Assignment 1

Statistical Learning Theory and Data Science
Course Code: BI3424/DS3214

The assignment submission deadline is February 13, 2025, at 11:59 PM.

Question

Suppose data is generated from a target function $y = ((x-2)^2 + 4 + \text{noise})$, where the noise is defined as $\mathcal{N}(\mu = 0, \sigma = 0.2)$ (Here, $\mathcal{N}$ is the normal distribution). You can use the command `np.random.normal()` in Python for noise. One dataset of size $N = 3$ can be obtained from this process by sampling uniformly over $x \in [-10, 10]$: $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3). This data is to be fit using one of two model classes:

1. $\mathcal{H}_0$: The set of all constant functions $h(x) = b$.
2. $\mathcal{H}_1$: The set of all linear functions $h(x) = ax + b$.

Thus, a training set $\mathcal{D}$ has only 3 points, picked independently, and the learning algorithm determines the hypothesis that minimizes the in-sample least squared error, MSE. This process can be repeated for another three data points, and so on.

For each of the model classes $\mathcal{H}_0$ and $\mathcal{H}_1$ compute:

1. The hypothesis that best approximates f in the average sense.
2. Its bias and variance components.
3. The expected out-of-sample error.

What do you conclude about whether $\mathcal{H}_0$ or $\mathcal{H}_1$ is the more appropriate class for prediction? Why?

Instructions

1. Simulate the above experiment in Python and plot appropriate graphs.
2. Submit your code with appropriate extensions (.py or .ipynb).
3. Submit a document with your code to provide a brief explanation of your observations and results.
4. Your file should be named in the following manner: course-code_registration-number_name