02. Self-assessment questions

January 17, 2022

1. Why, if we have N features, we represent each i-th sample with a vector $\mathbf{x}^{(i)}$ of $N+1$ elements?

2. The following table[1] shows the coefficients of a Ordinary Least Square (OLS) model to predict house prices based on the neighborhood characteristics.

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

   Write the equation of such a model. What is the meaning of each coefficient? If you look at the sign of each coefficient, are they the signs you would expect?

3. What is more important to judge the quality of your model, the loss on the training or the test set?

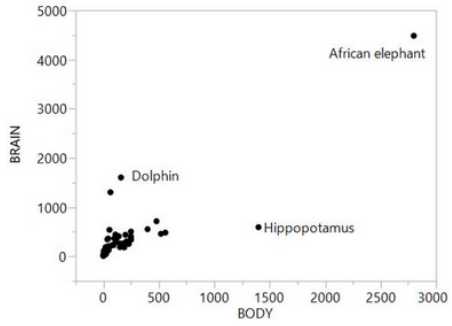
4. Suppose you have two regression models. The first has Root Mean Square Error (RMSE) 55.8 on the training set and 20.4 on the test set. The second has RMSE 73.2 on the training set and 15.9 on the test set. Which one would you prefer? Why?

5. Why is it useful to set a seed (parameter random_state) when using train_test_split function of scikit learn?

6. Suppose you want to train a linear regression to predict the brain size based on the body size of animals. The dataset is[2] as the one below

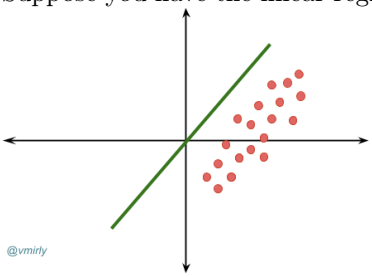
[1] From Bulteau, Julie, Thierry Feuillet, and Rémy Le Boennec. "Spatial Heterogeneity of Sustainable Transportation Offer Values: A Comparative Analysis of Nantes Urban and Periurban/Rural Areas (France)." Urban Science 2.1 (2018): 14.

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



Is there any pre-processing you can do, before starting training, to improve the accuracy of the prediction?

7. Suppose you have the linear regression model depicted by the green line.[3]



Can you improve it by just changing one parameter of your model? How?

8. What is the advantage of using the Root Mean Square Error (RMSE) instead of the Mean Square Error (MSE)?

9. Do the parameters obtained from the Ordinary Least Square (OLS) change if we change the training set?

10. Suppose you are deciding which features to include in an OLS model and you observe that the error explodes when adding more features. What could be the cause?

11. Explain how cross-validation works

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