TP 05

Exercice 1 : Une première regression linéaire

La régression linéaire est sans aucun doute l'algorithme de *Machine Learning* (ML) le plus simple qui soit. Comme tous les algorithmes de ML, il est instancié à partir d'exemples , et permet ensuite de généraliser à de nouvelles données.

Ce modèle est suffisamment simple pour qu'une solution mathématique immédiate existe dans tous les cas. Il "suffit" de calibrer le modèle et non de l'entraîner comme avec les réseaux de neurones ou les approches basées sur l'apprentissage par renforcement.

En guise d'illustration on prendra ici un exemple trivial avec quelques données, qui pourraient par exemple correspondre à prédiction du poids d'un individu à partir de sa taille, la température en fonction de l'altitude, ou les ventes d'un produit en fonction de sa qualité.

Toutes les méthodes du framework que nous allons utiliser, `ScikitLearn`, fonctionnent avec des `np.array`. La technique d'utilisation est toujours la même :

1. récupérer les données sous forme de `np.array`
2. créer un modèle
3. calibrer ce modèle avec la méthode `fit(...)`
4. généraliser (faire des prédictions) avec la méthode `predict(...)`

Q1. Mettez en forme les observations (x_i, y_i) suivantes afin qu'elles soient utilisables : (5, 5), (15,20), (25, 14), (35, 32), (45, 22), (55, 38). Affichez-les comme nuage de points sur une grille.

Q2. Créez le modèle de régression qui représente le mieux ces observations, c'est-à-dire la droite minimisant l'écart avec les points.

Q3. Quel est l'équation de la droite obtenue ainsi ? Autrement dit, quelles sont les valeurs de la pente et de l'ordonnée à l'origine ?

Q4. Puisque nous disposons d'un modèle, effectuons quelques prédictions. Quelles sont les valeurs prédites pour $x_1 = 20$ et $x_2 = 40$?

Q5. Tracez la droite de régression ainsi que le nuage de points. Vérifiez que vous retrouvez bien les valeurs prédites pour les nouvelles observations données précédemment.

Q6. Quantifiez la qualité du modèle en affichant les scores R2, MSE, RMSE et MAE.

Exercice 2 : Le cas polynomial

La régression polynomiale se gère de la même manière que la régression linéaire, mais avec une étape supplémentaire. Il est en effet nécessaire de transformer le tableau des entrées pour inclure des termes non linéaires.

Q1. Mettez en forme les observations (x_i, y_i) suivantes afin qu'elles soient utilisables : (5, 15), (15, 11), (25, 2), (35, 8), (45, 25), (55, 32). Affichez-les comme nuage de points sur une grille.

Q2. Créez le modèle de régression et calibrez-le avec les données précédentes.

Q3. Puisque nous disposons d'un modèle, effectuons quelques prédictions. Quelles sont les valeurs prédites pour $x_1 = 20$ et $x_2 = 40$?

Q4. Affichez les coefficients du polynôme, puis calculez les scores R2, MSE, RMSE et MAE afin d'évaluer la qualité du modèle.

Q5. Affichez les données initiales ainsi que le polynôme calculé.

Q6. Malheureusement, vous n'avez que quelques points pour établir votre modèle.

Q6.1. Récupérez le fichier `data_mm05_additional.csv`. Incorporez ces nouvelles données et évaluez la précision de votre modèle avec ces nouvelles données (R2, MSE, RMSE, MAE). Affichez finalement tous les points et le polynôme de régression.

Q6.2. Trouvez un nouveau modèle si nécessaire qui obtiendrait de meilleurs résultats.

Exercice 3 : Prédiction de loyers

Vous êtes consulté par une agence immobilière pour prédire les loyers des différents arrondissements de Paris, afin de les aider à prendre des décisions d'achat d'appartements.

Q1. Récupérez les données du fichier `data_mm05_house.csv`. Affichez les 5 premières lignes du fichier pour comprendre sa structure ainsi que le nombre de lignes qu'il contient.

Q2. Affichez la description statistique des données, et représentez les données sous forme d'un nuage de points pour déterminer la présence éventuelle d'*outliers*.

Q3. Nettoyez les données et ignorez les *outliers*. Ré-affichez le nuage de points correspondant. Faites-en sorte de pouvoir distinguer les arrondissements.

Q4. Représentez la répartition des prix par arrondissement. Est-elle plus ou moins semblable dans les différents arrondissements ? Où les prix sont-ils le moins élevés ?

Q5. Vaut-il mieux faire un unique modèle de régression commun quelque soit l'arrondissement, ou un modèle différent par arrondissement ?

Q5.1. Afin d'évaluer correctement les modèles, divisez votre jeu de données. Conservez 30% des données pour les tests.

Q5.2. Créez le modèle de régression unique pour l'ensemble des arrondissements. Tracez le et évaluer ses scores.

Q5.3. Créez les modèles propres à chaque arrondissement. Tracez-les et évaluer leurs scores.