COMP 4983: Lab Exercise #4 Mark: /50

[Due: Sep 30, 2022 @1730
Assignment Submission
Folders]

Instructions:

In this lab, you will explore the bias and variance of a polynomial regression model using simulated data. Recall from Module 3 that the observed data points, $\hat{y}$, are generated by a true function f that maps an input x, plus some observation noise, ϵ, i.e., $\hat{y} = f(x) + \epsilon$.

In this lab, we will assume the following:

- $f(x) = x + sin(3x)$
- ϵ is a random variable following a zero-mean Gaussian distribution with standard deviation, σ = 0.25

You will approximate the true function f with a polynomial regression model $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_i^j$, $\forall i = 1, 2, \ldots, N$, where p represents the complexity (p-th degree polynomial) of the polynomial regression model.

The goal of this lab is to illustrate how the bias and variance of a polynomial regression model vary as a function of p. This will be done by

- generating 1000 independent observation datasets
- fitting the polynomial regression model of degree p = $[1, 3, 5, 9, 15]$ to each dataset
- measuring the bias and variance at a given point, $x = 5$

Steps:

1) Download the simulated data generator script, *GenData_Lab4.py*, from BCIT Learning Hub (Content | Laboratory Material | Lab 4) and save it in your working directory. This script contains two functions that you will use for this lab:
   - **f(x)** - returns the value of $f(x)$
   - **genNoisyData()** - generates a random set of x (with 50 elements) and associated noisy observation, $\hat{y} = f(x) + \epsilon$

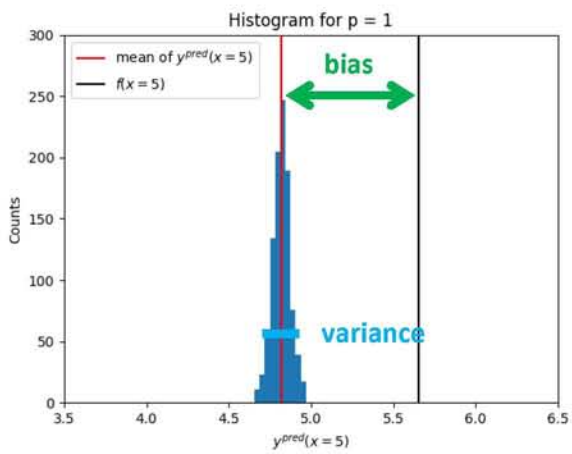
2) Create a new Python script using the filename B*iasVariance_Lab4.py* and save it in your working directory.

3) Include the following line at the top of your script, B*iasVariance_Lab4.py*:

```
import GenData_Lab4 as lab4
```

4) For each $p = [1,3,5,9,15]$, you will implement the following in your script, *BiasVariance_Lab4.py*:

   a. Use genNoisyData() to generate 1000 datasets, where each dataset contains $N = 50$ samples.
   b. For each dataset m, $\forall m = 1,2,\ldots,1000$,
      i. train a polynomial regression model of degree p on the data.
      ii. evaluate the trained model at $x = 5$, i.e., $y_m^{pred}(x = 5)$.
   c. Compute and output the bias for $x = 5$, which can be computed by $\overline{y^{pred}}(x = 5) - f(x = 5)$. Note that $\overline{y^{pred}}(x = 5) = \frac{1}{1000}\sum_{m=1}^{1000} y_m^{pred}(x = 5)$ is the average of $y_m^{pred}(x = 5)$ over the 1000 datasets.
   d. Compute and output the variance of $y^{pred}(x = 5)$, which can be computed by $\text{Var}\left(y^{pred}(x = 5)\right) = \frac{1}{1000}\sum_{m=1}^{1000}\left(y_m^{pred}(x = 5) - \overline{y^{pred}}(x = 5)\right)^2$, over the 1000 datasets.
   e. Plot the distribution, using a histogram, of $y_m^{pred}(x = 5)$ from all 1000 datasets. In addition, plot a vertical line to indicate $\overline{y^{pred}}(x = 5)$ (red line) and another vertical line to indicate $f(x = 5)$ (black line) in the example shown below for $p = 1$. Note that your exact values may differ as the noise is randomly generated.



5) Indicate in the output which p-th degree polynomial gives the smallest (lowest absolute) bias and the lowest variance, respectively, at $x = 5$.

<u>Deliverable</u>:

All work submitted is subject to the standards of conduct as specified in BCIT Policy 5104. No late assignments will be accepted.

[Sep 30, 2022 @1730] Ensure that your source code is adequately commented and submit using the filename *BiasVariance_Lab4.py* to BCIT Learning Hub (Laboratory Submission | Lab 4).