39754 LAFIN Emérance

Calcul de risque à l'aide la régression logistique binaire

La régression logistique binaire permet, à partir de certains paramètres, de prédire une variable. On peut par exemple s'intéresser à la prévision de problème de santé. Ici, on se concentrera sur la probabilité pour des femmes enceintes d'avoir ou non un enfant prématuré.
J'ai trouvé très intéressant, qu'à l'aide de la régression logistique, on puisse prédire la probabilité d'un événement concret à l'aide de plusieurs variables. Cette modélisation est la plus adaptée dans le cadre de mon étude car je cherche à prédire une probabilité à l'aide d'un ensemble de variables.

Positionnement thématique (ETAPE 1)

INFORMATIQUE (Informatique pratique), MATHEMATIQUES (Mathématiques Appliquées).

Mots-clés (ETAPE 1)

Mots-Clés (en français)	**Mots-Clés** (en anglais)
Régression logistique binaire	*Binary logistic regression*
Calcul de probabilités	*Calculate probability*
Fonction logit	*Logit function*
Algorithme de Newton-Raphson	*Newton-Raphson algorithm*
Méthode d'apprentissage	*Learning method*

Bibliographie commentée

La régression logistique binaire est utilisée pour déterminer la probabilité d'un résultat (dans notre étude : si un enfant est prématuré ou non) en fonction de plusieurs variables (dans le cas étudié : grossesse simple ou multiple, présence de diabète...) [1]. Cette probabilité est calculée à l'aide de données déjà existantes, dite données d'apprentissage. On peut dans ce cas comparer cette méthode à d'autres méthodes d'apprentissage telles que les réseaux de neurones. Au contraire de cette méthode, la régression logistique est bien plus transparente. On comprend pourquoi les décisions sont prises tandis que les réseaux de neurones fonctionnent comme une boîte noire. [2] De plus, la régression logistique permet de nous renseigner sur l'impact des variables sur le résultat final des prédictions. Elle nécessite toutefois des échantillons de grandes tailles pour être fiable. [1]

Quand on se place dans ce type de classement binaire, on préfère la régression logistique à la régression linéaire. En effet, la régression linéaire ne permet de comprendre l'impact que d'une seule variable et non d'un ensemble. En outre les valeurs ne sont pas bornées entre 0 et 1 ce qui ne permet pas d'obtenir une probabilité.[3]

Le modèle de la régression logistique s'appuie sur la fonction logit - qui permet de conserver les valeurs entre 0 et 1, correspondant donc à une probabilité – et sur la méthode du maximum de vraisemblance – permettant de déterminer un vecteur a classant le maximum de données dans la bonne catégorie. Différentes méthodes existent pour optimiser la log-vraisemblance, dont

l'algorithme de Newton-Raphson. Il consiste à initialiser le vecteur a que l'on fait ensuite varier par récurrence jusqu'à un certain critère d'arrêt à déterminer (nombre d'itérations, précision...). [4]

La régression logistique peut être appliquée dans des domaines très différents et permet de prendre la meilleure décision. Les banques peuvent s'en servir pour décider si oui ou non elles peuvent accorder un prêt selon la probabilité que le client a de le rembourser. [5] On peut également déterminer la probabilité qu'un enfant soit ou non prématuré. En cas de risque important, une administration de corticoïdes peut réduire le risque de décès. De plus, la maternité choisie pour l'accouchement pourra être mieux adaptée : celles de type II ou III ont les équipements nécessaires pour les naissances de prématurés. [6]

Problématique retenue

Comment la régression logistique peut aider à déterminer la probabilité d'avoir un enfant prématuré ?

Objectifs du TIPE

- Comprendre la théorie mathématique de la régression logistique
- Implémenter la régression logistique à l'aide de bibliothèques Python et à l'aide de la méthode de Newton Raphson
- Savoir exploiter la régression logistique après le traitement des données
- Application à la prévention d'un enfant prématuré

Références bibliographiques (ETAPE 1)

[1] CareerFoundry : Un guide pour débutant sur la régression logistique : *http://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/*

[2] ActiveWizards : Exemples concrets de régression logistique : *https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application*

[3] towards data science : Construire un modèle pour déterminer si une journée est ou non productive : *https://towardsdatascience.com/logistic-regression-in-real-life-building-a-daily-productivity-classification-model-a0fc2c70584e*

[4] Ricco Rakotomalala : Pratique de la Régression Logistique : *http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf*

[5] IBM : Régression logistique : *https://www.ibm.com/fr-fr/topics/logistic-regression*

[6] Inserm : Prématurité : *https://www.inserm.fr/dossier/prematurite/*

DOT

[1] *23/07/2021 : Développement d'un intérêt dans le calcul du risque pour déterminer le prix d'une assurance mais sans trouver de modèle mathématique expliquant le processus. Au début de l'année mon professeur de maths m'oriente vers la régression logistique.*

[2] *12/10/2021 : Découverte d'un tutoriel vidéo clair expliquant comment coder la régression logistique à l'aide de modules de python.*

[3] *8/11/2021 : Tentative de mettre mon fichier sous forme de liste mais impossible de convertir les nombres directement en flottant à cause d'un problème de normalisation au niveau de la virgule et impossibilité de remplacer cette virgule directement dans le fichier donc création d'un morceau de code faisant cela.*

[4] *15/11/2021 : initialisation au hasard de β par mon programme mais impossibilité de calculer la log-vraisemblance car on se retrouve à calculer ln(1 - π) avec $\pi = 1$ à chaque initialisation différente. Donc le vecteur β n'est plus initialisé au hasard.*

[5] *11/04/2021 : En relançant les 2 modèles de la régression logistique on trouve cette fois-ci des valeurs de coefficients β complétement différents. Cela s'expliquait en fait par les données d'entrainement utilisées qui étaient différentes.*

[6] *mars 2022 : Découverte de vidéos sur des exemples de régression logistiques avec différentes interprétations, mais pas sur Python. Essaie d'analogie avec python mais en vain. Le module statsmodels.api aurait pu aider notamment pour calculer la p-value mais malgré de nombreux essais, impossible de l'installer.*

[7] *16/05/2022 : Trouvé un modèle avec un taux de succès de 83% que j'ai donc cherché à garder en mémoire. Cependant d'après mes fonctions le taux de succès est différent de 83%. Les valeurs dans la matrice de confusion sont également placées différemment.*

[8] *16/05/2022 : Compréhension de l'origine de l'expression de la vraisemblance après de nombreuses recherches infructueuses (documents pas clairs, incomplets, formules différentes).*