TP 2 : Régression Logistique

1 Introduction

La régression logistique est une technique statistique très utilisée dans le domaine biomédical pour la classification. Dans ce travail pratique, nous appliquons cette méthode à un jeu de données médicales afin de prédire la présence ou l'absence d'un cancer du sein. Les variables explicatives incluent plusieurs descripteurs statistiques dérivés des images radiologiques des tumeurs.

2 Méthodologie

2.1 Chargement des Données

- Importer les bibliothèques nécessaires : `numpy`, `pandas`, `matplotlib`, `seaborn`, `scikit-learn`.
- Charger le jeu de données biomédical (`load_breast_cancer(return_X_y=True)`).

2.2 Prétraitement des Données

- Séparer les données en ensemble d'entraînement et de test (80%-20%) `test_train_split`.
- Normaliser les variables explicatives pour améliorer la convergence du modèle.

3 Entraînement et Évaluation du Modèle

3.1 Entraînement du Modèle

- Utiliser la classe `LogisticRegression` de `scikit-learn`.
- Ajuster le modèle sur les données d'entraînement.

3.2 Évaluation du Modèle

- Prédire les classes sur l'ensemble de test.
- Calculer et afficher la matrice de confusion.
- Évaluer les performances à l'aide du score de précision, rappel et F1-score.

4 Validation Croisée

La validation croisée est une technique utilisée pour évaluer la robustesse d'un modèle en répartissant les données en plusieurs sous-ensembles et en entraînant le modèle sur chaque sous-ensemble à tour de rôle.

4.1 Mise en Œuvre

- Utiliser la classe `KFold` de `scikit-learn` pour diviser les données en k sous-ensembles.
- Entraîner le modèle sur chaque sous-ensemble et évaluer ses performances sur le sous-ensemble restant.
- Calculer la moyenne des performances sur les k itérations pour obtenir une estimation plus robuste des capacités du modèle.
- Pourquoi la validation croisée est importante ?

5 Régularisation de Lasso

Les chemins de Lasso permettent de visualiser l'impact de la régularisation L1 sur les coefficients du modèle à mesure que le paramètre de régularisation C varie.

5.1 Mise en Œuvre

- Utiliser la classe `Lasso` de `scikit-learn` pour entraîner des modèles avec différentes valeurs de C.
- Tracer les coefficients des variables explicatives en fonction de C pour observer leur évolution.
- Identifier les variables les plus importantes pour le modèle en fonction de leur persistance dans le chemin de Lasso.

6 Résultats

Présenter les résultats de l'entraînement et de l'évaluation du modèle, y compris les graphiques et les tableaux pertinents.

7 Discussion

Interpréter les résultats, discuter des limites du modèle et proposer des pistes pour l'amélioration future.

8 Conclusion

Résumer les principales conclusions et souligner les implications pratiques de l'étude.