meritis

Semaine 4: Régression Logistique

Prédiction des maladies cardiaques

L'Organisation mondiale de la santé a estimé à 12 millions le nombre de décès chaque année dus à des maladies du cœur. Presque la moitié des décès dans les pays développés sont dus à des maladies cardio-vasculaires.

Cette recherche a pour but d'identifier les facteurs de risque les plus pertinents / de risque de maladie cardiaque et de prédire le risque global à l'aide d'une régression logistique.

De nombreux paramètres sont considérés pour cette étude :

- Le sexe de la personne (H/F)
- L'âge
- L'éducation
- Fumeur ou Non
- Nombre de cigarettes fumées par jour
- Médicaments contre l'hypertension
- Accident vasculaire cérébral (**AVC**)
- Hypertrophie cardiaque préalable
- Diabète
- Niveau Cholestérol
- La pression artérielle systolique
- La pression artérielle diastolique
- Indice de masse corporelle
- Rythme cardiaque
- Glucose

L'ensemble de données que nous utilisons provient de l'étude Framingham Hearts qui a débuté en 1948 et qui a mis au jour de nombreuses associations ou facteurs de risque liés à la maladie coronarienne.

La base de données est basée sur 4 238 sujets, chacun avec 15 mesures descriptives (caractéristiques) telles que le cholestérol, la pression artérielle et la fréquence cardiaque en plus d'une étiquette de sortie - si un diagnostic de coronaropathie a été posé plus de 10 ans.

Annexe : http://www.who.int/mediacentre/factsheets/fs317/en/

I. Préparation et Apprentissage

1. Importer les packages python sklearn, numpy et pandas et préparer votre environnement
2. Préparer vos données pour l'analyse en chargeant votre ensemble de données de l'étude Framingham Hearts et en séparant vos features (X) des labels (Y)
3. Écrire une fonction qui vous permet de visualiser vos données sous forme d'histogramme.
4. Utiliser les fonctions de scikit-learn pour apprendre votre modèle en se basant sur la régression logistique.

II. Analyse des résultats

Dans cette 2ème partie, nous examinerons la précision, la sensibilité, la spécificité et l'ASC (Aire sous la courbe) à l'aide du tracé de la courbe ROC.

En Machine Learning, la mesure de la performance est une tâche essentielle. Ainsi, s'agissant d'un problème de classification, nous pouvons compter sur une courbe AUC - ROC.

Lorsque nous devons vérifier ou visualiser les performances du problème de classification multi-classes, nous utilisons la courbe ROC AUC (Area Under The Curve).

Petit Rappel :

- La sortie d'une Régression logistique est une probabilité
- Si cette probabilité est supérieure à 0.5, les données sont étiquetées 1 (anormal) sinon 0 (normal).
- Le seuil de régression logistique par défaut est de 0.5
- **ROC** est la caractéristique de fonctionnement du récepteur. Sur cette courbe, l'axe des ***x*** est le taux de faux positif et l'axe des ***y*** est le taux de vrai positif.
- Si la courbe du tracé est plus proche du coin supérieur gauche, le test est plus précis.
- Le score de la courbe de ROC est la surface de calcul sous la courbe à partir des scores de prédiction.
- Nous voulons que auc se rapproche 1
    - fpr (False Positive Rate) = taux de Faux Positifs
    - tpr (True Positive Rate) = Taux de Vrai Positifs

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

1. Calculer le taux de précision, la spécificité et la sensibilité des résultats
2. Visualiser la courbe ROC et calculer l'ASC (Aire sous la courbe).
    a. Conclusions ?