TP Machine Learning : Régression Linéaire et Polynomiale

Enseignant : Mouheb Mehdoui

Partie 1 : Instructions

1. **Préparation de l'environnement**
   Ouvrir Google Colab et créer un nouveau notebook.

2. Dans le menu *Fichier*, importer le fichier `medical-insurance-cost-with-linear-regression.ipynb` et le fichier `insurance.csv`.

3. Dans chaque cellule, vous trouverez des explications pour chaque tâche que vous devez terminer. Chaque ligne en pointillés est une ligne manquante que vous devez compléter par les instructions dans le commentaire qui la précède.

Partie 2 : Bibliothèques utilisées

Dans ce TP, nous utiliserons plusieurs bibliothèques Python essentielles pour l'analyse des données et la visualisation :

- **Pandas** : `import pandas as pd`
  Pandas est une bibliothèque puissante pour la manipulation et l'analyse de données. Elle fournit des structures de données flexibles, comme les DataFrames, qui facilitent le traitement des données.

- **NumPy** : `import numpy as np`
  NumPy est une bibliothèque fondamentale pour le calcul scientifique en Python. Elle fournit un support pour les tableaux multidimensionnels et des fonctions mathématiques de haut niveau.

- **Matplotlib** : `import matplotlib.pyplot as plt`
  Matplotlib est une bibliothèque de visualisation de données qui permet de créer des graphiques statiques, animés et interactifs en Python. Elle est particulièrement utile pour tracer des graphiques de données.

- **Seaborn** : `import seaborn as sns`
  Seaborn est une bibliothèque de visualisation basée sur Matplotlib qui simplifie la création de graphiques attrayants et informatifs. Elle offre des fonctionnalités avancées pour l'analyse exploratoire des données.

- **scikit-learn** : `import sklearn`
  scikit-learn est une bibliothèque de machine learning qui fournit des outils simples et efficaces pour l'analyse prédictive. Elle inclut des algorithmes pour la régression, la classification, le clustering, et bien plus encore.

Partie 3 : Rappel sur la régression linéaire et la régression polynomiale

La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle vise à établir une équation de la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

où : - y est la variable dépendante (cible). - β_0 est l'ordonnée à l'origine (intercept). - $\beta_1, \beta_2, \ldots, \beta_n$ sont les coefficients de régression (pentes) pour chaque variable indépendante $x_1, x_2, \ldots, x_n$. - ϵ est l'erreur aléatoire ou résidu.

Principes de la Régression Linéaire: La régression linéaire repose sur plusieurs hypothèses :

- La relation entre les variables est linéaire.
- Les résidus (erreurs) sont normalement distribués.
- Les résidus ont une variance constante (homoscédasticité).
- Les observations sont indépendantes.

Étapes de la Régression Linéaire: Voici les étapes typiques pour réaliser une régression linéaire :

1. **Collecte des données** : Rassembler les données pertinentes pour la variable dépendante et les variables indépendantes.
2. **Prétraitement des données** : Nettoyer et préparer les données, gérer les valeurs manquantes et normaliser les variables si nécessaire.
3. **Division des données** : Séparer les données en ensembles d'entraînement et de test pour évaluer le modèle.
4. **Ajustement du modèle** : Utiliser l'ensemble d'entraînement pour estimer les coefficients β en minimisant la somme des carrés des résidus, ce qui donne la formule :

$$Minimize \quad \sum (y_i - \hat{y}_i)^2$$

   où $\hat{y}_i$ est la valeur prédite pour l'observation i.
5. **Évaluation du modèle** : Évaluer la performance du modèle en utilisant des métriques comme le coefficient de détermination R^2, l'erreur quadratique moyenne (RMSE), etc.
6. **Interprétation des résultats** : Interpréter les coefficients pour comprendre l'impact des variables indépendantes sur la variable dépendante.

Formules Importantes: En plus de l'équation de la régression, voici quelques formules clés :

- **Erreur quadratique moyenne (RMSE)** :

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}$$

- **Coefficient de détermination (R^2)** :

  Le coefficient de détermination, noté R^2, mesure la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes. Il varie entre 0 et 1, où :

  - $R^2 = 1$ indique que le modèle explique 100% de la variance des données.
  - $R^2 = 0$ signifie que le modèle n'explique aucune des variations des données.

  La formule pour R^2 est :

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

  où :

  - SS_{res} (somme des carrés des résidus) est la somme des carrés des différences entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$:

$$SS_{res} = \sum (y_i - \hat{y}_i)^2$$

  - SS_{tot} (somme totale des carrés) mesure la somme des carrés des différences entre les valeurs observées y_i et la moyenne des valeurs observées $\bar{y}$:

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

Détermination des Coefficients: Les coefficients de la régression linéaire ($\beta_0, \beta_1, \ldots, \beta_n$) représentent l'impact de chaque variable indépendante sur la variable dépendante. Il existe plusieurs méthodes pour estimer ces coefficients :

1. Méthode des Moindres Carrés

La méthode des moindres carrés est la méthode la plus courante pour estimer les coefficients. Elle vise à minimiser la somme des carrés des résidus (SS_{res}). Cette méthode est basée sur l'idée que la meilleure ligne de régression est celle qui a la plus petite distance (erreur) par rapport à tous les points de données. Les coefficients sont calculés en résolvant le système d'équations suivant :

$$\beta = (X^T X)^{-1} X^T y$$

où : - X est la matrice des variables indépendantes (avec une colonne de 1 pour l'ordonnée à l'origine). - y est le vecteur des valeurs observées de la variable dépendante. - β est le vecteur des coefficients de régression.

2. Méthode de Gradient Descent

La méthode de gradient descent (descente de gradient) est une technique itérative pour optimiser les coefficients. Elle est particulièrement utile lorsque le nombre de variables est très élevé ou lorsque les données sont trop volumineuses pour être traitées par la méthode des moindres carrés. La procédure consiste à ajuster les coefficients en suivant le gradient de la fonction de coût (somme des carrés des résidus) :

$$\beta = \beta - \alpha \nabla J(\beta)$$

où : - α est le taux d'apprentissage (learning rate). - $\nabla J(\beta)$ est le gradient de la fonction de coût.

3. Méthode des Moindres Carrés Pondérés

Cette méthode est une extension de la méthode des moindres carrés, où chaque observation a un poids différent. Cela peut être utile lorsque certaines données sont plus fiables que d'autres. Les coefficients sont estimés en minimisant la somme des carrés des résidus pondérés :

$$\beta = (X^T W X)^{-1} X^T y$$

où W est une matrice diagonale contenant les poids des observations.