Machine Data and Learning
Assignment 2

Maximum Marks: 100

Deadline: 11:55 PM, 17th February, 2023

1 Introduction

1.1 Bias-Variance trade-off

When we discuss model prediction, it is important to understand the various prediction errors - bias and variance. There is a trade-off between a model's ability to minimise bias and variance. A proper understanding of these errors would help in distinguishing a layman and an expert in Machine Learning. Before using different classifiers, it is important to understand how to select a classifier to use.

Let us get started and understand some basic definitions that are relevant.

For basic definitions, when $\hat{f}$ is applied to an unseen sample, x refer here.

- **Bias** is the difference between the average prediction of our model and the correct value that we are trying to predict. A model with high bias does not generalise the data well and oversimplifies the model. It always leads to a high error on training and test data.

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

  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 realisations of the model.

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

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

- **Noise** is any 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 a faulty capturing process.

- **Irreducible error** is the error that cannot be reduced by creating good models. It is a measure of the amount of noise in the data. Here, it is important to understand that no matter how good we make our model, our data will have a certain amount of noise or irreducible error that cannot be removed.

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

where $f(x)$ represents the true value, $\hat{f}(x)$ represents the predicted value, $E[(f(x) - \hat{f}(x))^2]$ is the mean squared error(MSE) and σ^2 represents the irreducible error.

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 is going to have high variance and low bias. So we need to find the right (or good) balance without overfitting and underfitting the data.

1.2 Linear Regression

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

- Simple regression
- Multivariable regression

For a more detailed definition refer this article.

For a simple linear regression model with only one feature the equation is:

$$y = w_1 x + b$$

where,

- y = Predicted value/Target Value
- x = Input
- w_1 = Gradient/slope/Weight
- b = Bias

For a Multivariable regression model the equation is:

$$y = b + \sum_{i=1}^{n} w_i x_i$$

Once we have the prediction function we need to determine the value of weight/s and bias. To see how to calculate the value of weight/s and bias, refer this article.

2 Tasks

2.1 Task 1: Linear Regression

Write a brief about what function the method `LinearRegression().fit()` performs.

2.2 Task 2: Gradient Descent

Explain how gradient descent works to find the coefficients. For simplicity, take the case where there is one independent variable and one dependent variable.

2.3 Task 3: Calculating Bias and Variance

A large multinational corporation recently underwent a round of layoffs, affecting thousands of employees across different locations. The HR department wants to understand why some employees were laid off while others were retained. They have the data of the performance metric of the employees and the risk of them getting fired from the company. In this task, you need to help the HR to find the bias and variance of a trained model which can help her to analyse the risk factor of the employees.

2.3.1 How to Re-Sample data

The HR is given with two datasets, i.e, train set and test set, consisting of pairs $(x_i; y_i)$. x_i corresponds to the performance score of the employee, while y_i corresponds to the risk score of the employee. This data can be loaded into your python program using the *pickle. load()* function. You then need to divide the train set into 20 equal parts randomly, so that you get 20 different train datasets to train your model.

2.3.2 Task

After re-sampling the data, you have 21 different datasets (20 train sets and 1 test set). Train a linear classifier on each of the 20 train sets separately so that you have 20 different classifiers or models. Now you can calculate the bias and variance of the model using the test set. You need to repeat the above process for the following class of functions,

- $y = mx + c$
- $y = ax^2 + bx + c$
- $y = ax^3 + bx^2 + cx + d$

And so on, up till polynomials of degree 15. The only two functions that you are allowed to use are (from `sklearn`):

- `linear model.LinearRegression().fit()`
- `preprocessing.PolynomialFeatures()`

These functions will help you find 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 change as you vary your function classes.

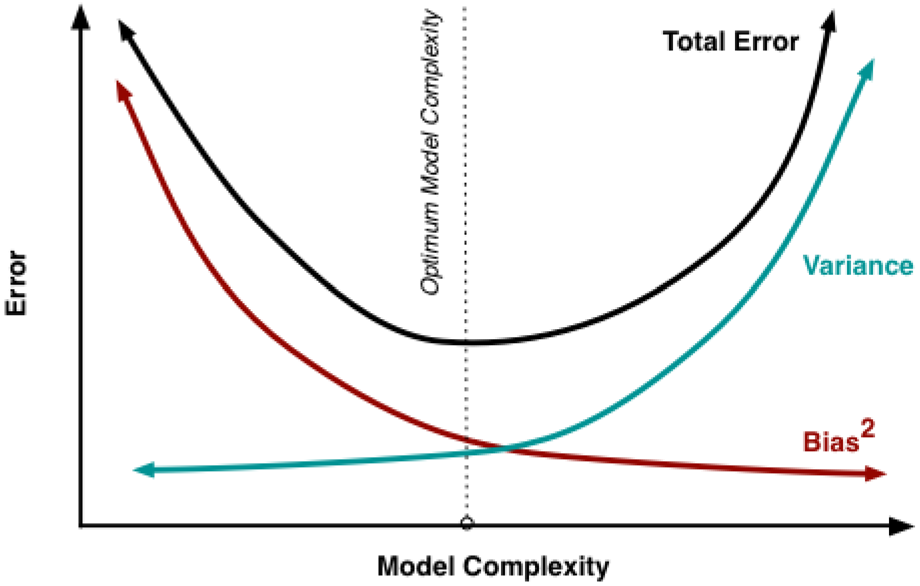
Note: Whenever we are talking about the bias and variance of the model,
it refers to the average bias and variance of the model over all the test points.

2.4 Task 4: Calculating Irreducible Error

Tabulate the values of irreducible error for the models in Task 2 and also write a detailed report explaining why or why not the value of irreducible error changes as you vary your class function.

2.5 Task 5: Plotting *Bias*^{2} - *Variance* graph

Based on the variance, bias and total error calculated in earlier tasks, plot the $Bias^2 - Variance$ tradeoff graph and write your observations in the report with respect to underfitting, overfitting and also comment on the type of data just by analysing the $Bias^2 - Variance$ plot. The below figure shows the balance between model framework error and model complexity.



Plot variation of $Bias^2$, Variance and MSE against degree of polynomial in the same graph.

Note: The formula for $Bias^2$ and Variance are for a single input, but as the testing data contains more than one input, take the mean wherever required. You need to plot the graph for polynomials of up to degree 10 only. (Plotting higher degrees makes the graph difficult to interpret).

3 Bonus

We have provided you with the data of a discharging capacitor together with a loop containing a resistor. Charge on the capacitor (dependent variable) is a function of time(independent variable) and varies exponentially according to the following equation:

$$Q = CV_0e^{\frac{-t}{RC}}$$

Given $V_0 = 5V$, perform linear regression on the data and report the values of Capacitance(C) and Resistance(R).

Note: You cannot directly perform linear regression since the function is an exponential one. You have to figure out another way to use linear regression on the dataset.

4 General Instructions

- The data is in numpy array format.
- Submit a zip file name rollnumber_assgn2.zip containing source code and the report:
  - code.ipynb
  - bonus.ipynb (if done)
  - report.pdf
  - readme.md (if any assumptions)
- All coding has to be done in Python3 only, using Jupyter Notebook.
- Report should include all details needed for evaluation. Please include relevant graphs, tables, analysis, observations and writeup as required for each of the tasks above.
- Get familiar with numpy, matplotlib, pickle, pandas dataframe and sklearn.
- You should write vectorized code which performs much better compared to individual iteration.
- Plagiarism will be penalised heavily.
- Manual evaluations will be held regarding which further details will be announced later.

5 Marking Scheme

- Task 1: 5 marks
- Task 2: 5 marks
- Task 3: 30 marks
- Task 4: 10 marks
- Task 5: 20 marks
- Viva: 30 marks
- Bonus: 20 marks

Note: Marks lost in any task can be covered by bonus. However, bonus will not compensate for any marks lost in Viva. The maximum marks is **100** for this assignment.