CSAI 801 Lab 3: Bias and Variance

1. Calculating Bias and Variance

In this question you are going to calculate the bias and variance of your trained model.

1.1 How to Re-sample data

You have been provided a dataset consisting of pairs (x_i, y_i). It can be loaded into your python program using pickle.load() function. Split the dataset into training and testing(90:10 split). Now divides the train set into 10 equal parts randomly, so that you will get 10 different dataset to train your model.

1.2 Task

After re-sampling data, you have 11 different datasets (10 train sets and 1 test set). Train a linear classifier on each of the 10 train set separately, so that you have 10 different classifiers or models. You have 10 different models or classifiers trained separately on 10 different training set, so now you can calculate the bias and variance of the model. You need to repeat the above process for the following class of functions.

1.2.1 $y = mx + c$

1.2.2 $y = ax^2 + bx + c$

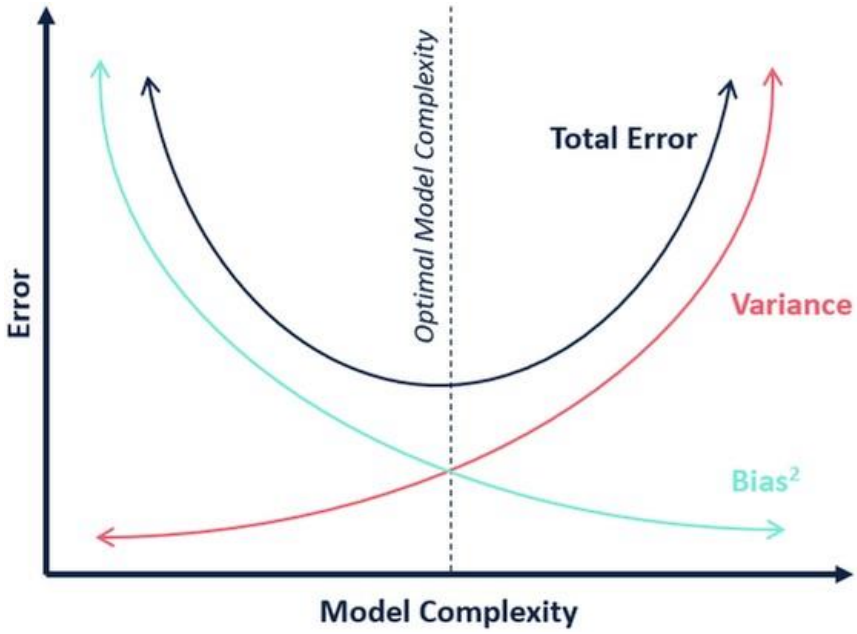
1.2.3 $y = ax^4 + bx^3 + cx^2 + dx + e$

And so on up till polynomial of degree 9. **You are only supposed to use sklearn's linear_model.LinearRegression().fit() for finding the appropriate coefficients with the default parameters.** Tabulate the values of bias and variance and also write a detailed report explaining how bias and variance changes as you vary your function classes.
Note: Whenever we are talking about the bias and variance of model, it refers to the average bias and variance of the model over all the test points.

2. Bias-Variance

Task: You have been provided with a training data and a testing data. You need to fit the given data to polynomials of degree 1 to 9 (both inclusive). You are only supposed to use *sklearn.linear_model.LinearRegression().fit()* with the default parameters to fit the model to your data.

You might need to play around with the data for this. Check out sklearn.preprocessing.PolynomialFeatures.



Specifically, you have been given 20 subsets of training data containing 400 samples each. For each polynomial, create 20 models trained on the 20 different subsets and find the variance of the predictions on the testing data. Also, find the bias of your trained models on the testing data. Finally plot the bias-variance trade-Off graph.

Note: You do not need to plot the curve for total error. The formula for bias and variance are for a single input but as the testing data contains more than one input, take the mean wherever required. Write your observations in the report with respect to underfitting, overfitting and also comment on the type of data just by looking at the bias-variance plot.