Introduction to Machine Learning

Homework 02

Atharva Patil- B01014288

LFD Exercise 2.6

A data set has 600 examples. To properly test the performance of the final hypothesis, you set aside a randomly selected subset of 200 examples which are never used in the training phase; these form a test set. You use a learning model with 1, 000 hypotheses and select the final hypothesis g based on the 400 training examples. We wish to estimate Eout(g). We have access to two estimates: Ein(g), the in sample error on the 400 training examples; and, Etest(g), the test error on the 200 test examples that were set aside.

(a) Using a 53 error tolerance (8 = 0.05), which estimate has the higher 'error bar' ?

Ans-

Problem 2.6

a). Given Hypothesis set of size $|H|$ a sample size N and a tolerance δ, the Hoeffding's inequality is given by:

$$P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2|H| e^{-2\epsilon^2 N}$$

In this case, we have two estimates
$E_{in}(g)$ from the 400 training
$E_{test}(g)$ from 200 test sample.

The error bar for each estimate will be depend on sample size (N) the no. of hypothesis $(|H|)$ and the tolerance (δ)

For $E_{in}(g)$, $N = 400$, $(|H|) = 1000$
For $E_{test}(g)$, $N = 200$, $(|H'|) = 1$

For $E_{in}(g)$: $P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 * 1000 * e^{-2\epsilon^2 * 400}$

For $E_{test}(g)$: $P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2 * 1 * e^{-2\epsilon^2 \cdot 200}$

As we see, error bar for $E_{in}(g)$ would be smaller than error bar for $E_{test}(g)$ because $E_{in}(g)$ is calculated with larger sample size ($N = 400$ vs $N = 200$)

(b) Is there any reason why you shouldn't reserve even more examples for testing?

<u>Ans</u>-

Yes, there are several reasons why we might not want to reserve more examples for testing:

<u>Reduced Training Data</u>: The more data we reserve for testing, the less data we have for training. This could potentially lead to a worse model if our model doesn't have enough data to learn effectively.

<u>Overfitting to Test Data</u>: If we use a large test set and use it frequently to evaluate the model, we might end up overfitting the test data. This is because we're indirectly using the test data to make decisions about the model, which can lead to an overly optimistic estimate of out-of-sample performance.

<u>Unnecessary Complexity</u>: If the model performs well on a smaller test set, reserving more data for testing might not provide much additional benefit. It could also make the evaluation process more complex and time-consuming.

<u>Data Distribution</u>: If the data is not evenly distributed, reserving more examples for testing might skew the distribution of classes in the training and test sets. This could make the model perform poorly on underrepresented classes.

In general, it's important to find a balance between the size of the training set and test set. A common practice is to split the data into 80% for training and 20% for testing.

<u>LFD Problem 2.12</u>

For an *H* with dvc = 10, what sample size do you need (as prescribed by the generalization bound) to have a 95% confidence that your generalization error is at most 0.05?

<u>Ans</u>-

Problem 2.12

We have for N,
$d_{vc} = 10, \quad \epsilon = 0.05$ and $\delta = 0.05$

$$N \geq \frac{8}{0.05^2} \ln\left[\frac{4[(2N)^{10}+1]}{0.05}\right]$$

To determine N, we will use an iterative process with an initial guess at $N = 1000$ in RHS, we get.

$$N \geq \frac{8}{0.05^2} \ln\left(\frac{4.[(2 \cdot 1000)^{10}+1]}{0.05}\right) \approx 2.572 \times 10^5$$

We then try the new value $N = 2.5725 \times 10^5$ in RHS & iterate this process rapidly converging to an estimate of $N \approx 4.529 \times 10^5$

LFD Problem 2.24

(a) Give the analytic expression for the average function g(x).

Ans-

Problem 2.24

(a)

Ans- The analytic expression for average function $\bar{g}(x)$ is given below:

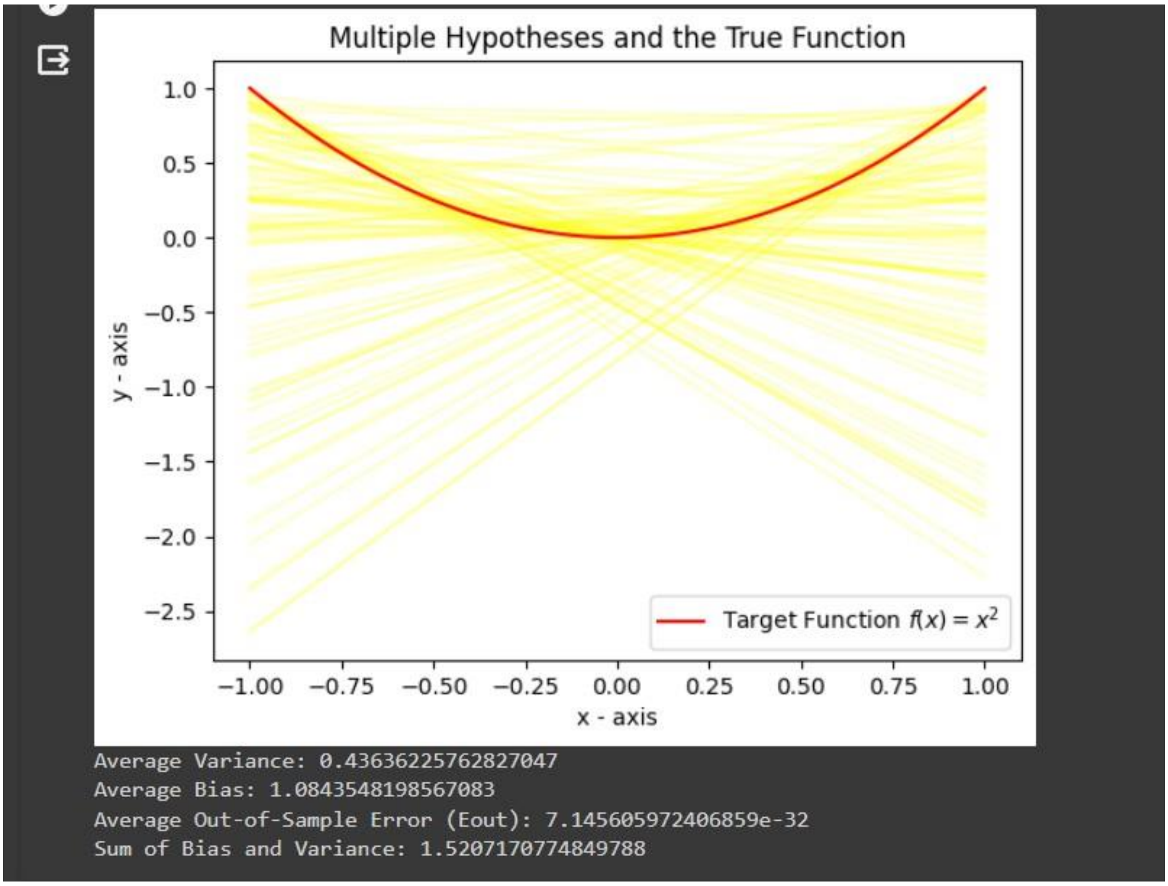
$$\begin{aligned}
\bar{g}(x) &= E_D[g(x)] \\
&= E_D\left[\frac{y_1 - y_2}{x_1 - x_2} x + \frac{x_1 y_2 - x_2 y_1}{x_1 - x_2}\right] \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} \frac{x_1^2 - x_2^2}{x_1 - x_2}\, dx_1\, dx_2 . x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} \frac{x_1 x_2^2 - x_2 x_1^2}{x_1 - x_2}\, dx_1\, dx_2 \\
&= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} (x_1 + x_2)\, dx_1\, dx_2 . x - \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} (x_1 x_2)\, dx_1\, dx_2 \\
&= \frac{1}{4} . 0 - \frac{1}{4} . 0 = 0
\end{aligned}$$

(b) Describe an experiment that you could run to determine (numerically) $g(x)$, E_{out}, bias, and var.

Ans- The experiment and the code is given in the google colab file and the result is attached in the next question answer section.

(c) Run your experiment and report the results. Compare *E*out with bias+var. Provide a plot of your g(x) and f(x) (on the same plot).

Ans-



(d) Compute analytically what *E*out. bias and var should be.

Ans-

Problem 2.24

(d)

Ans- To compute $E[E_{out}]$, we will first determine E_{out}.

$$E_{out} = E_x[(g(x) - f(x))^2] = E_x[(ax + b - x^2)^2]$$
$$= E_x[x^4] - 2aE_x[x^3] + (a^2 - 2b)E_x(x^2) + 2abE_x[x] + b^2$$
$$= \frac{1}{2}\int_{-1}^{1} x^4 dx - 2a\frac{1}{2}\int_{-1}^{1} x^3 dx + (a^2-2b)\frac{1}{2}\int_{-1}^{1} x^2 dx + 2ab\frac{1}{2}\int_{-1}^{1} x\, dx + b^2$$
$$= \frac{1}{5} + \frac{(a^2-2b)}{3} + b^2.$$

Then, we take expectation w.r.t. D to get test performance & we replace a and b by (x_1+x_2) & $(-x_1x_2)$ respectively, we get,

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[(x_1+x_2)^2 + 2x_1x_2] + E_D[x_1^2x_2^2]$$
$$= \frac{1}{5} + \frac{1}{3}\cdot\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2 + x_2^2 + 4x_1x_2)\,dx_1 dx_2 + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2\, dx_1 dx_2$$
$$= \frac{1}{5} + \frac{1}{3}\cdot\frac{1}{4}\cdot\frac{8}{3} + \frac{1}{4}\cdot\frac{4}{9} = \frac{8}{15}.$$

Next we compute bias.

$$\text{bias}(x) = (\bar{g}(x) - f(x))^2 = f(x)^2 = x^4;$$

then we get,

$$\text{bias} = E_x[x^4] = \frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{5}.$$

Finally, we compute variance.

$$\text{Var}(x) = E_D[(g(x) - \bar{g}(x))^2] = E_D[a^2x^2 + 2abx + b^2]$$
$$= E_D[a^2]\cdot x^2 + 2E_D[ab]\cdot x + E_D[b^2]$$
$$= E_D[(x_1+x_2)^2]\cdot x^2 - 2E_D[(x_1+x_2)x_1x_2]\cdot x + E_D[x_1^2x_2^2]$$
$$= E_D[x_1^2 + 2x_1x_2 + x_2^2]\cdot x^2 - 2E_D[x_1^2x_2 + x_1x_2^2]\cdot x + E_D[x_1^2x_2^2]$$
$$= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2 + 2x_1x_2 + x_2^2)\,dx_1dx_2\cdot x^2 - \frac{2}{4}\int_{-1}^{1}\int_{-1}^{1}(x_1^2x_2 + x_1x_2^2)\,dx_1dx_2\cdot x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2\,dx_1dx_2$$
$$= \frac{1}{4}\left(\frac{4}{3} + 0 + \frac{4}{3}\right)\cdot x^2 - 0\cdot x + \frac{1}{4}\cdot\frac{4}{9}$$
$$= \frac{2}{3}\cdot x^2 + \frac{1}{9}$$

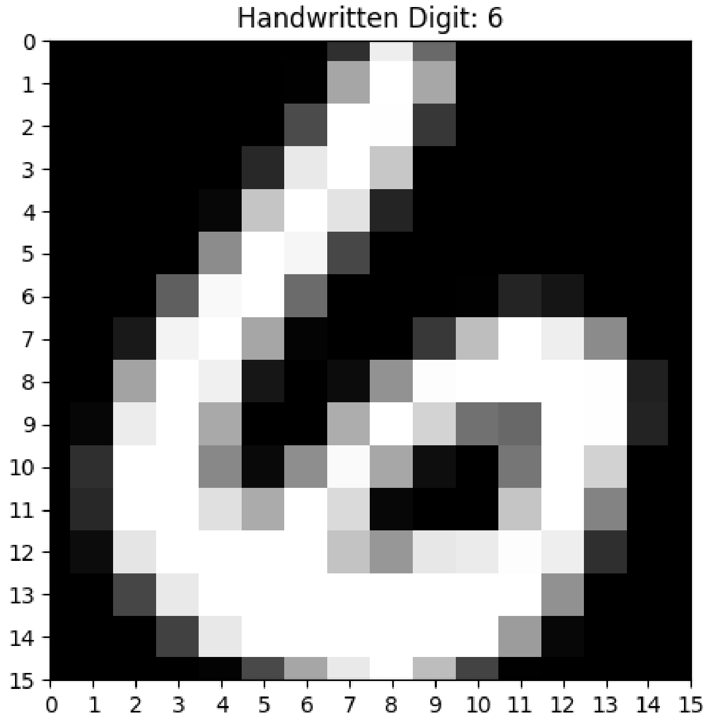
Then we get,

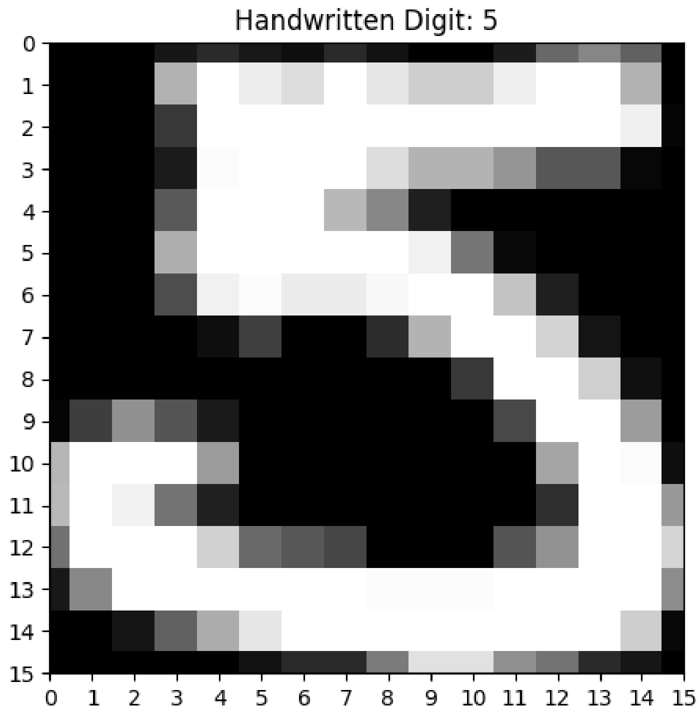
$$\text{Var} = E_x\left[\frac{2}{3}x^2 + \frac{1}{9}\right] = \frac{2}{3}\cdot\frac{1}{2}\int_{-1}^{1} x^2 dx + \frac{1}{9}$$
$$= \frac{1}{3}.$$

Q4.

(a) Familiarize yourself with the dataset by giving a plot of the first two digits in ZipDigits.train

Ans -





(b) Develop two features to measure properties of the image that would be useful in distinguishing between the digits 1 and 5. You may use average intensity and symmetry (defined in LFD Example 3.1) as your two features, or define and compute any other features you think are better suited to help distinguish between 1 and 5. Provide a mathematical definition of the two features you compute using the notation discussed in class.

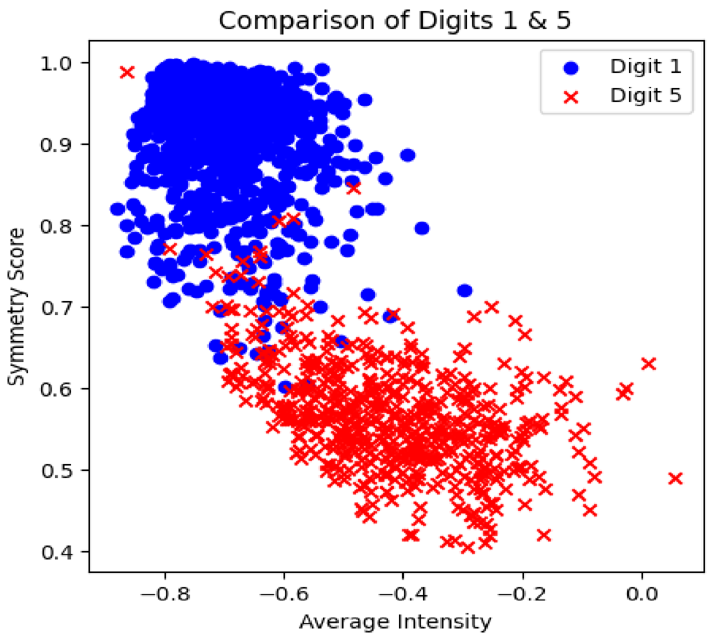
Ans-

```
[[ 1.         -0.75391406  0.93355109]
 [ 1.         -0.77228125  0.86450322]
 [ 1.         -0.76925781  0.97489716]
 ...
 [ 1.         -0.60979297  0.77710542]
 [ 1.         -0.66230859  0.91543961]
 [ 1.         -0.44755859  0.88223511]]
[7, 9, 12, 13, 19, 29, 34, 44, 70, 82, 92, 112, 124, 126, 129, 137, 144, 175, 179, 184, 197, 206, 221, 233, 234, 236, 245, 246, 250, 251, 253, 256, 258, 259, 273, 2
Number of 1s: 1005
[1, 39, 119, 143, 147, 166, 211, 261, 262, 265, 276, 299, 326, 348, 351, 359, 360, 376, 377, 391, 407, 415, 420, 441, 444, 462, 490, 562, 605, 646, 675, 718, 722, 7
Number of 5s: 556
Total number of 1s and 5s: 1561
```

```
 [ 1.         -0.76014453  0.99466305]
 [ 1.         -0.63766797  0.8521528 ]
 [ 1.         -0.6675625   0.9741973 ]
 [ 1.         -0.73596094  0.81295794]
 [ 1.         -0.80748047  0.9513939 ]
 [ 1.         -0.72017578  0.9751024 ]
 [ 1.         -0.77021484  0.62981569]
 [ 1.         -0.61441797  0.87896067]
 [ 1.         -0.64599609  0.82728818]
 [ 1.         -0.62204687  0.86770629]]
[16, 42, 45, 52, 66, 69, 77, 79, 128, 133, 156, 158, 180, 188, 190, 192, 201, 206, 210, 226, 233, 238, 243, 248, 250, 273, 288, 314, 316, 317, 318, 328, 341, 343, 3
Number of 1s: 264
[32, 36, 51, 56, 116, 121, 126, 131, 147, 195, 204, 239, 241, 242, 246, 247, 306, 311, 323, 332, 351, 437, 448, 461, 481, 489, 518, 556, 561, 599, 601, 604, 611, 61
Number of 5s: 160
Total number of 1s and 5s: 424
FeaturesTest shape: (424, 3), LabelsTest shape: (424, 1)
```

(c) Provide a 2-D scatter plot of the examples in ZipDigits.train w.r.t. the two features you defined in Part (b), similarly to the scatter plot in LFD Example 3.1 and elsewhere in LFD Chapter 3. For each example, plot the values of the two features with a red '×' marker if it is a 5 and and a blue '◦' marker if it is a 1. You must clearly label each axis with the feature it represents, and it should be possible to determine for each data point, the values of the two features you computed. You must also include a legend on the upper right corner of your scatter plot which clearly identifies that data points marked with '×' represent examples of the digit 5 those marked '◦' marker represent examples of the digit 1.

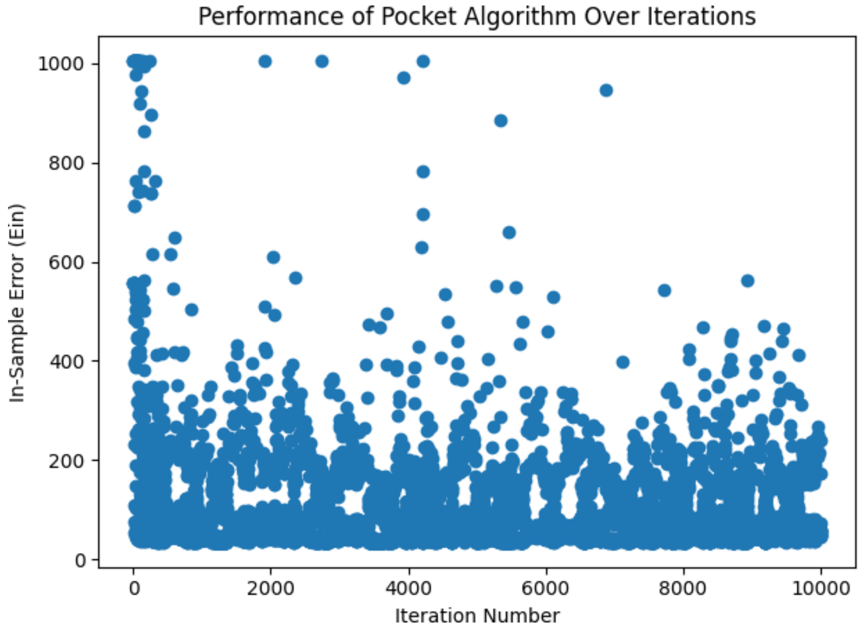
Ans-



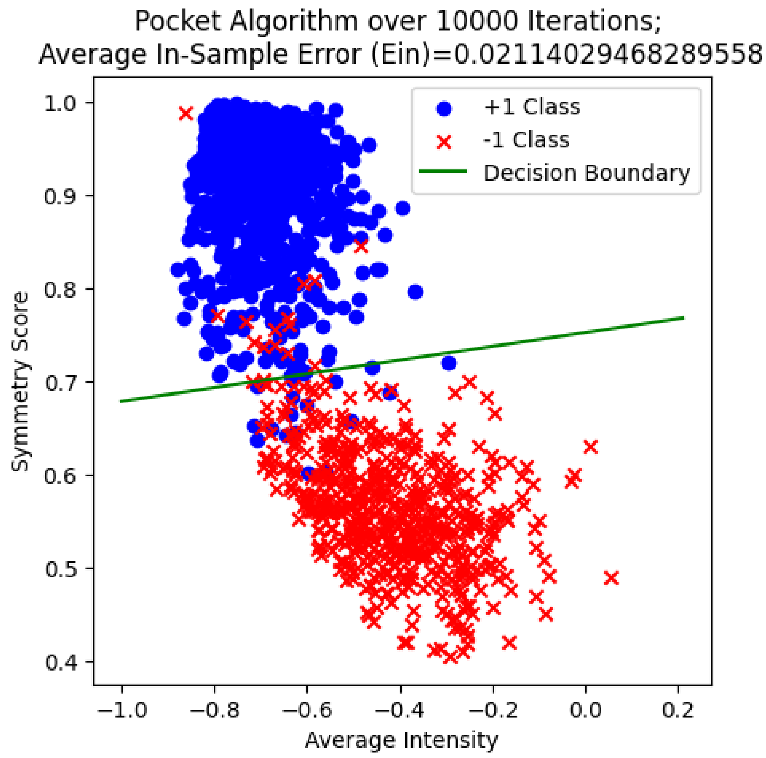
Q5.

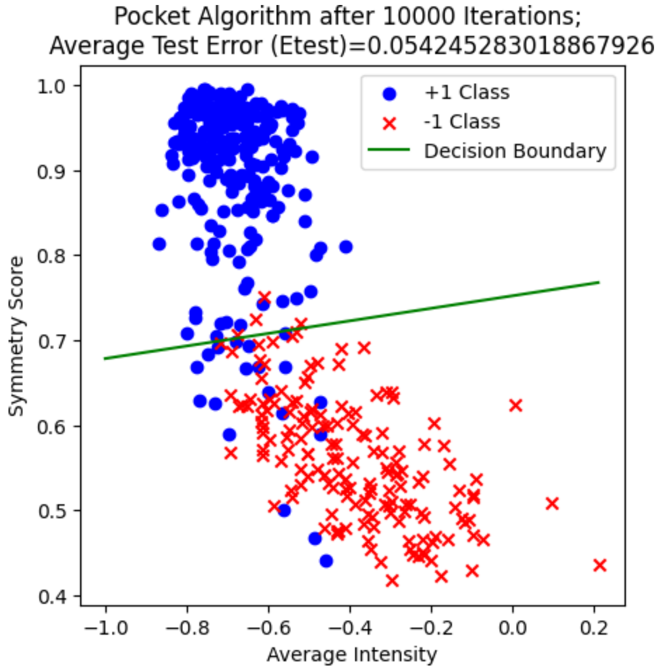
(a) (b) (c)

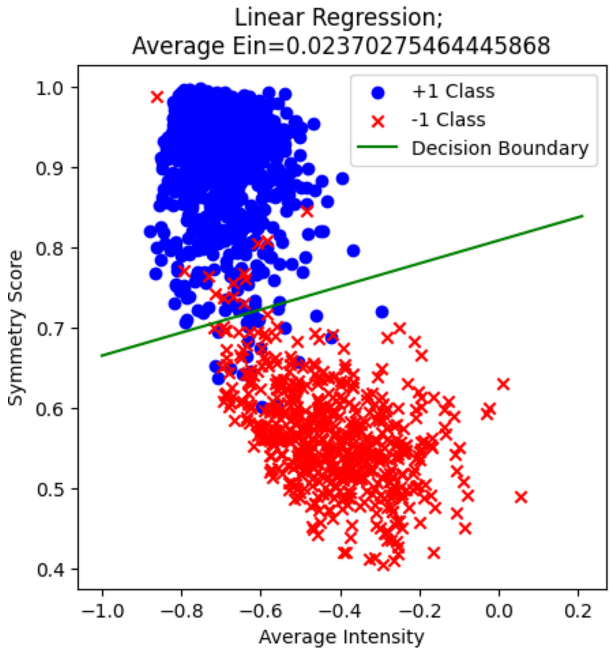
<u>Ans</u>-



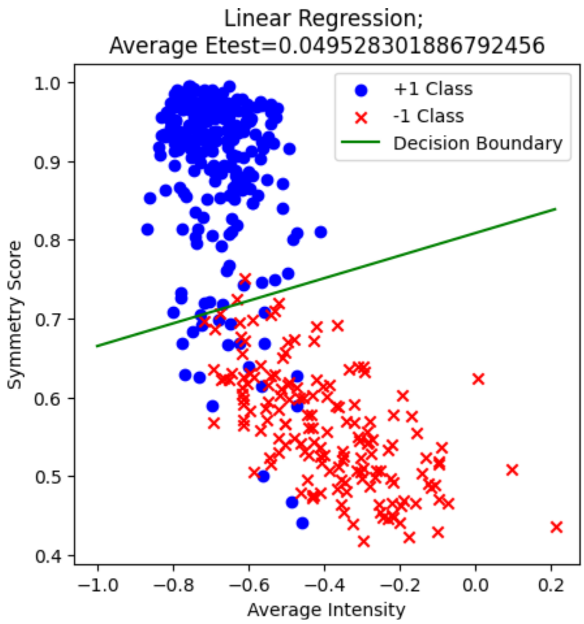

```
Lowest In-Sample Error (Ein_best): 33.0
Optimal Weights (w_best):
[[-11.99      ]
 [ -1.17691406]
 [ 15.94093084]]
```



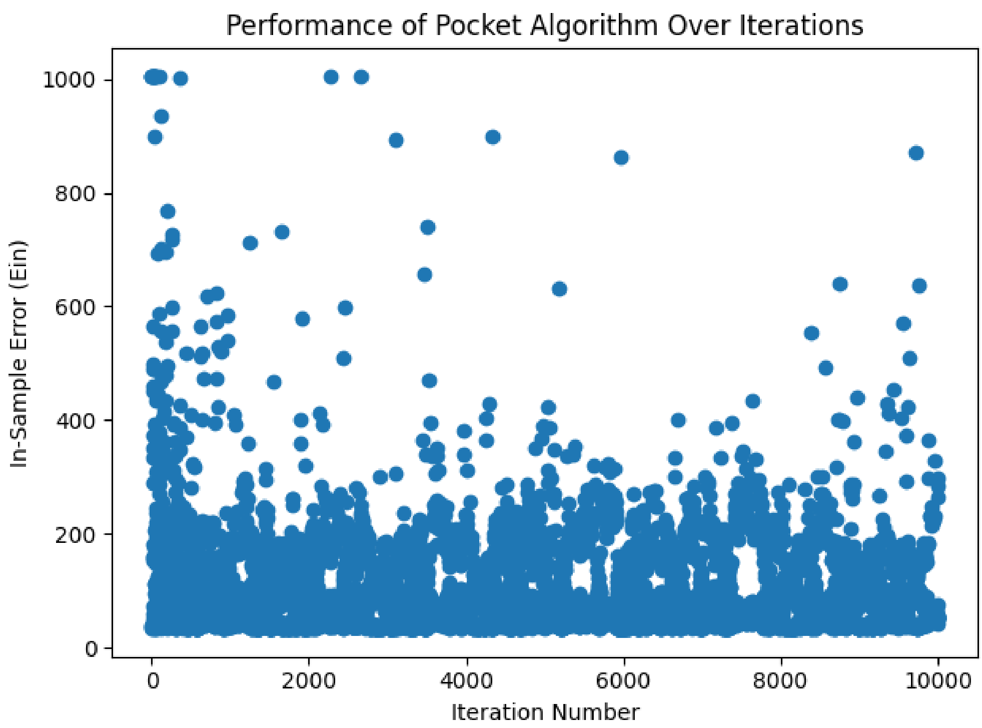




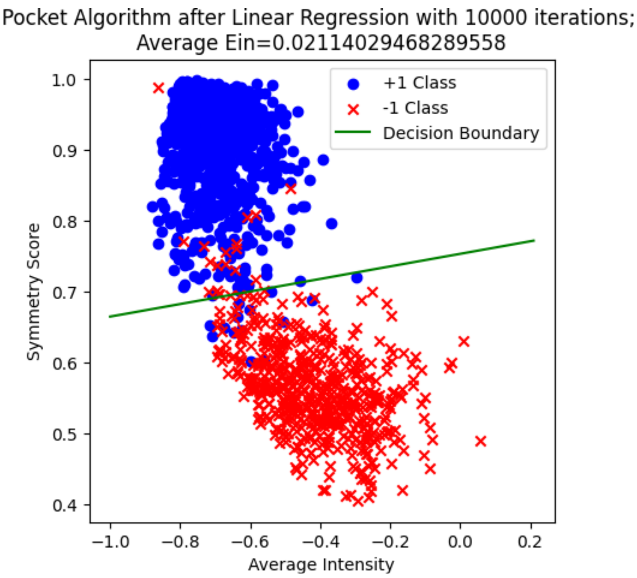

```
Estimated Eout based on Ein:
0.05807681082095556
```



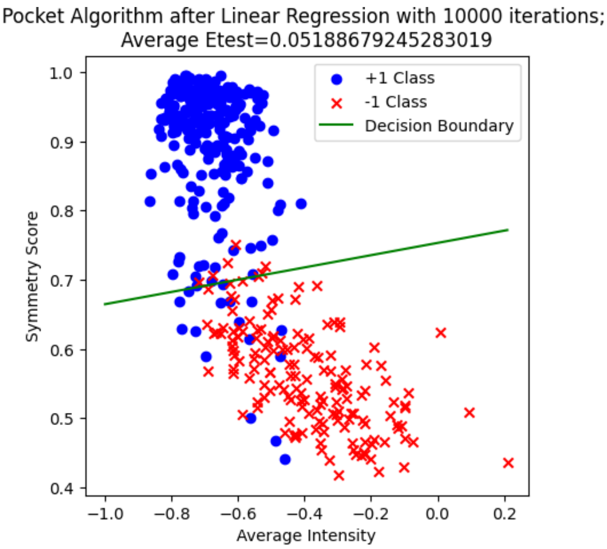

```
Estimated Eout based on Etest:
0.11548354198103422
```




```
Lowest In-Sample Error (Ein_best): 33.0
Optimal Weights (w_best):
[[-11.57805185]
 [ -1.3603941 ]
 [ 15.37312342]]
```




```
Estimated Eout based on Ein after Pocket:
0.055514350859392464
```




```
Estimated Eout based on Etest after Pocket:
0.11784203254707196
```

Google Colab:- https://colab.research.google.com/drive/1xP1n1JFRHjZoAKG-CFl95c2DSC7QdMrh#scrollTo=3piLY9st4fh1

Collaborator – Shreyas Kadam