Names:

DATA 311 – Lecture 19 – Data Splits, Regularization, Hyperparameter Tuning

Suppose you aim to predict a student's final exam score given their scores on all the assignments in the first half of the quarter. For each of the following, say whether it is most likely to **decrease** (a) bias, (b) variance, or (c) irreducible risk.

1. Add additional information, such as amount of time spent studying, to your feature set.

2. Use a more powerful/flexible model.

3. Use a less powerful/flexible model.

The following questions pertain to the Jupyter notebook (`L19.ipynb`) linked from the course webpage. Download the notebook, open it up in JupyterHub, and answer the following questions. The notebook has 4 code cells with headings numbered 0 through 3. Answer the following questions without modifying the code.

4. Read the code for cell 0. What fraction of the entire labeled dataset will be used for training?

5. Read the code for Cell 1. How many different classifiers are trained?

6. Explain why `best_C` is chosen to be the value of `C` that performs on the validation set instead of on the training set.

7. Run the code for Cells 0, 1, and 2, and read the code for Cell 2. Which line (orange or blue) do you think is training accuracy, and which line is validation accuracy?

8. What do you expect the value of `best_C` to be?

9. In the region to the **right** of `best_C` (higher values of C), is the model overfitting or underfitting?

10. In the region to the **left** of `best_C` (higher values of C), is the model overfitting or underfitting?

11. Explain the relationship between the model's C hyperparameter and the model's flexibility.

12. Read the code for Cell 3 and run it. Comment on the results: is this what you expected to see? Why or why not?