02. Fiche d'auto-évaluation

September 14, 2023

1. Pour quelle raison, qund on a N caractéristiques, on représente l'i-me exemple avec un vecteur $\mathbf{x}^{(i)}$ de $N+1$ éléments?

2. Le tableau suivant[1] montre les coefficients obtenus par la méthode des moindres carrées ordinaire (Ordinary Least Square) afin de prédire les prix des logements à partir des caractéristiques des logements et des quartiers.

Log(DensPop)	0.040 *,
Unemployment rate	−0.005 °
Social diversity index	−0.469
log(dist_school)	−0.030 *,
log(dist_park)	0.017
log(dist_shop)	0.017

   Écrivez l'équation de tel modèle. Quelle est la signification de chaque coefficient? Est-ce que les coefficient ont tous le signe (positif/négatif) comme pourrait s'y attendre?

3. Pour évaluer la qualité d'un modèle, est-il plus important de regarder la valeur de la fonction de perte sur le jeu de données d'entraînement ou de test?

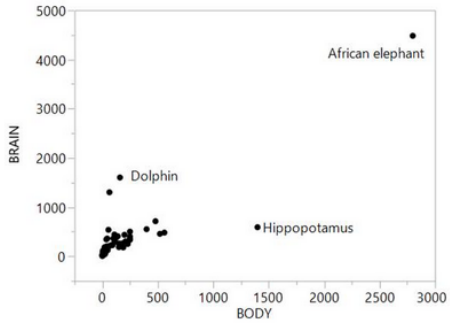
4. Imaginez de comparer deux modèles de régression. Le premier donne une erreur quadratique moyenne (root mean square error) de 55.8 sur les données d'entraînement et 20.4 sur les données de test. Le deuxième modèle donne une erreur de 73.2 sur les donnés d'entraînement et 15.9 sur les données de test. Lequel choisiriez-vous? Pourquoi?

5. Pourquoi est-il utile de fixer un *seed*, c'est-à-dire le paramètre random_state, quand on utilise la fonction train_test_split sur scikit learn?

6. Imaginez d'entraîner une regréssion linéaire pour inférer la taille du cerveau des animaux en fonction de leur taille. Le jeu de données et [2] as the one

[1] From Bulteau, Julie, Thierry Feuillet, and Rémy Le Boennec. "Spatial Heterogeneity of Sustainable Transportation Offer Values: A Comparative Analysis of Nantes Urban and Periurban/Rural Areas (France)." Urban Science 2.1 (2018): 14.

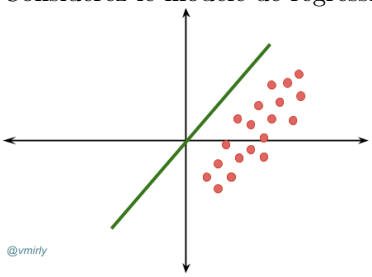
[2] From `https://statweb.calpoly.edu/bchance/stat302/hw/hw5sols.html`



below

Y a-t-il des traitement de données qu'on peut appliquer, avant de commencer l'entraînement, afin de améliorer la précision de l'inférence?

7. Considérez le modèle de régression linéaire représenté par la ligne verte.[3]



Pouvez-vous l'améliorer en changeant seulement un paramètre? Si oui, comment?

8. Quel est l'avantage d'utiliser l'erreur quadratique moyenne (root mean squared error - RMSE) à la place du carré moyen des erreurs (mean squared error - MSE)?

9. Est-ce que les coefficients obtenus par la méthodes des moindres carrées ordinaire (ordinary least square - OLS) change à chaque fois qu'on change les données d'entraînement?

10. Imaginez que vous devez décider quelle caractéristiques il faut inclure dans un modèle obtenu par la méthode des moindres carrées ordinaire (ordinary least square - OLS). Supposons que l'erreur explose quand on inclut plus de caractéristiques. Quelle pourrait en être la cause?

11. Expliquez comment la cross-validation marche.

[3] Picture from Vahid Mirjalili - `https://qr.ae/pN9G1S`