Regression Logistique

Wiem Ktari et Maram Jlassi

Novembre 2024

1 Introduction

1. Contexte et Problématique

La régression logistique est un modèle statistique largement utilisé dans le domaine de l'apprentissage automatique et de la statistique pour résoudre des problèmes de classification. Dans un monde de plus en plus centré sur les données, l'analyse prédictive est devenue essentielle pour anticiper des comportements, prendre des décisions et automatiser des processus. La régression logistique est particulièrement pertinente dans ce contexte car elle permet de prévoir des résultats binaires ou multi-classes, comme la prédiction de l'acceptation d'un prêt, le diagnostic médical... Ce modèle est fréquemment employé pour son interprétabilité et son efficacité dans les cas de classification.

2. Importance de la Régression Logistique

Bien que de nombreux algorithmes avancés existent (tels que les réseaux de neurones, les arbres de décision et les machines à vecteurs de support), la régression logistique reste une méthode de base en machine learning pour plusieurs raisons. Elle est simple à implémenter, rapide à entraîner, et ses coefficients sont facilement interprétables. De plus, la régression logistique est moins sensible aux sur-ajustements (overfitting) dans les petits ensembles de données par rapport à des méthodes plus complexes, ce qui en fait un choix privilégié dans des contextes où la simplicité et l'explicabilité sont essentielles.

3. Objectifs du Rapport

L'objectif de ce rapport est de fournir une analyse approfondie de la régression logistique, en expliquant son fondement théorique, ses techniques d'optimisation, et les méthodes d'évaluation de la performance de ce modèle. Le rapport explorera également ses applications pratiques dans des contextes variés ainsi que ses limitations et extensions pour répondre à des problématiques plus complexes.

4. Structure du Rapport

Le rapport s'organisera en plusieurs parties :

- **Compréhension de la Régression Logistique** : Défintion de la regression logistique et exploitation de ses types.
- **Théorie de la régression logistique** : Présentation des concepts mathématiques et statistiques sous-jacents.
- **Méthodes d'optimisation** : Explication des méthodes utilisées pour ajuster le modèle aux données.
- **Evaluation de la performance du modèle** : Explication des métriques de performance.
- **Application théorique** : Exercice d'application : Prédiction de la probabilité d'admission dans une université
- **Applications pratiques** : Mise en œuvre d'un modèle de régression logistique, interprétation des coefficients et exemples d'applications réelles.

2 Compréhension de la Régression Logistique

2.1 Définition

La régression logistique est une méthode de classification qui permet de modéliser la probabilité d'appartenance à une classe en fonction d'un ensemble de variables explicatives. Contrairement à la régression linéaire, qui produit une valeur continue, la régression logistique utilise une fonction logistique (ou sigmoïde) pour contraindre les valeurs de sortie entre 0 et 1, ce qui permet d'interpréter le résultat comme une probabilité. Cette approche est largement utilisée pour les problèmes de classification binaire (ex. : réussite/échec, oui/non) et peut être étendue à des situations multi-classes.

2.2 Types de Régression Logistique

- **Régression Logistique Binaire** : C'est le type le plus courant de régression logistique, où la variable dépendante est binaire. Par exemple, elle est utilisée pour prédire si un e-mail est un spam ou non, ou si un patient a une certaine maladie.
- **Régression Logistique Multinomiale** : Cette variante permet de traiter des problèmes de classification multi-classes, où la variable cible peut prendre plusieurs valeurs discrètes. Par exemple, elle peut être utilisée pour prédire la catégorie d'un article (politique, sport, technologie).

- **Régression Logistique Ordinale** : Ce type est utilisé lorsque les classes sont ordonnées, mais non continues. Par exemple, on peut s'en servir pour modéliser des niveaux de satisfaction (faible, moyen, élevé) dans un sondage.

2.3 Exemples d'Applications

La régression logistique est utilisée dans de nombreux domaines pour résoudre des problèmes de classification binaire. Voici quelques exemples concrets d'applications de la régression logistique :

- **Prédiction de défaut de crédit** : Les banques utilisent la régression logistique pour prédire la probabilité qu'un client fasse défaut sur un prêt en fonction de diverses caractéristiques telles que les revenus, l'historique de crédit, l'âge, etc.
- **Diagnostic médical** : En médecine, la régression logistique est couramment utilisée pour prédire la probabilité qu'un patient souffre d'une maladie, par exemple en fonction de symptômes ou de tests médicaux.
- **Prévision de churn client** : Les entreprises utilisent la régression logistique pour prédire la probabilité qu'un client quitte leur service en fonction de son comportement, de son utilisation des produits, etc.

Ces applications montrent la flexibilité de la régression logistique dans des contextes variés où les résultats sont de type binaire, comme la prédiction d'un événement ou le diagnostic de maladies.

3 Théorie de la Régression Logistique

3.1 Principe Mathématique

La régression logistique repose sur la fonction logistique (ou sigmoïde) pour transformer une combinaison linéaire de variables explicatives en une probabilité.

Partons de l'équation de régression linéaire classique qui modélise une relation entre une variable dépendante y et un ensemble de variables explicatives $x_1, x_2, \ldots, x_n$:

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

Dans un problème de classification binaire, la variable y représente une probabilité et doit être contrainte entre 0 et 1, ce qui n'est pas le cas de la régression linéaire. Pour résoudre ce problème, on applique une transformation logistique.

L'idée de la régression logistique est de modéliser le rapport des cotes (odds) et non la probabilité directe. Le rapport des cotes est défini comme le ratio entre

la probabilité d'appartenance à la classe positive ($P(y = 1)$) et la probabilité d'appartenance à la classe négative ($P(y = 0)$) :

$$odds = \frac{P(y = 1)}{P(y = 0)} = \frac{P(y = 1)}{1 - P(y = 1)}$$

Afin de linéariser cette relation, on applique le logarithme du rapport des cotes :

$$\log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Cette transformation permet de rendre la relation entre les variables explicatives et le log du rapport des cotes linéaire.

Une fois la transformation logarithmique appliquée, on peut obtenir la probabilité $P(y = 1)$ en inversant le logarithme. Pour cela, on applique l'exponentielle à la relation précédente :

$$\frac{P(y = 1)}{1 - P(y = 1)} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}$$

Ensuite, on résout pour $P(y = 1)$:

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

Cette expression est la fonction sigmoïde, qui contraint la sortie entre 0 et 1, permettant ainsi d'interpréter cette sortie comme une probabilité. La fonction sigmoïde est donc la clé pour transformer une combinaison linéaire en une probabilité. L'équation générale de la régression logistique devient donc :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

où $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$. Cette fonction sigmoïde assure que la probabilité $P(y = 1)$ est toujours entre 0 et 1, ce qui permet de classer les observations en fonction d'une probabilité d'appartenance à la classe positive. En régression logistique binaire, si $\sigma(z) \geq 0.5$, l'observation est classée dans la classe positive ($y = 1$), sinon dans la classe négative ($y = 0$).

3.2 Hypothèses Sous-jacentes

Pour que la régression logistique soit un modèle efficace, plusieurs hypothèses sont généralement posées :

- **Relation log-linéaire** : Le logarithme des rapports de cotes (odds) est une combinaison linéaire des variables explicatives.
- **Indépendance des observations** : Les observations dans l'ensemble de données doivent être indépendantes les unes des autres.

- **Absence de multicolinéarité** : Les variables explicatives ne doivent pas être fortement corrélées entre elles, car cela peut nuire à la stabilité des coefficients.
- **Grande taille de l'échantillon** : Avec un échantillon de grande taille, les estimations des coefficients sont plus stables et l'approximation de la distribution des erreurs est meilleure.

3.3 Fonction de Coût et Optimisation

La fonction de coût dans la régression logistique est basée sur la vraisemblance, et non sur l'erreur quadratique comme en régression linéaire. La fonction de log-vraisemblance est définie par :

$$L(\beta) = \sum_{i=1}^{m} [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))]$$

L'objectif est de maximiser cette fonction de vraisemblance pour obtenir les meilleurs coefficients β. Cette optimisation est souvent réalisée via la méthode de la descente de gradient, qui met à jour les coefficients dans le sens de la pente ascendante de la fonction de vraisemblance, jusqu'à ce que la convergence soit atteinte.

4 Méthode d'Optimisation

4.1 Optimisation par Descente de Gradient

L'optimisation par descente de gradient est une méthode courante pour minimiser une fonction de coût. Elle consiste à ajuster les paramètres β dans la direction opposée au gradient de la fonction de coût pour trouver les valeurs optimales.

En régression logistique, la fonction de coût est la log-vraisemblance, et le gradient de cette fonction par rapport aux paramètres β est calculé. La mise à jour des paramètres se fait à chaque itération en suivant la règle suivante :

$$\beta_j \leftarrow \beta_j - \alpha \frac{\partial L(\beta)}{\partial \beta_j}$$

où α est le taux d'apprentissage, et $\frac{\partial L(\beta)}{\partial \beta_j}$ est le gradient de la fonction de log-vraisemblance par rapport à β_j. Cette méthode est itérative et continue jusqu'à ce que la fonction de coût converge vers un minimum.

4.2 Algorithmes Courants

4.2.1 Newton-Raphson

L'algorithme de Newton-Raphson est une méthode d'optimisation qui utilise l'information de la dérivée seconde (la matrice Hessienne) pour trouver plus rapi-

dement le minimum de la fonction de coût. En régression logistique, l'algorithme de Newton-Raphson peut être utilisé pour ajuster les paramètres β en utilisant la mise à jour suivante :

$$\beta = \beta - H^{-1}\nabla L(\beta)$$

où H est la matrice Hessienne (la matrice des dérivées secondes) et $\nabla L(\beta)$ est le gradient de la fonction de log-vraisemblance. L'algorithme de Newton-Raphson converge plus rapidement que la descente de gradient pour des problèmes bien conditionnés, mais il peut être coûteux en termes de calcul, surtout pour des ensembles de données volumineux.

4.2.2 Gradient Stochastique

Le gradient stochastique (SGD) est une variante de la descente de gradient qui met à jour les paramètres en utilisant un seul échantillon de données à la fois plutôt que l'ensemble complet. Cette méthode est particulièrement utile lorsque les données sont très volumineuses. La mise à jour des paramètres est effectuée après chaque échantillon de données, selon la règle suivante :

$$\beta_j \leftarrow \beta_j - \alpha \frac{\partial L(\beta; x_i, y_i)}{\partial \beta_j}$$

où (x_i, y_i) est un échantillon de données. Bien que la convergence soit plus bruyante que dans la descente de gradient classique, le gradient stochastique permet de traiter efficacement des ensembles de données massifs et de réduire le coût computationnel par itération.

5 Évaluation de la Performance du Modèle

5.1 Métriques de Performance

L'évaluation de la performance d'un modèle de régression logistique repose sur diverses métriques qui permettent de mesurer sa capacité à faire des prédictions correctes. Parmi les métriques les plus courantes, on trouve :

- **Précision (Accuracy)** : La précision mesure la proportion de prédictions correctes parmi toutes les prédictions effectuées. Elle est définie par :

  $$\textit{Précision} = \frac{TP + TN}{TP + TN + FP + FN}$$

  où TP est le nombre de vrais positifs, TN est le nombre de vrais négatifs, FP est le nombre de faux positifs et FN est le nombre de faux négatifs.

- **Rappel (Recall)** : Le rappel, ou sensibilité, mesure la capacité du modèle à identifier toutes les instances positives. Il est défini par :

  $$\textit{Rappel} = \frac{TP}{TP + FN}$$

- **F1-score** : Le F1-score est la moyenne harmonique entre la précision et le rappel, fournissant une mesure équilibrée de la performance du modèle. Il est défini par :
$$F1 = 2 \times \frac{Précision \times Rappel}{Précision + Rappel}$$

- **Courbe ROC-AUC** : La courbe ROC (Receiver Operating Characteristic) trace le taux de vrais positifs (TPR) contre le taux de faux positifs (FPR) pour différentes valeurs de seuil. L'aire sous la courbe (AUC) quantifie la capacité du modèle à distinguer les classes. Une AUC proche de 1 indique un bon modèle, tandis qu'une AUC proche de 0,5 indique un modèle qui fait des prédictions aléatoires.

5.2 Validation Croisée

La validation croisée est une méthode essentielle pour évaluer la robustesse d'un modèle en le testant sur plusieurs sous-ensembles de données. Elle permet de s'assurer que le modèle ne surajuste pas les données d'entraînement et qu'il généralise bien sur de nouvelles données.

La validation croisée consiste à diviser les données en k sous-ensembles (ou plis), en entraînant le modèle sur $k-1$ plis et en testant le modèle sur le pli restant. Ce processus est répété k fois, chaque pli servant une fois de jeu de test. La moyenne des performances sur ces k tests est ensuite utilisée pour évaluer la performance du modèle.

6 Application Théorique de la Régression Logistique

Exercice d'application : Prédiction de la probabilité d'admission dans une université

Énoncé

Une université souhaite étudier la probabilité d'admission d'étudiants en fonction de leurs scores au test d'entrée. Vous disposez d'un ensemble de données contenant des informations sur plusieurs étudiants, avec deux variables :

- **Score au test d'entrée** (entre 0 et 100)
- **Admission** (variable binaire : 1 si l'étudiant a été admis, 0 sinon)

Les données sont les suivantes :

Score au test	Admission
45	0
50	0
65	1
70	1
85	1
95	1
35	0
80	1
55	0
60	1

Questions

1. Modéliser la probabilité d'admission en fonction du score au test en utilisant la régression logistique. Estimer les coefficients du modèle β_0 et β_1 en utilisant les données fournies.
2. Interpréter les coefficients β_0 et β_1 dans le contexte de l'admission.
3. Calculer la probabilité d'admission pour un étudiant ayant un score de 75.
4. En utilisant un seuil de 0,5 pour la probabilité d'admission, évaluer l'exactitude des prédictions du modèle sur l'ensemble de données fourni.

Correction

1. Modélisation

La régression logistique a pour but de prédire la probabilité p d'admission en fonction du score au test x. Le modèle de régression logistique s'écrit :

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

où β_0 est l'ordonnée à l'origine (intercept) et β_1 est le coefficient associé au score x.

En utilisant un logiciel de calcul pour ajuster le modèle de régression logistique, supposons que les valeurs estimées soient :

amsmath

$$\beta_0 = -8 \quad et \quad \beta_1 = 0.12$$

2. Interprétation des coefficients

- **Intercept (β_0) :** Cette valeur indique le log-odds de l'admission quand le score au test est nul. Un intercept négatif signifie qu'avec un score très faible, la probabilité d'admission est très faible.
- **Coefficient associé au score (β_1) :** Ce coefficient positif montre que, plus le score au test augmente, plus la probabilité d'admission augmente.

3. Prédiction

Pour un étudiant ayant un score de 75, nous calculons la probabilité d'admission en utilisant le modèle :

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times 75)}}$$

En remplaçant les valeurs :

$$p = \frac{1}{1 + e^{-(-8 + 0.12 \times 75)}}$$

Calcul :

$$p = \frac{1}{1 + e^{-1}}$$

$$p \approx \frac{1}{1 + 0.3679} \approx 0.73$$

Ainsi, pour un étudiant avec un score de 75, la probabilité d'admission est d'environ 0,73 (ou 73 %).

4. Évaluation du modèle

Pour évaluer la précision du modèle, on utilise un seuil de 0,5 pour classifier les prédictions :

- Si $p \geq 0.5$, on prédit que l'étudiant est admis (1).
- Si $p < 0.5$, on prédit que l'étudiant n'est pas admis (0).

En utilisant le modèle pour prédire l'admission des étudiants dans l'ensemble de données et en comparant avec les valeurs réelles, nous obtenons les résultats suivants :

L'exactitude (accuracy) est donnée par la proportion de bonnes prédictions :

$$Exactitude = \frac{Nombredebonnesprédictions}{Nombretotaldeprédictions} = \frac{10}{10} = 1.0$$

Score	Admission réelle	Prédiction modèle
45	0	0
50	0	0
65	1	1
70	1	1
85	1	1
95	1	1
35	0	0
80	1	1
55	0	0
60	1	1

Conclusion

Le modèle a une exactitude de 100 % sur cet ensemble de données, mais il serait nécessaire de tester le modèle sur un autre ensemble de données pour vérifier sa performance réelle.

7 Application Pratique de la Régression Logistique

7.1 Préparation des Données

La préparation des données est une étape cruciale avant de créer un modèle de régression logistique. Elle inclut plusieurs sous-étapes comme le nettoyage des données, l'encodage des variables catégorielles et la normalisation des données.

- **Nettoyage des données** : Cette étape consiste à traiter les valeurs manquantes, éliminer ou imputer les valeurs aberrantes et s'assurer que les données sont cohérentes et prêtes à être utilisées pour la modélisation.
- **Encodage des variables catégorielles** : Les variables catégorielles doivent être converties en variables numériques pour être utilisées dans les modèles de régression logistique. Cela peut être effectué par des méthodes telles que l'encodage par *one-hot encoding* ou l'encodage ordinal, selon la nature des données.
- **Normalisation des données** : Il est souvent recommandé de normaliser les variables numériques afin que toutes les caractéristiques aient la même échelle. Cela est particulièrement important lorsque les variables ont des unités différentes.

7.2 Mise en Œuvre du Modèle

Pour mettre en œuvre un modèle de régression logistique, plusieurs outils peuvent être utilisés comme `Scikit-Learn` en Python, une bibliothèque très pop-

ulaire pour le machine learning.

7.3 Interprétation des Coefficients

Les coefficients d'un modèle de régression logistique ont une signification importante en termes d'influence sur la probabilité de la classe positive. Chaque coefficient β_j représente l'effet de la variable x_j sur la log-odds de la probabilité d'appartenance à la classe positive.

- Un coefficient β_j positif signifie que lorsque x_j augmente, la probabilité d'appartenance à la classe positive (y=1) augmente.
- Un coefficient β_j négatif signifie que lorsque x_j augmente, la probabilité d'appartenance à la classe positive diminue.

Les coefficients peuvent être interprétés en termes de *odds ratio*, qui est l'exponentielle du coefficient e^{β_j}. Cela représente le facteur d'augmentation (ou de diminution) des chances d'appartenir à la classe positive pour une unité d'augmentation de x_j, toutes choses égales par ailleurs.

$$Oddsratio = e^{\beta_j}$$

Par exemple, si $\beta_j = 0.5$, l'odds ratio sera $e^{0.5} \approx 1.65$, ce qui signifie qu'une augmentation de x_j de 1 unité augmente les chances d'appartenir à la classe positive d'environ 65%.

Il est aussi possible d'obtenir une probabilité en transformant les log-odds avec la fonction sigmoïde :

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}}$$

Cela donne la probabilité estimée d'appartenir à la classe positive pour une combinaison donnée de variables explicatives $x_1, x_2, \ldots, x_n$.

8 Conclusion

8.1 Synthèse

La régression logistique, une méthode statistique largement utilisée pour résoudre des problèmes de classification binaire. La fonction sigmoïde, au cœur de la régression logistique, permet de transformer une combinaison linéaire de variables explicatives en une probabilité d'appartenance à une classe. Des méthodes d'estimation et d'optimisation, ainsi que des algorithmes sont couramment utilisés, tels que la descente de gradient et la méthode de Newton-Raphson.

8.2 Perspectives

Bien que la régression logistique soit robuste et interprétable, elle présente certaines limitations. Par exemple, elle suppose une relation log-linéaire entre les variables explicatives et le logarithme des rapports de cotes (odds), ce qui n'est pas toujours le cas dans la pratique. Des améliorations potentielles incluent l'utilisation de techniques de régularisation, comme la régression Lasso ou Ridge, pour prévenir le surapprentissage. De plus, pour des ensembles de données plus complexes ou non linéaires, des algorithmes plus avancés tels que les forêts aléatoires, les machines à vecteurs de support (SVM), ou les réseaux de neurones peuvent offrir de meilleures performances.