DATA7703 Practical 10

2021 Semester 2

1. We study how the hyper-parameters of the RANSAC regressor affect its performance on the `diabetes` dataset in this question.

(a) Load the `diabetes` dataset from `sklearn.datasets` and construct a random 70/30 train-test split.

In addition, read the documentation of the `load_diabetes` function closely, and find a way to print out the full description of the dataset. According to the dataset description, how many features are there, and what is the target variable?

For the entire dataset, compute the means and standard deviations of the features, and the range of the target variable.

(b) Train an OLS model and report its training and test R^2 values and MSEs.

(c) Train a RANSAC regressor using the default hyper-parameters, and measure its training and test MSEs. Repeat this for 10 times and report the averages and standard errors of the training and test MSEs.

Comment on the performance of RANSAC as compared to the OLS model in (c).

(d) Repeat (c) by setting `min_samples` to each of 0.05, 0.10, 0.15, ..., 1.0 for the RANSAC regressor. Plot the mean training and test MSEs against `min_samples`. Show the error bars on the plots, with the error bar sizes being the standard errors. Comment on the effect of `min_samples`.

(e) Train a support vector regressor on the `diabetes` dataset. Tune the hyper-parameters of your model so as to obtain good generalization performance. Describe how you do this, and note down the hyper-parameters of the best support vector regression model. Report the training and test MSEs for the chosen model.

(f) Repeat (d) using the support vector regressor as the basis models. Use the hyper-parameters that you choose in (e) for the support vector regressors. Compare the plot with that in (d).