Introduction à la Science des données

Travail pratique 05 – Régression linéaire

Professeurs: Carlos Peña et Stephan Robert
Assistants: Thibault Schowing, Régis Houssou
Emails : prenom.nom@heig-vd.ch

Objectifs:

- Appliquer l'algorithme de régression linéaire simple pour créer des modèles de régression
- Utiliser la méthode de validation hold-out pour évaluer un modèle de régression
- Utiliser les bibliothèques d'apprentissage automatique pour construire des modèles de régression multiple
- Évaluer les modèles et analyser les résultats

1. Régression linéaire simple

1) Installez la bibliothèque sklearn
2) Créez un notebook pour lire la base de données «Boston housing dataset » et générez un dataframe.

   Aide : *from sklearn.datasets import load_boston*
   boston_dataset = load_boston()

3) Se familiariser avec la base de données (p.ex., nombre d'observations, des classes, d'attributs, statistiques des attributs, données manquantes). Générez un histogramme de la distribution de la variable indépendente.
4) Visualiser des boxplots des variables MEDV, RM, LSTAT, RAD, DIS et AGE. Commentez vos observations.
5) Créez des modèles en utilisant l'algorithme de régression linéaire simple vue en cours pour estimer la valeur MEDV à partir des variables RM, RAD et LSTAT, chaque fois utilisant seulement une des variables.
   Programmez les fonctions pour faire une validation hold-out (e.g., 80% train + 20% test x 100 fois) et pour calculer le coefficient de détermination afin d'évaluer la performance moyenne du modèle. Commentez vos résultats.
6) Explorez les hyper-paramètres : learning_rate et nombre d'epochs d'entraînement. Calculez la performance moyenne (c.a.d., la moyenne de x100 fois le coefficient de détermination calculé sur le test set) du meilleur modèle trouvé par validation hold-out. Indiquez la valeur des hyper-paramètres qui donnent les meilleurs résultats et la meilleure performance trouvée. Générez un scatter-plot affichant les données (variable indépendante, variable dépendante)

et la ligne droite trouvée par régression linéaire. (Note : la variable dépendante est MEDV).

2. Régression linéaire multiple

1) Utilisez la bibliothèque sklearn pour entraîner un modèle de régression linéaire multiple qui exploite plusieurs variables indépendantes (p.ex. RM et LSTAT) de la base de données « Boston Housing dataset ».
2) Utilisez la bibliothèque sklearn pour préparer les données pour faire une validation hold-out ainsi que pour le calcul du coefficient de détermination (r2_score) des modèles.
3) Présentez les résultats de vos expériences (points 1 et 2) avec la base de données proposée et commentez ceux-ci.
4) Générez un scatter-plot (valeur prédite, variable dépendante) et une ligne droite correspondant à l'équation valeur prédite = valeur dépendante (Note : la variable dépendante est MEDV).

Rapport

Préparer un fichier compressé nommé nom1_nom2_ISD_TP5.zip intégrant UN SEUL notebook et le télécharger sur Cyberlearn avant la date limite indiquée. Veuillez bien séparer les différentes parties dans le notebook et veuillez également intégrer vos commentaires par rapport aux résultats obtenus dans les différents points.