Programming Assignment 1

Due Date: Sunday, October 13, 2024

1 Problem 1

In this problem, you'll fit sampled data using a polynomial function to demonstrate the bias-variance trade-off and overfitting. For the given objective function, sample the function values over a specific range, incorporating noise into the sampling process.

$$y = \sin(x) + \frac{x}{2}, \quad x \in [0, 6\pi]$$

The task is divided into two parts:

1. **Demonstrating Overfitting**

   Show an example of overfitting. As the degree (model complexity) increases, train and test errors initially decrease to a certain point. However, after that point, while the train error decreases, the test error starts increasing slightly. Visualize this as seen in Fig. 1, where the overfitting is evident.

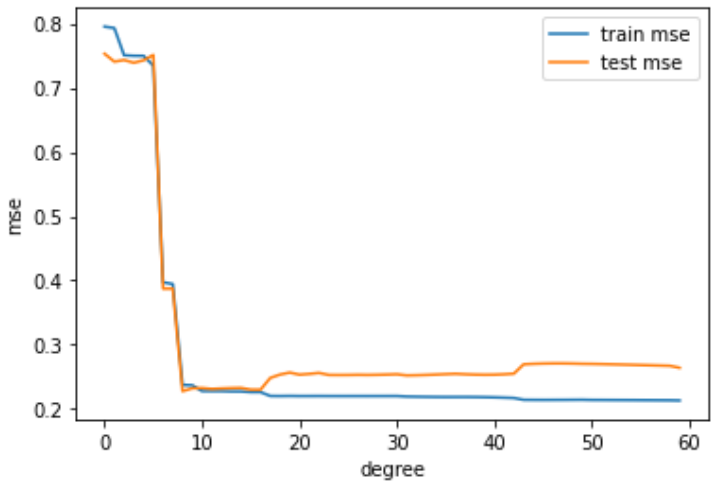

Figure 1: Example of overfitting

2. **Visualizing the Bias-Variance Trade-off and Error**

   Refer to the Fig. 2 and Fig. 3. Show that the total error can be expressed as:

$$Error = Bias^2 + Variance + Noise$$

Visualize the changes in bias, variance, and error as the degree (model complexity) increases. Explain the meaning of these changes. Additionally, visualize the polynomial function that corresponds to the degree with the smallest error, along with the answer function.

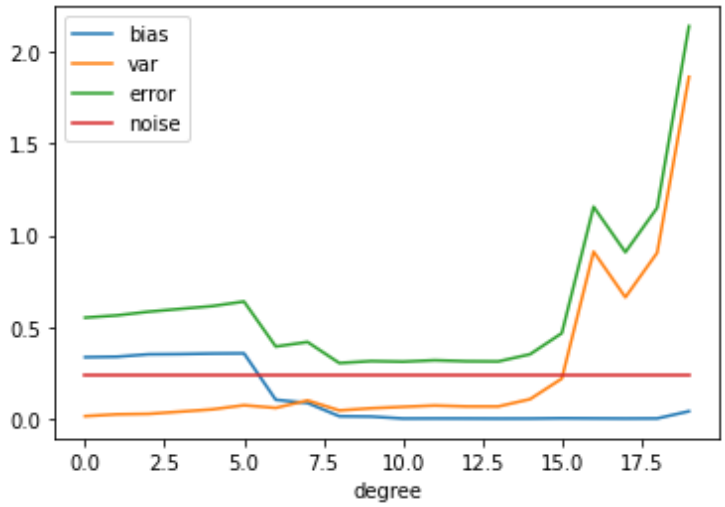

Figure 2: Example of error

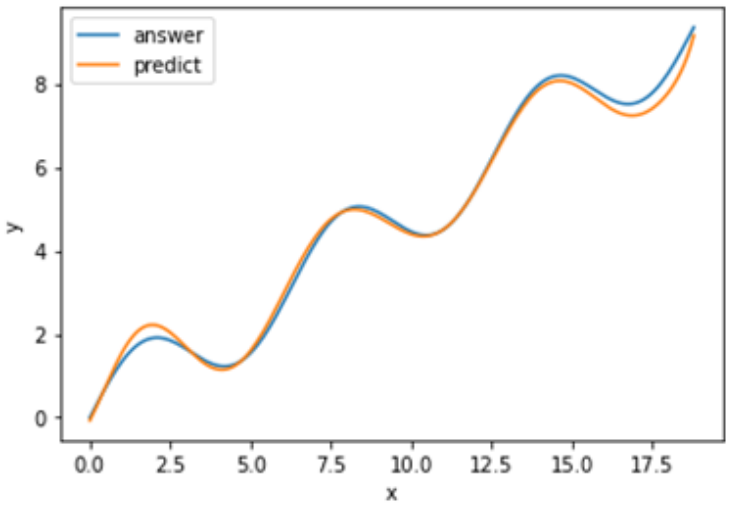

Figure 3: Example of visualizing an objective function and a function obtained through regression

Requirements

1. For sampling, determine the parameters through direct adjustment. Include reasoning for the chosen parameters in the report. Here are some example variables (the names

don't have to match exactly):

- `n_repeat`: The number of repetitions, determines how many times one 'train set' is sampled.
- `n_sample`: Total number of samples to extract.
- `ratio`: The proportion of the extracted samples to use as training data.
- `n_train`, `n_test`: The number of samples for training and testing, respectively, determined by `ratio`.
- `noise`: The range of noise (follows a normal distribution).

2. Sample extraction is done uniformly, and noise follows a normal distribution. Fig. 4 is an example of it (e.g., `n_sample` = 250 , `noise` = 0.5).

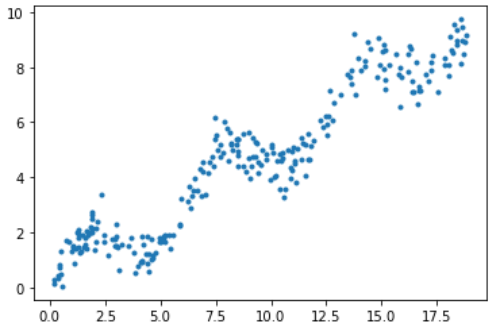

Figure 4: Example of sampling

3. For calculations and fitting, it is highly recommended to use `numpy` and `matplotlib` only. If you use packages other than `numpy` or `matplotlib` (e.g., `sklearn`), include a detailed description of those packages in the report. Please note that there will be a slight penalty for using additional packages.

2 Problem 2

In this problem, you will visualize the model using ROC and PR curves and find the appropriate threshold.
First, generate a binary classification dataset (Hint: use `make_classification` from `sklearn`.) and train a regression model. Then, visualize the model's performance using both the ROC curve and the PR curve. Use the F1 score from the PR curve to identify the optimal threshold. The visualizations should be presented as shown in Fig. 5, Fig. 6, and Fig. 7. You may use any package for this problem.

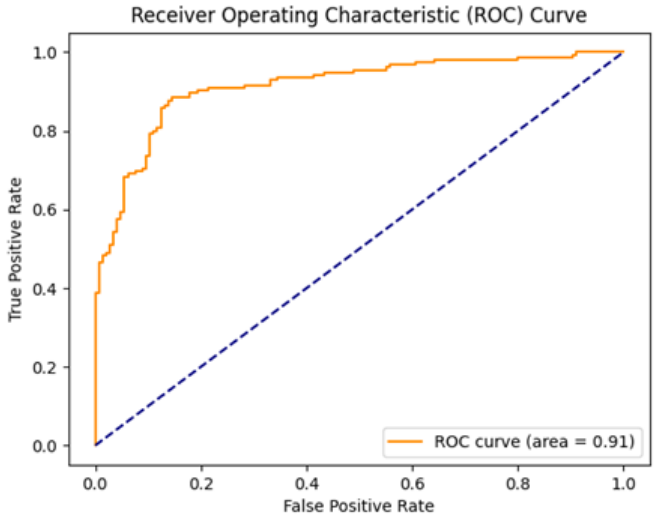

Figure 5: Example of ROC curve

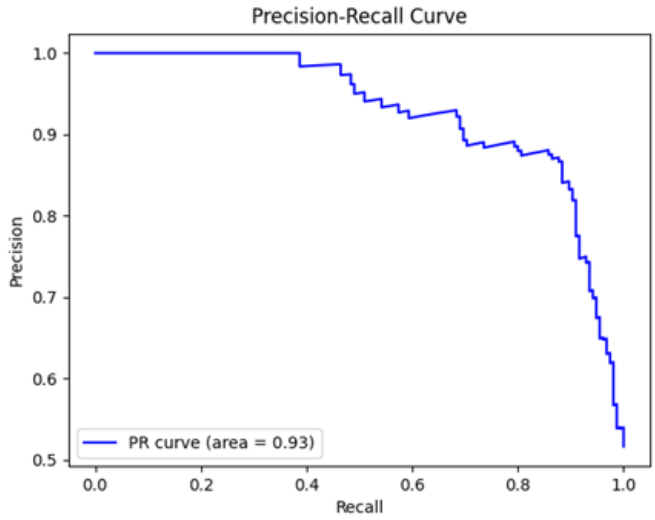

Figure 6: Example of PR curve

```
Optimal threshold for PR curve: 0.4292
```

Figure 7: Example of result output

Submission Guide

- **File Format** : PA1_*2024-12345_firstname_LASTNAME*.zip
- Inside the compressed file, there should be a report `.pdf` file and two `.py` files, one for Problem 1 and another one for Problem 2.
- For your report,
  - Must be converted to PDF.
  - Include basic elements such as a cover page, table of contents, etc.
  - Provide a problem definition, approach (visual aids like flowcharts are recommended), background information, input/output (conceptual explanation, distinct from output results), code analysis, results and discussion, and references. Explain the key concepts and the process followed to solve the assignment.
  - Ensure all graphs included in the results are attached. If there are any additional pieces of information or results you consider relevant, feel free to include them as well.
  - It is not recommended to copy and paste the entire code into the report.
  - Handwriting, Microsoft Word, or LaTeX formats are all acceptable for the report.
- A 10% deduction per day will be applied for late submissions.