Assignment Eleven

BSDS Spring 2021

(due 4/30, 11:59pm PST)

Note: You must submit this assignment as BOTH a Jupyter Notebook file (.ipynb) and an (.html) file on Canvas. All of the following must be satisfied.

- The filenames must be of the form

  `[last_name]_HW11.ipynb` and `[last_name]_HW11.html`

- You must include your name and the assignment number in a Markdown cell at the beginning of the notebook
- You must separate questions using Markdown cells
- If a question requires a short answer rather than code, use a markdown cell.

1. We will be investigating Petal Width versus Petal Length of various irises to construct a linear model that best predicts Petal Length (y) versus Petal Width (x).

(a) Import the `iris_data.csv` from Canvas for the following problems. This is subset of the famous `iris` dataset from the UCI Machine Learning Repository (available in base R).

(b) Before running any other commands, set a random seed using `set.seed()` to ensure that your results are reproducible.

2.

(a) Split the `iris_data` dataset into a training set and a validation set (80% goes in training set and 20% goes in validation set).

(b) Fit a linear model to your training data using Root Mean Squared Error (RMSE) and create a scatter plot showing the linear model against the training set. (Let Petal Width be "x" and Petal Length be "y")

(c) Evaluate the RMSE on the validation set and create a scatter plot showing the linear model against the validation set.

3.

(a) Create new training set and validation set split, but this time only choose iris of Species `Iris-virginica` for your validation set. (Hint: you can do this by taking the first 80% of the dataset for training)

(b) Repeat (1b) and (1c) for your new training and validation sets.

(c) Is the RSME score better or worse? Why do you think so?

4. Repeat Question 2b and 2c after filtering the dataset to consist of only flowers with Species `Iris-virginica`.

5.

(a) What's the benefit of using the full dataset to create a linear model? What's the benefit of creating a linear model for each Species?

(b) Why might it be bad practice to compare the RMSE scores on the validation set obtained in (2c) to those obtained in (4)?

(c) Create a scatter plot of Petal Width versus Petal Length for the entire dataset, coloring each point by Species. Based on this plot, do you think it is better to have a separate linear model for flowers of Species `Iris-virginica` or use a single linear model fit to the whole dataset?

(d) Import the test set and compute the RMSE for your models from Question 2 and 4 on the full dataset. Which does the best?

(e) Create a new test set from the old test, filtering for only flowers of Species `Iris-virginica`. Compare the RSME for the linear model from Question 2 with the model from Question 4. Did you confirm your answer in part c?