Machine Learning
TP03 – Régression Linéaire
TP noté (Contrôle Continu)

Vous devez soumettre vos solutions de ce travail sur Moodle, en respectant la date limite fixée par votre enseignant(e).
Vous devez travailler en binôme, mais une seule soumission par groupe est requise.
La triche ne sera pas tolérée et sera sanctionnée.

Régression Linéaire

La régression linéaire est un algorithme d'apprentissage supervisé utilisé pour modéliser la relation entre deux variables, généralement appelées la variable dépendante et la variable indépendante. Elle vise à trouver l'équation linéaire la mieux adaptée qui décrit la relation entre les variables. L'objectif de la régression linéaire est de minimiser les résidus, c'est-à-dire les différences entre les valeurs observées et les valeurs prédites par l'équation linéaire.

Ce modèle est utilisé pour prédire une variable dépendante à partir d'un ensemble de variables indépendantes lorsque la sortie prédite est continue.

Dans la régression linéaire, la variable dépendante est prédite en fonction d'une ou plusieurs variables indépendantes, qui sont supposées avoir une relation linéaire avec la variable dépendante. L'équation linéaire est représentée par $Y = mX + b$, où Y est la variable dépendante, X est la variable indépendante, m est la pente et b est l'interception. La pente (m) représente le changement de la variable dépendante pour une variation d'une unité de la variable indépendante, tandis que l'interception (b) représente la valeur prédite de la variable dépendante lorsque la variable indépendante est égale à zéro.

La régression linéaire est couramment utilisée dans divers domaines tels que les statistiques, l'économie, la finance, les sciences sociales et l'apprentissage automatique pour des tâches telles que la prédiction des prix des actions, des prix de l'immobilier, des ventes et du comportement des clients. C'est une technique simple mais puissante qui fournit des informations sur la relation entre les variables et peut être étendue à la régression linéaire multiple avec plusieurs variables indépendantes.

Lecture préliminaire:

Rappelez-vous : l'apprentissage se fait avec une pédagogie axée **sur le travail pratique et l'autonomie.**
Avant de commencer à coder et à implémenter votre modèle d'apprentissage automatique, il est nécessaire de commencer par les lectures recommandées :

1. Leçon – Apprentissage Supervisé & Évaluation de Modèle
2. Documentation Scikit-Learn sur la Régression Linéaire : https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

3. Tout ce que vous devez savoir sur la régression linéaire, Kavita Mali, Analytics Vidhya
https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

sklearn: Bibliothèque Scikit-learn

Pour cette activité de laboratoire et pour les suivantes, nous allons utiliser sklearn, qui est la bibliothèque scikit-learn. C'est une bibliothèque open source d'apprentissage automatique qui prend en charge l'apprentissage supervisé et non supervisé. Elle fournit également divers outils pour l'ajustement de modèles, le prétraitement des données, la sélection et l'évaluation de modèles, ainsi que de nombreuses autres utilités.

Scikit-learn suppose une connaissance de base des pratiques en apprentissage automatique (ajustement de modèles, prédiction, validation croisée, etc.). Elle fournit des dizaines d'algorithmes et de modèles intégrés, appelés estimateurs. Chaque estimateur peut être ajusté à des données à l'aide de sa méthode fit.

La méthode fit accepte généralement deux entrées :

- La matrice des échantillons X. La taille de X est généralement (n_samples, n_features), ce qui signifie que les échantillons sont représentés en lignes et les caractéristiques en colonnes.
- Les valeurs cibles y qui sont des nombres réels pour les tâches de régression, ou des entiers pour la classification (ou tout autre ensemble de valeurs discrètes).

X et y doivent généralement être des tableaux numpy ou des types de données similaires, bien que certains estimateurs puissent fonctionner avec d'autres formats comme les matrices creuses.

Une fois que l'estimateur est ajusté, il peut être utilisé pour prédire les valeurs cibles de nouvelles données avec la méthode predict(test_X_data). D'autres méthodes sont disponibles à des fins d'évaluation et bien d'autres.

Explorez https://scikit-learn.org/stable/ pour trouver la documentation et des exemples d'implémentation.

Cas d'Usage Ecommerce – Se concentrer sur le Site Web ou l'Application Mobile ?

Dans ce TP, le problème que vous allez résoudre concerne une entreprise de e-commerce. L'entreprise possède un magasin où les clients bénéficient de séances avec un styliste personnel, puis rentrent chez eux et peuvent commander via une application mobile ou le site web pour les vêtements qu'ils souhaitent. L'entreprise essaie de décider si elle doit concentrer ses efforts sur l'expérience mobile ou sur le site web.

Vous allez utiliser le modèle **LinearRegression** de la bibliothèque **Scikit-learn**. Ce modèle est un modèle de régression linéaire par moindres carrés ordinaires. Il ajuste un modèle

linéaire avec des coefficients w = (w1, …, wp) afin de minimiser la somme des carrés des résidus entre les cibles observées dans le jeu de données et les cibles prédites par l'approximation linéaire.

Vous travaillerez avec le fichier **Ecommerce_Customers.csv** de l'entreprise. Il comprend des informations sur les clients, telles que l'Email, l'Adresse et la couleur de leur Avatar, ainsi que d'autres caractéristiques numériques comme :

- **Avg. Session Length** : Durée moyenne des sessions de conseils de style en magasin.
- **Time on App** : Temps moyen passé sur l'application en minutes.
- **Time on Website** : Temps moyen passé sur le site web en minutes.
- **Length of Membership** : Nombre d'années depuis l'adhésion du client.

Avez-vous besoin de tous ces détails pour réaliser votre étude ? Réfléchissez aux colonnes que vous souhaitez inclure comme caractéristiques. L'étiquette de sortie est le **"yearly amount spent"**, que vous allez prédire.

Ouvrez le notebook **LinReg_Ecommerce.ipynb** et complétez les parties manquantes du code. Suivez les instructions du notebook pour résoudre ce problème :

1. **Commencez par charger le jeu de données. (5 points)**
2. **Faites une analyse exploratoire. (10points)**
3. **Créez et ajustez le modèle de régression linéaire sur le jeu d'entraînement. (30 points)**
4. **Faites des prédictions. (10 points)**
5. **Évaluez la performance du modèle. (20 points)**
   Notez que si votre code fonctionne correctement, vous devriez obtenir des valeurs proches de :
   MAE : 7.2
   MSE : 79.8
   RMSE : 8.9
   R^2 : 0.989
6. **Terminez par l'interprétation des résultats basée sur les coefficients de corrélation. (15 points)**
7. **À la fin, essayez de répondre à la question : se concentrer sur l'application mobile ou sur le site web ? (10 points)**

Notez que comme c'est votre première mise en œuvre de l'apprentissage automatique, dans cette activité de laboratoire, une partie du code vous est fournie pour vous aider à débuter.

NB : Veillez à ajouter à chaque étape des commentaires internes expliquant le code et interprétant les résultats. Cela est obligatoire et sera pris en compte dans l'évaluation. L'absence de ces commentaires à une étape entraînera une perte de points correspondante.

GOOD LUCK !