INDIAN INSTITUTE OF TECHNOLOGY, ROPAR

CS503: Machine Learning

Lab 1: PAC Learnability, Hypothesis, Regression

SUJAL SABAVAT

2021MCB1249

TASK 1

SUBTASK1

The task involved analysing the empirical and true errors in a classification problem where data points are sampled from a uniform distribution and assigned binary labels. The hypothesis space consists of countable unions of non-overlapping partitions, and the goal is to evaluate the errors associated with this hypothesis space.

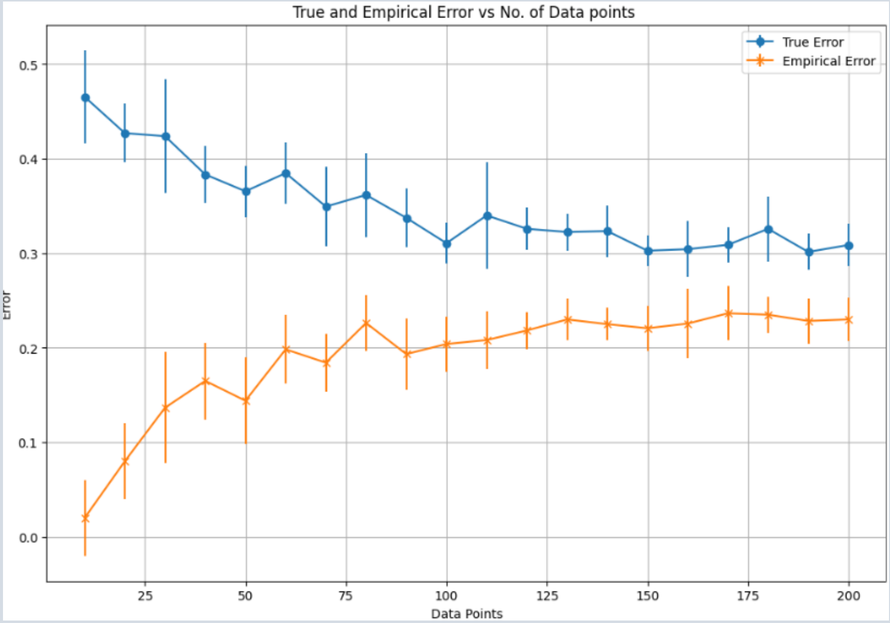


Figure: SUBTASK 1 Error plot

We implemented a procedure to compute the true error and empirical error for a given set of partitions. True error is calculated based on the provided probability distribution of labels for each data point, while empirical error is computed from the actual labelled data.

We generated synthetic data points with varying sample sizes (n) ranging from 10 to 200. Each data point consists of a feature value sampled uniformly from the range [0, 1] and a binary label determined probabilistically based on the provided distribution.

For each sample size, we performed 10 independent runs. In each run, we generated data points, computed the true error and empirical error, and recorded the results. We then calculated the mean and standard deviation of these errors across the 10 runs.

We plotted the mean empirical error and mean true error against the number of data points (n). Additionally, we included error bars representing the standard deviation to visualize the variability of errors.

True errors exhibit a decreasing trend as the number of data points increases. This observation suggests that larger datasets lead to more accurate classification. Where as empirical error is on a rise, tending to possible overfitting scenarios.

Empirical errors tend to be slightly higher than true errors across all sample sizes. This phenomenon indicates that the hypothesis space might not perfectly capture the underlying distribution of labels, leading to some discrepancies between predicted and actual labels.

SUBTASK2

As the number of partitions increase, true error and empirical error first converge due to underfitting, at k=3 they converge and as the partitions further increase the errors are diverging due to overfitting.

At k=3 (k*) the difference between true error and empirical error is minimal due to optimal fit.

At k=10 the empirical error is minimum.

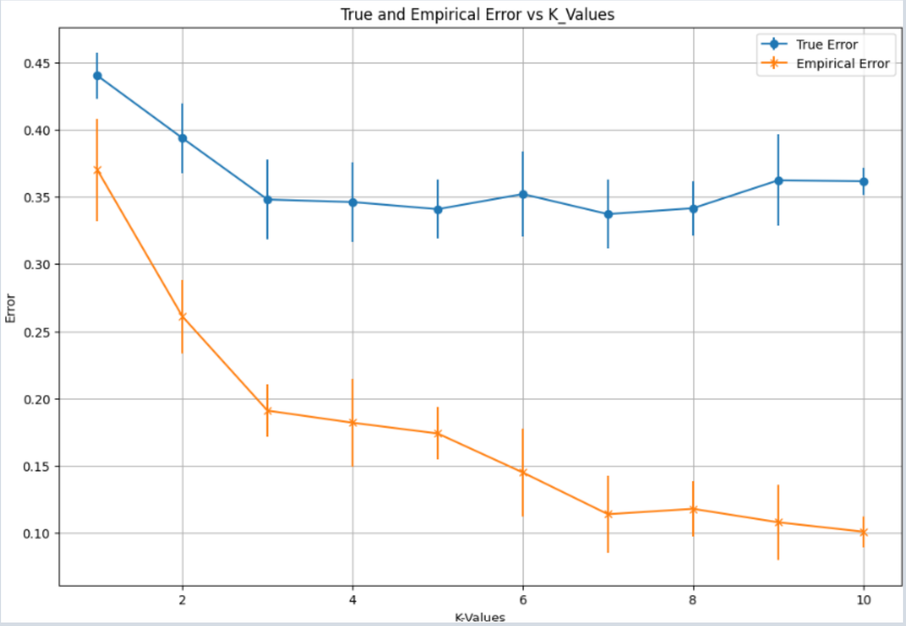


Figure 2: Error Plots for SUBTASK2

TASK 2

Gaussian Noise

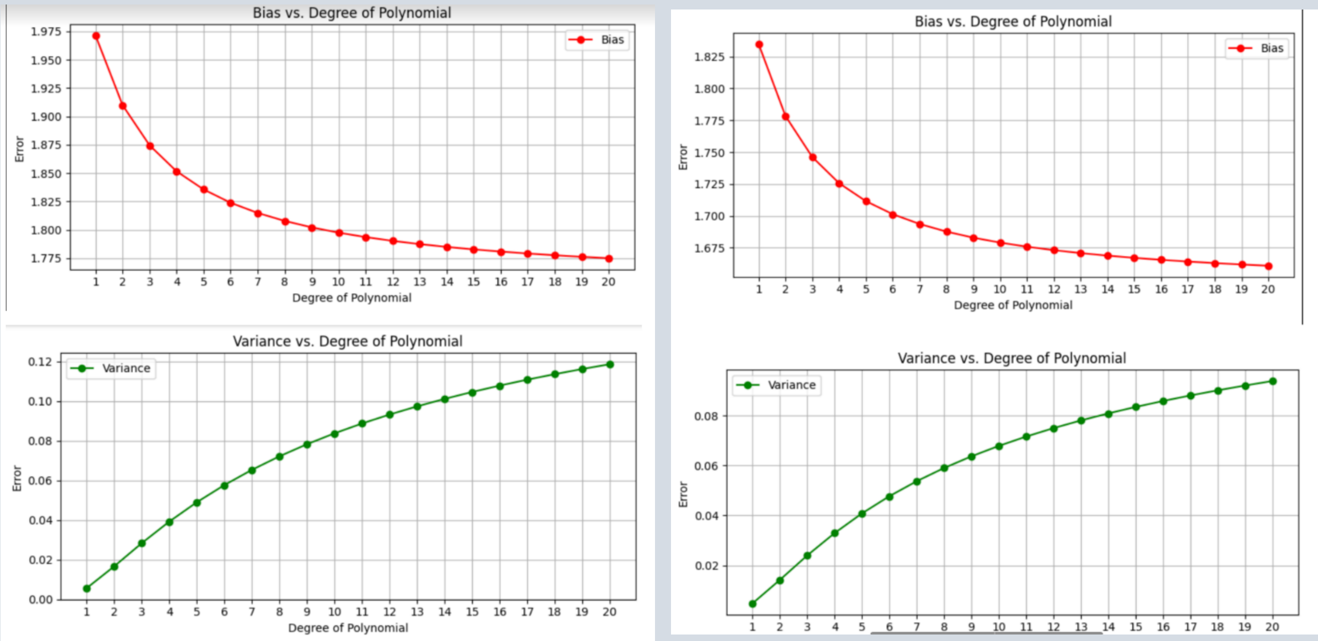


Figure 3.a and 3.b: Plots for Gaussian Noise

The bias and variance overlap for smaller degree and diverge for higher degree of polynomial regression.

Poisson Noise

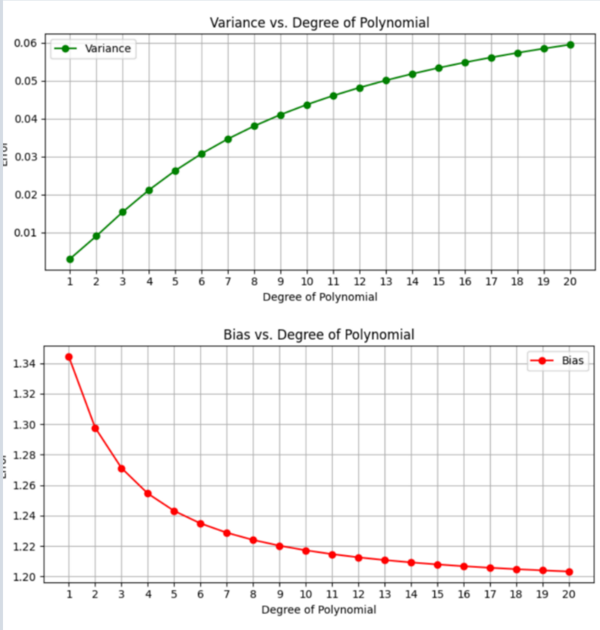
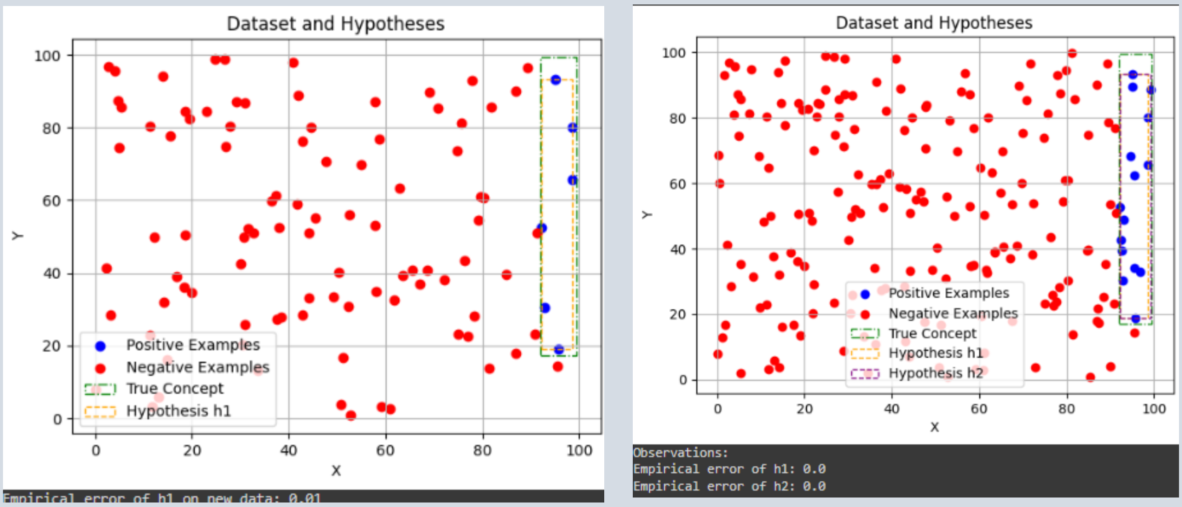


Figure 4: Plots for Poisson Noise

Bias tends to decrease with increasing model complexity (higher polynomial degrees) as the model can better fit the training data. Variance starts to increase when the model overfits.

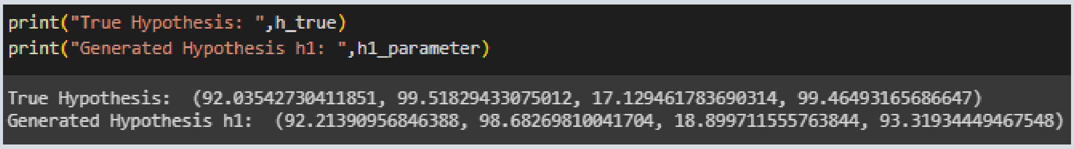
TASK 3



```
Empirical error of h1 on new data: 0.01
```

```
Observations:
Empirical error of h1: 0.0
Empirical error of h2: 0.0
```

Figure 5.a and 5.b: Plots representing the rectangles



```
print("True Hypothesis: ",h_true)
print("Generated Hypothesis h1: ",h1_parameter)
```

```
True Hypothesis:  (92.03542730411851, 99.51829433075012, 17.129461783690314, 99.46493165686647)
Generated Hypothesis h1:  (92.21390956846388, 98.68269810041704, 18.899711555763844, 93.31934449467548)
```

Figure 6: divergence of the hypothesis

In plot 5.a, the true hypothesis is represented by the green rectangle, while the yellow rectangle illustrates the hypothesis generated from the given data.

Moving to plot 5.b, the red rectangle signifies the true hypothesis, while both the blue rectangle and the green rectangle represent hypotheses generated from datasets with 50 and 200 data points respectively. Notably, the gap between the true hypothesis and the hypothesis generated with 200 data points (green rectangle) is notably smaller compared to the gap between the true hypothesis and the hypothesis generated with only 50 data points (blue rectangle). This observation underscores that as the number of data points increases, the generated hypothesis converges closer to the true hypothesis, indicating an improvement in the model's fidelity to the data distribution.