TP 3 - Régression Logistique

Exercice 1 : Régression Logistique Binaire - Prédiction du Diabète

Nous allons prédire si un patient est diabétique ou non à partir de données médicales.

1. Comprendre et explorer les données

Question 1 : Pourquoi est-il crucial d'explorer les données avant toute modélisation ? Importer les données depuis `diabetes.csv` et afficher les premières lignes.

Question 2 : Analyser les informations générales (`.info()`) et statistiques (`.describe()`). Quels indices vous permettent de détecter d'éventuels problèmes dans les données ?

Question 3 : Y a-t-il des valeurs manquantes ou aberrantes ? Proposer une méthode de traitement adaptée.

2. Préparer les données pour l'apprentissage

Question 4 : Expliquer pourquoi la séparation entre variables explicatives (X) et cible (y) est importante. Effectuer cette séparation.

Question 5 : Pourquoi est-il souvent nécessaire de normaliser les données en régression logistique ? Appliquer une normalisation via `StandardScaler`.

3. Créer un modèle de régression logistique

Question 6 : Pourquoi devons-nous séparer nos données en ensemble d'entraînement et de test ? Effectuer la séparation (80%-20%).

Question 7 : Construire un modèle de régression logistique et expliquer le rôle de chaque l'hyperparamètre `utilisé`.

Question 8 : Interpréter les coefficients du modèle : quelles variables semblent avoir un impact positif ou négatif ?

4. Évaluer les performances du modèle

Question 9 : Pourquoi la simple précision (`accuracy`) n'est-elle pas suffisante pour juger un modèle de classification ? Calculer et interpréter :

- Matrice de confusion
- Précision, Rappel, F1-score
- Courbe ROC et Aire sous la courbe (AUC)

5. Interprétation

Question 10 : Comment interpréter la courbe ROC d'un modèle parfait, moyen et mauvais ? Décrire brièvement.

Exercice 2 : Régression Logistique Multiclasse - Classification des Fleurs Iris

Nous allons prédire l'espèce d'une fleur parmi trois catégories.

1. Explorer un problème multiclasse

Question 1 : Quelle est la différence entre un problème binaire et multiclasse ? Charger le dataset `Iris` depuis `sklearn.datasets` et afficher ses caractéristiques.

Question 2 : Visualiser la répartition des classes et les corrélations entre variables par un nuage de points.

2. Préparer les données

Question 3 : Séparer les données en variables explicatives (X) et cible (y). Comment choisir un bon pourcentage pour l'ensemble de test ?

Question 4 : Diviser les données en ensemble d'entraînement et de test (70%-30%).

3. Modéliser une régression logistique multinomiale

Question 5 : Quelle est la différence entre les stratégies `one-vs-rest` et `multinomial` en régression logistique ? Configurer un modèle en `multi_class='multinomial'` et `solver='lbfgs'`.

Question 6 : Entraîner le modèle et afficher les poids associés aux différentes classes.

4. Évaluer le modèle multiclasse

Question 7 : Pourquoi la matrice de confusion est-elle encore plus utile en multiclasse ? Afficher et interpréter la matrice de confusion.

Question 8 : Calculer et analyser le rapport de classification : quelles classes sont mieux prédites ?

5. Visualiser la séparation entre classes

Question 9 : Tracer les frontières de décision pour deux variables au choix. Comment interpréter ces frontières par rapport aux performances du modèle ?

6. Analyse critique

Question 10 : Comparer les défis entre régression logistique binaire et multiclasse. Quels éléments techniques doivent être ajustés pour mieux gérer des données multiclasse ?