MDL Assignment 1

Deadline: 14^{th} February

1 Bias-Variance trade off

Whenever we discuss model prediction, it's important to understand prediction errors (bias and variance). There is a trade-off between a model's ability to minimize bias and variance. Gaining a proper understanding of these errors would help to distinguish a layman and an expert in ***Machine Learning****. So, instead of playing around with a number of classifiers, let's understand how to select which classifier to use.*
Let's get started and understand some of the basic definition.

- **Bias** is the difference between the average prediction of our model and the correct value which we are trying to predict. A model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to a high error on training and test data.For a more detailed definition refer this article

$$Bias^2 = (E[\hat{f}(x)] - f(x))^2$$

where $f(x)$ represents the true value, $\hat{f}(x)$ represents the predicted value

- **Variance** is the variability of a model prediction for a given data point. Again, imagine you can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.For a more detailed definition refer this article

$$Variance = E\left[(\hat{f}(x) - E[\hat{f}(x)])^2\right]$$

where $f(x)$ represents the true value, $\hat{f}(x)$ represents the predicted value

- **Noise** is a unwanted distortion in data. Noise is anything that is spurious and extraneous to the original data, that is not intended to be present in the first place, but was introduced due to faulty capturing process.For a more detailed definition refer this article

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand, if our model has a large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data. Now moving on to the second crucial part of the assignment "Data Visualisation".

2 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. The problem is, it's often challenging to choose the right visualization for the data you want to show.The best way to show the bias-variance trade off is through the bull's eye diagram as shown below.

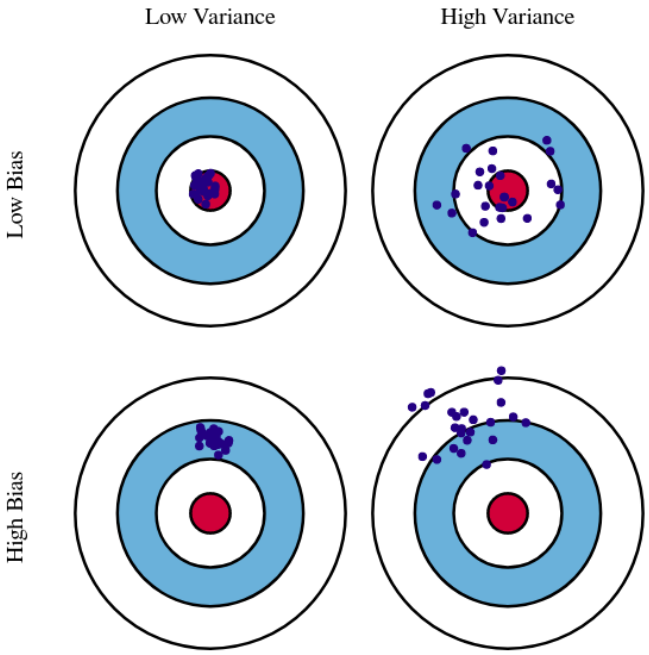


Figure 1: Graphical illustration of bias and variance.

3 Calculating Bias and Variance

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

3.1 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.

3.2 Task

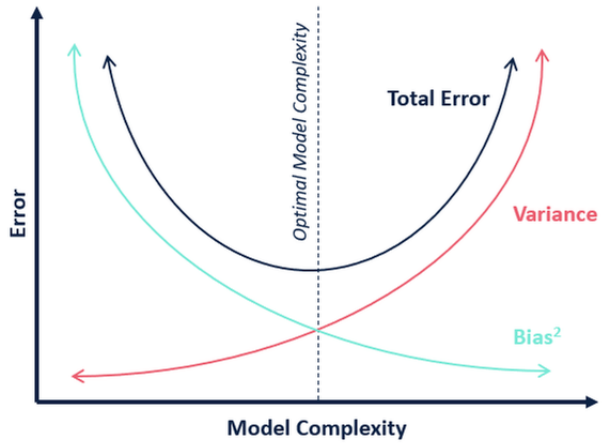
After re-sampling data, you have 11 different datasets (10 train sets and 1 test set). Train a linear classifier on each of the 10 train set separately, so that you have 10 different classifiers or models. You have 10 different models or classifiers trained separately on 10 different training set, so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

- $y = mx + c$
- $y = ax^2 + bx + c$
- $y = ax^4 + bx^3 + cx^2 + dx + e$

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.

4 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.

5 General Instructions

- Marks distribution
  - Question 1: 33%
  - Question 2: 33%
  - Viva: 34%
- The data is in numpy array format.
- Submit a zip file name **TeamNumber_assgn1.zip** containing source code and the report.
- All coding questions have to be done in Python3 only.
- Get familiar with numpy, matplotlib, pickle, pandas dataframe and sklearn.
- You are expected to write vectorized codes, as it will be a part of grading.
- Plagiarism will be penalized heavily.
- **Report should contain details of algorithm implementation, results, tables, plots, observations and answers to the subjective questions (if any)**.
- No deadline extension.
- Manual evaluations will be held regarding which further details will be announced later.