HELHa
Haute École Louvain en Hainaut

KNN sur des voitures

Phase initiale

Durant cette première phase, vous préparerez vos données afin de les traiter ultérieurement avec l'algorithme du KNN.

Tout d'abord, vous chargerez le fichier en mémoire centrale au sein d'un référent nommé "cars". Juste après, vous afficherez les 15 premières lignes de la base de données.

Ensuite, vous créerez une variable nommée « features_name » qui contiendra un tableau qui exposera les noms suivants : "doornumber", "horsepower", "carheight" et "carwidth".

À ces variables-là, vous allez devoir ajouter la colonne "fueltype". Attention, il s'agit d'une colonne qui contient des chaînes de caractères représentant des catégories. Pour vous aider dans l'accomplissement de cette tâche, vous pouvez utiliser ce lien. Exemple : gaz ou diesel. Il faudra trouver un moyen pour pouvoir les exploiter. Si vous ne parvenez pas à résoudre ce point-là, ignorez cette étape (une pénalité sera de mise).

Lorsque vous aurez terminé la précédente étape de transformation, créez une variable "output_name" qui contiendra juste la valeur "price".

Grâce aux variables "features_name" et "output_name", créez les variables X et y.

Phase de création du modèle

Vous produirez une instance d'un " KNeighborsRegressor". L'algorithme sera fixé à "brute". Concernant le nombre de voisins, ne fournissez aucune valeur.

Toujours concernant le nombre de voisins, vous emploierez une instance de "GridSearchCV". Concernant sa création, voici toutes les données dont vous aurez besoin :

- "cv" : cela sera une instance de KFold ("n_splits = 5, shuffle = True, random_state = 1"),
- "scoring" : MSE,
- "param_grid" : un nombre de voisins pouvant aller de 5 jusqu'à 50.

Entrainez le meilleur modèle avec vos X et y.

Phase de prédiction

Prédisez le prix des voitures suivantes :

- Une voiture possédant 150 chevaux, 5 portes, 54 de hauteur, 70 de longueur et qui fonctionne au diesel.
- Une voiture possédant 114 chevaux, 3 portes, 50 de hauteur, 60 de longueur et qui fonctionne au gaz.

Phase d'argumentation

Répondez aux questions suivantes :

- À quoi correspondent les variables "features" et "output" ?
- Que signifie le paramètre "cv" lorsque l'on crée une instance de "GridSearchCV" ? En quoi intervient-il ?
- Avons-nous testé notre modèle ? Justifiez votre réponse.

Phase d'approfondissement

Nous sommes parvenus à identifier d'une manière automatique le nombre de voisins.

Par contre, l'énoncé indique les paramètres sur lesquels l'algorithme doit se baser pour réaliser des prédictions. Existe-t-il, dans la libraire "sklearn", un moyen de les trouver automatiquement ? Si oui, trouvez le nom de ce mécanisme.