ESE 542: Statistics for Data Science

Chapter 3 Project

In addition to answering the **bolded** questions on Coursera, also attach your notebook, both as `.ipynb` and `.html`.

In the following exercise, we will perform linear regression to fit various data sets and to predict outputs. Perform the following analyses by starting a new notebook.

1 Part A

First, we will use the Ch3PartA dataset to generate polynomial regressions using `scikitlearn`. This dataset contains 100 observations of points x and their corresponding response, y. The data is divided into a training set (x_{tr}, y_{tr}) and a test set (x_{te}, y_{te}).

1. Load `Ch3PartA.csv` into your notebook.
2. Create a scatter plot of: (a) y_{tr} against x_{tr} and another of (b) y_{te} against x_{te}. Notice the similarities and differences between the plots. **Answer on Coursera: What is the maximum value of y in the training set? In the test set?**
3. Generate the necessary features to fit polynomial regressions up to the 20^{th} degree (up to and including the x^{20} term) on the training data. *Hint*: You will be fitting multivariate linear regression models with polynomial features of x. Familiarize yourself with `sklearn.preprocessing.PolynomialFeatures`.
4. Calculate the training MSE and the test MSE for 20 polynomial models up to degree 20. Store these as lists named `mse_train` and `mse_test`respectively. *Hint*: Familiarize yourself with `sklearn.metrics.mean_squared_error` and try to automate the process.
5. Generate a plot of both the training MSE and test MSE against flexibility for degrees 1 to 20. **Answer on Coursera: What is the minimum training MSE? Test MSE?**
6. From your plot, make an educated guess about the polynomial degree of the function that was used to generate the data. Then, give an estimate of the irreducible error Var(ϵ) for the

optimal model on both the training set and test set. *Hint*: Revisit the section on hypothesis testing. Think about the relationship between MSE, RSS, and RSE. **Enter your answers on Coursera.**

2 Part B

Next, we will use the Ch3PartB dataset to observe the effects of collinearity using `statsmodels`. This dataset contains 100 observations of points x_1, x_2, and y, the response variable.

1. Load the data from `Ch3PartB.csv` into a `pandas DataFrame`.
2. Show a scatterplot displaying the relationship between x_1 and x_2? **Answer on Coursera: What is the correlation coefficient between x_1 and x_2?**
3. Using the data, fit a least squares regression to predict y using x_1 and x_2. Describe your results. *Hint*: Familiarize yourself with `statsmodels.formula.api.ols`. **Answer on Coursera: What are the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? At a 95% confidence level, can you reject the null hypothesis $H_0 : \beta_1 = 0$? What about $H_0 : \beta_2 = 0$?**
4. Now fit a least squares regression to predict y using only x_1. Comment on your results. **Answer on Coursera: Can you reject the null hypothesis $H_0 : \beta_1 = 0$?**
5. Fit a least squares regression to predict y using only x_2. Comment on your results. **Answer on Coursera: Can you reject the null hypothesis $H_0 : \beta_2 = 0$?**
6. Do Part B Questions 3-5 contradict each other? Explain why.