensemble des données pour chacun des 5 fichiers :

1. orders_distance_stores_softmax (3.4m lignes, 206k utilisateurs) :
   - user_id: identification du consommateur.
   - store_id: identication du magasin.
   - distance_id: distance euclidienne entre différents magasins et l'utilisateur.
   - order_id: identification de l'ordre.
   - eval_set: à quel ensemble d'évaluation appartient cet ordre (voir SET décrit ci-dessous).
   - order_number: le numéro de séquence de l'ordre pour cet utilisateur (1 = premier, n = nième).
   - order_dow: le jour de la semaine où la commande a été passée.
   - order_hour_of_day: l'heure de la journée à laquelle la commande a été passée.
   - days_since_prior: jours depuis la dernière commande, plafonnés à 30 (avec NAs pour order_number = 1).

2. orders_products__prior_specials (3.4m lignes, 206k utilisateurs) :

[1] Tahiri, N., Mazoure, B. and Makarenkov, V., 2019. An intelligent shopping list based on the application of partitioning and machine learning algorithms. In PROC. OF THE 18th PYTHON IN SCIENCE CONF.(SCIPY 2019) Pp (pp. 85-92).

- order_id: identification de l'ordre.
- product_id: numéro unique du produit.
- add_to_cart_order: indique la commande dans lequel le produit a été ajouté.
- reordered: la nouvelle commande est égale à 1 si le produit a été commandé par cet utilisateur dans le passé, 0 sinon.
- special: est le pourcentage, par intervalle, appliqué au prix du produit au moment de l'achat.

3. products (50k lignes):
   - product_id: identication du produit.
   - product_name: nom du produit.
   - aisle_id: clé étrangére.
   - department_id: clé étrangére.

4. aisles (134 lignes):
   - aisle_id: identifiant de l'allée
   - aisle: nom de l'allée

5. deptartments (21 lignes):
   - department_id: identication du département
   - department: nom du département

Dans le cadre de ce travail pratique, vous n'utiliserez que les deux premiers fichiers à savoir (orders_distance_stores_softmax et orders_products__prior_specials). Les données sont segmentées en 2 classes (« 0 » et « 1 ») pour l'attribut reorderred que l'on souhaite prédire.

L'objectif du TP est de prédire la probabilité que le produit i soit inclus dans le panier order_{t+1} de u en fonction de l'utilisateur u et de l'historique d'achat de l'utilisateur (order_{t-h}:t , $h > 0$).

1. Analyse des données :

Cet ensemble de données anonymes contient un échantillon de plus de 3 millions de commandes de produits d'épicerie provenant de plus de 200 000 utilisateurs d'Instacart. Pour chaque utilisateur, nous fournissons entre 4 et 100 de ses commandes, avec la séquence des produits achetés dans chaque commande. Nous fournissons également la semaine et l'heure à laquelle la commande a été passée, ainsi qu'une mesure relative du temps entre les commandes.

Toutes les caractéristiques utilisées dans l'étude sont présentées ci-dessous.

(a) Dans cette étude, vous avez trop de données, mais le nombre d'attributs est raisonnable. Comparativement au premier TP, vous allez réduire uniquement les données (lignes) pour ne conserver que 5%, ceci afin de réduire le temps de calcul. Attention de prendre des données aussi diverses que possible.

(b) Préparer vos données. Vous allez maintenant diviser vos données en deux groupes (train et test) pour des proportions de 80 et 20 respectivement. Attention, la même remarque que le

[1] Tahiri, N., Mazoure, B. and Makarenkov, V., 2019. An intelligent shopping list based on the application of partitioning and machine learning algorithms. In PROC. OF THE 18th PYTHON IN SCIENCE CONF.(SCIPY 2019) Pp (pp. 85-92).

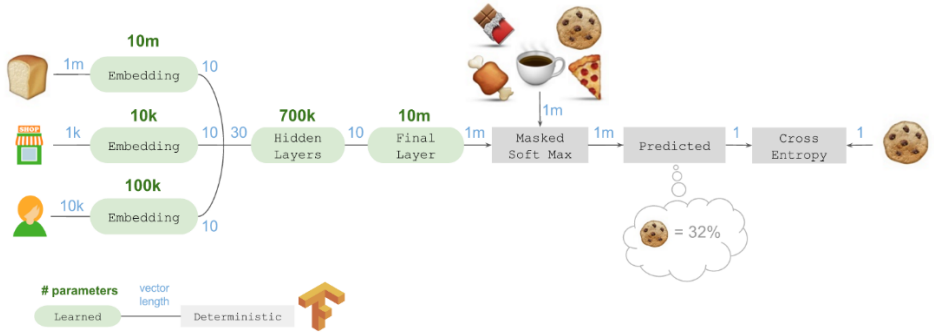
point précédent.

(c) Veuillez visualiser vos données, idéalement en mettant des couleurs différentes pour le train et le test. Veuillez analyser vos visualisations et interprétez.

2. Choix du modèle de prédiction :

(a) Veuillez implémentez en python un modèle de prédiction (voir prochain cours – 2024/03/15 et 2024/03/22).

- CNN (réseau de neurones convolutifs) avec 2 couches internes cachées avec la fonction ReLU et finalement la couche de sortie avec deux neurones (0 et 1) utilisant la fonction soft max.



Source : [fast ai]

- SVM (machines à vecteurs de support) en recherchant de l'hyperplan optimal.

(b) Veuillez implémentez le score F_1 pour connaître la performance de vos deux modèles en fonction de vos données.

3. Conclusion :

Veuillez tirer des conclusions de votre études.

(a) Avez-vous analysé le comportement des ventes de produits. Est-ce qu'il a des comportements pouvant se déduire du comportement des consommateurs (par exemple : achète chips avec salsa).

(b) Veuillez identifier les hyperparamètres que vous avez utilisés et/ou indiqués par l'énoncé même.

(c) Le choix du modèle était-il judicieux. Veuillez élaborer votre réponse.

(d) Veuillez analyser vos résultats du score F_1. Que pouvez-vous en dire ?

Remise du travail

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur Moodle. Chargez votre fichier nomEquipe_tp2.ipynb, fichier des données nettoyés et soumettez-le. Indiquez bien les noms des deux membres de l'équipe dans le fichier. Ne faites qu'une seule soumission par équipe.

Bon travail ☺

[1] Tahiri, N., Mazoure, B. and Makarenkov, V., 2019. An intelligent shopping list based on the application of partitioning and machine learning algorithms. In PROC. OF THE 18th PYTHON IN SCIENCE CONF.(SCIPY 2019) Pp (pp. 85-92).