COMP3222/6246 Coursework 1

Machine Learning with Python & Scikit-Learn

Consider the diabetes data set from the scikit-learn python package. The structure of the data set is as follows: Each data point consists of 10 baseline variables (i.e., feature vector x): age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each diabetes patient, as well as the response of interest (i.e., output variable y), a quantitative measure of disease progression one year after baseline. Your task is to identify a good machine learning model to predict the value of y from a given feature vector x. To do so, you can use the above mentioned data set to train your model.

1. What is the python code to load the diabetes data set?
2. Uniformly randomly split the data set to training and test data with 80%for training and 20% for testing. Perform linear regression on the training data. What are the model parameters of the solution? What is the RMSE for this model on the training set? Apply this to the test data as well. What is the RMSE there?Repeat the uniformly random split 10 times and write down in a table each time the model parameters, RMSE for training, and RMSE for test data. How much do these values differ?
3. Investigate whether linear regression is an appropriate model for this data set. Does linear regression fit the data well? Explain your answer in 1-2 paragraphs (i.e., what can you say from the plot?). Hint: gradually increase the size of training data set, and plot the RMSE for both training and test data.
4. Use polynomial regression with degrees $d = 2, 5$ and 10 (still with 80%of data for training and 20% for testing). Plot their predictions in the same plot. Explain what you can see within 1-2 paragraphs.
5. Assuming that we still use 80% of the diabetes data set for training and 20% for testing, evaluate and explain the performance of Lasso with different α values within 1-2 paragraphs. (You can use up to 1 figure if it is necessary.)
6. What would be the optimal value of α for Lasso on this data set? What are the resulted model parameters? Hint: use cross-validation and grid search to find the optimal α.
7. Implement a decision tree as a regression model for this data set. Repeat this 10 times. Visualise the best tree up to depth 5 and report the RMSE on both training and test data. What is the reason that we might get different trees and results? Please explain within 1-2 paragraphs.