Franck Lin
flin@hpsj.fr

Examen : Analyse de Données 2

ADD2
18/03/2020

Date limite de rendu : **26/04/2020**

Travail à faire par groupe de 1,2 ou 3.
Format de fichier accepté : pynb, doc, pdf
Que ce soit en notebook ou doc/pdf, joindre le code dans une rubrique Annexe à la fin. Dans vos réponses, vous ferez référence au code utilisé à l'aide d'une numérotation. Des points sont attribués pour la lisibilité du code, l'orthographe et la présentation générale.

Exercice 1 (30%)

C'est un exercice de classification d'images. L'objectif est d'abord de réduire la dimension avec l'ACP puis de classifier les images. Vous pouvez utilisez l'ACP de scikit-learn. Le K plus proche voisin est à implémenter sans l'utilisation de packages (excepté numpy,pandas).

1. Décrivez les 4 datasets. Représentez graphiquement le visage moyen (moyenne de toutes les lignes) de l'échantillon train.
2. Appliquez une ACP pour réduire la dimension du dataset pour k=5, 10 et 50.
3. Choisissez un visage et représentez le graphiquement pour k=5,10,50.
4. L'ACP est plutôt longue lorsque la quantité de colonnes est très très grande, proposez une solution.
5. En utilisant le dataset réduit par ACP, implémentez un K plus proche voisin, pour pouvoir classifier les images du test et calculer la précision et la sensibilité de votre algorithme.
6. Offrez des recommandations pour améliorer le modèle.

Exercice 2 (50%)

Le fichier train.csv contient un échantillon de produits appartenant aux 20 plus grandes catégories d'un catalogue de e-commerce. Pour chaque produit on a les champs:

- category_id : identifiant de la catégorie du produit (entier entre 0 et 19)
- category : nom de la catégorie du produit
- title : titre du produit description : description du produit

1. Ecrivez un algorithme de classification supervisée qui prévoit la catégorie d'un produit à partir de son titre et de sa description. Vous mettrez de côté un ensemble de validation et donnerez le taux de classification correcte sur ledit ensemble. Vous appliquerez l'algorithme sur les données de test contenues dans le fichier test.csv

2. Le fichier purchases.csv contient la liste des achats réalisés le 1er septembre 2017. Les champs sont les suivants : time : date et heure d'achat du produit amount : prix en euros du produit acheté On note le nombre d'heures écoulées de la journée et le prix d'achat, considérés comme des variables aléatoires dont le fichier ci-dessus donne un échantillon. Ecrivez un programme pour estimer la fonction de densité de probabilité conjointe, par la méthode d'estimation par noyau. Illustrez la fonction sous forme de carte de chaleur.

Exercice 3 (20%)

En prenant un exemple de jeu de données d'au moins 1000 observations, proposez une étude de votre choix incorporant des notions de cours et des notions extérieures si vous le souhaitez. Les travaux originaux, non plagiés, seront valorisés.