HELHa
Haute École Louvain en Hainaut

Régression sur des cerveaux

Phase initiale

Selon le fichier de données ***brains.txt***[1], vous devez développer un modèle qui permettra de prédire le poids d'un corps en fonction de la masse de son cerveau.

Soit : $bodyWeight = brainWeight * x + intercept$

Ce document est décomposé en 3 colonnes (ces dernières sont séparées par des *tabulations*) :

- L'indice de l'information ;
- Le poids du cerveau ;
- Le poids du corps.

Chargez les données dans une variable nommée "brains". À la suite de cette opération, supprimez la colonne "Index".

Phrase de compréhension

L'objectif est d'assimiler la nature de nos données.

D'abord, faites appel à quelques méthodes pour comprendre rapidement les principales informations.

Ensuite, grâce à un nuage de points, analysez la relation de vos données.

Remarquez-vous des données extravagantes ? Si oui, enlevez-les et réaffichez votre graphique.

Après cela, stockez dans la variable "x"[2] les données qui se rapportent à "Brain Weight" et dans "y" celles qui concernent "Body Weight".

Enfin, quelle fonction pourrait correspondre au mieux à la relation de nos x et y ?
Rien ne vous empêche de faire des tests sur vos données pour trouver la fonction optimale à employer.

---

[1] Source : http://people.sc.fsu.edu/~jburkardt/datasets/regression/x01.txt

[2] Nous mettons cette fois-ci la variable "x" en minuscule car il s'agit d'un vecteur et non pas d'une matrice.

Phase de création du modèle

Vous produirez deux instances d'une régression linéaire.

La première respectera le format qui vous a été donné lors de la phase initiale.

L'autre suivra la fonction que vous avez trouvée lors de la phase de compréhension.

Pour tester au mieux nos résultats, vous emploierez la fonction "cross_val_score" en lui envoyant ces paramètres :

- "cv" : cela sera une instance de KFold ("n_splits = 5, shuffle = True, random_state = 1"),
- "scoring" : à vous de le deviner.

Vous devez, **sans hésitation**, pouvoir constater quel est le meilleur modèle.

Phase de prédiction

Pour chaque modèle, prédisez le poids d'un corps en fonction de la taille des cerveaux suivants :

- 44.5
- 8.1
- 423

Gardez à l'esprit que vous allez certainement **transformer ces données** pour entrainer votre second modèle.

Comparez les résultats de chaque modèle.

Phase de représentation

Pour chaque modèle, dessinez un graphique qui présente les informations suivantes :

- Un nuage de points (la relation entre Body et Brain Weight) ;
- Votre droite de régression.

Grâce à vos graphiques, que remarquez-vous ?

Phase d'argumentation

Répondez aux questions suivantes :

- Parmi nos modèles, l'un tire clairement son épingle du jeu, pourquoi ?
- La transformation de données est une solution pour répondre à quel problème ?
- Pourquoi n'avons-nous pas utilisé "GridSearchCV" ?