Homework 2

Write a computer program to evaluate the generalization error (GE), model prediction error (ME) and training error (TE) for the k-nearest neighbors learning approach. Use provided k-NN Python code.

Learning model:

In the k-nearest neighbors learning approach guess output, $\hat{Y}_k$, as:

$$\hat{Y}_k = \hat{f}_k\left(x\right) = \frac{1}{k} \sum_{i \in \left\{ i | x_i^{traning} \in N_k\left(x\right) \right\}} Y_i^{traning}$$

where $N_k\left(x\right)$ is a set of k nearest neighbors of x within the observed (training) data.

Data model:

Training data

Let us define observed (training) data model.

- First generate $N^{traning} = 50$ uniformly spaced feature samples in the interval from zero to one. That is $x_i = i\Delta x,\ i = 0,..,N^{traning} - 1$ where $\Delta x = 1/N^{traning}$.
- Next generate $N^{traning} = 50$ noise samples, n_i, as Normally distributed random process with a zero mean and 0.1 standard deviation.
  (Hint: Use npr.normal nympy function)

Now the observed noisy data can be calculated as:

$$Y_i^{traning} = f(x_i) + n_i,\ \ i = 1,..,N^{traning} \ \text{ where } \ f(x_i) = \sin\left(2\pi x_i\right).$$

Testing data

- To generate the testing data, $Y_j^{testing},\ \ j = 1,..,N^{testing}$, the same procedure described above can be used where N=300.

Evaluation:

The generalization error can be approximately calculated as:

$$GE\left(k\right) = E\left[\left(Y - \hat{Y}_k\right)^2\right] \simeq \frac{1}{N^{testing}} \sum_{j=1}^{N^{testing}} \left(Y_j^{testing} - \hat{f}_k\left(x_j^{testing}\right)\right)^2,$$

the model prediction error as:

$$ME\left(k\right) = E\left[\left(f\left(x\right) - \hat{Y}_k\right)^2\right] \simeq \frac{1}{N^{testing}} \sum_{j=1}^{N^{testing}} \left(f\left(x_j^{testing}\right) - \hat{f}_k\left(x_j^{testing}\right)\right)^2,$$

and the training error as:

$$TE\left(k\right) = \frac{1}{N^{traning}} \sum_{j=1}^{N^{traning}} \left(Y_j^{traning} - \hat{f}_k\left(x_j^{traning}\right)\right)^2.$$

Report:

Please provide a brief report, in a single pdf file, with a proper cover page. The cover page should contain: you name, ID number, homework number, class name and due date. In the report please show four graphs for $k \in \left\{1,5,10,15\right\}$ where each graph shows $Y_j^{testing}$,

$f\left(x_j^{testing}\right)$ and $\hat{f}_k\left(x_j^{testing}\right)$ as a function of x. Please provide a brief comment about these graphs. Next show a graphs of $GE(k), ME(k)$ and $TE(k)$ as a function of k, the number of nearest neighbors, followed by a brief comment of this graphs as well. Include your computer code at end of the report.