CS436/536: Introduction to Machine Learning
Homework 2

(2) LFD Exercise 2.6

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 $E_{\text{out}}(g)$. We have access to two estimates: $E_{\text{in}}(g)$, the in sample error on the 400 training examples; and, $E_{\text{test}}(g)$, the test error on the 200 test examples that were set aside.

(a) Using a 5% error tolerance ($\delta = 0.05$), which estimate has the higher 'error bar'?

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

Q2] Exercise 2.6

a) Given :- $N = 600$ examples (Sample data set)

$N_1 = 400$ (training data)

$N_2 = 200$ (test data)

$\mathcal{H} = 1000$

$\delta = 0.05$ (5% error)

To find :- Error bar for $E_{in}(g)$

Error bar for $E_{out}(g)$ → for test sample.

Formula :- $$E_{out} \leq E_{in} + \sqrt{\frac{1}{2N} \ln \frac{2\mathcal{H}}{\delta}}$$

Solution :-

Set Error bar$(\mathcal{E})$ = $$\sqrt{\frac{1}{2N} \ln \frac{2\mathcal{H}}{\delta}} \quad \text{—} \ (1)$$

lets use this Eq^n to find Error bar for Input samples.

$$\mathcal{E}_{(in)} = \sqrt{\frac{1}{2N_1} \ln \frac{2\mathcal{H}}{\delta}}$$

$$= \sqrt{\frac{1}{2 \times 400} \ln \frac{2 \times 1000}{0.05}}$$

$$= \sqrt{0.00125 \times \ln 40000}$$

$$= \sqrt{0.00125 \times 10.5966}$$

$$= \sqrt{0.01324}$$

$$\boxed{E_{in} \text{ (Error bar)} = 0.1150}$$

lets use the same eq^n to now solve for test sample.

$$\mathcal{E}_{test} \text{ (Error bar)} = \sqrt{\frac{1}{2N_2} \ln \frac{2M}{\delta}}$$

$$= \sqrt{\frac{1}{2\times 200} \ln \frac{2 \times 1}{0.05}} \quad (\because M = 1 \text{ for testing})$$

$$= \sqrt{\frac{1}{400} \ln 40}$$

$$= \sqrt{0.0025 \times 3.6888}$$

$$= \sqrt{0.009222}$$

$$\boxed{\mathcal{E}_{test} \text{ (Error bar)} = 0.0960}$$

From the calculated value we got.

$\mathcal{E}_{test} = 0.0960$ and $\mathcal{E}_{in}$ $\cancel{as}$ $= 0.1150$.

this shows that the in sample Error is greater than out sample error.

b) From the above output results we got the relation between the values of $\mathcal{E}_{in}$ and $\mathcal{E}_{test}$. If we change the sample size distribution i.e instead of 400:200, if we give more sample for testing then the sample data for training would get reduced. It would become 300:300. This may cause issue while choosing the correct hypothesis and we might end up with a hypothesis which is not as good as it could have been if we would have given more sample data for training.

This would result in higher out of sample error, as we have given less data to train our model.

That's why we shouldn't reserve even more examples for testing. And choose optimum distribution proportion.

(3) LFD Problem 2.12

Problem 2.12 For an $\mathcal{H}$ with $d_{\text{vc}} = 10$, 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?

3) Problem 2.12

Given :- $d_{vc} = 10$

Confidence $= 95\%$ so $(1-\delta) = 95\%$ so $\delta = 0.05$

$\varepsilon = 0.05$

Formula :- $$N \geq \frac{8}{\varepsilon^2} \ln\left(\frac{4(2N)^{d_{vc}} + 1}{\delta}\right)$$

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

Trying an initial guess of $N = 2000$ in RHS

$$N \geq \frac{8}{(0.05)^2} \ln\left(\frac{4(2 \times 2000)^{10} + 1}{0.05}\right)$$

$$N \geq \frac{8}{(0.05)^2} \times 87.322$$

$$N \geq \frac{698.5801}{0.0025}$$

$$N \approx 279432$$

~~So~~ Using this value of N now in RHS.

$$N \geq \frac{8}{(0.05)^2} \ln\left(\frac{4(2 \times 279432)^{10} + 1}{0.05}\right)$$

$$N \geq \frac{8}{(0.05)^2} \times 136.7186$$

$$\boxed{N \approx 437499}$$

When we ~~try the~~ tried the New value of N which was 279432 in the RHS, the value is rapidly converging to an estimate of $N \approx 437499$.

(4) LFD Problem 2.24

Problem 2.24 Consider a simplified learning scenario. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval $[-1, 1]$. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$. Thus, the full data set is $\mathcal{D} = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm returns the line fitting these two points as g ($\mathcal{H}$ consists of functions of the form $h(x) = ax + b$). We are interested in the test performance (E_{out}) of our learning system with respect to the squared error measure, the bias and the var.

(a) Give the analytic expression for the average function $\bar{g}(x)$.

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

(c) Run your experiment and report the results. Compare E_{out} with bias+var. Provide a plot of your $\bar{g}(x)$ and $f(x)$ (on the same plot).

(d) Compute analytically what E_{out}, bias and var should be.

Problem 2.24

a) Given :- $f(x) = x^2$

$D = \{(x_1, x_1^2), (x_2, x_2^2)\}$

data interval $[-1, 1]$.

Formula :- $E_{in}(g) = \frac{1}{N}\sum_{i=1}^{2}\left[f(x_i) - h(x_i)\right]^2$

To find :- $\bar{g}(x) = ?$, ~~var~~ variance $= ?$, bias $= ?$

$\mathbb{E}\left[E_{out}(g)\right] = ?$

Solution :-

$$E_{in}(g) = \frac{1}{2}\sum_{i=1}^{2}\left[f(x_i) - h(x_i)\right]^2$$

$$= \frac{1}{2}\sum_{i=1}^{2}\left[x_i^2 - (ax_i + b)\right]^2$$

Taking partial derivative w.r. a

$$\frac{\partial E_{in}(g)}{\partial a} = \frac{1}{\cancel{2}}\sum_{i=1}^{2} -\cancel{2}x_i\left[x_i^2 - ax_i - b\right]$$

$$\boxed{\frac{\partial E_{in}(g)}{\partial a} = -\sum_{i=1}^{2} x_i\left[x_i^2 - ax_i - b\right]} \quad (1)$$

Similarly lets take partial derivative w.r.b

$$\frac{\partial E_{in}(g)}{\partial b} = \frac{1}{2}\sum_{i=1}^{2}[(x_i^2 - ax_i - b)]^2$$

$$= \frac{1}{\cancel{2}}\sum_{i=1}^{2} -\cancel{2}[x_i^2 - ax_i - b]$$

$$\boxed{\frac{\partial E_{in}(g)}{\partial b} = -1\sum_{i=1}^{2}[x_i^2 - ax_i - b]} \quad — ②$$

from eq^n ①

$$\frac{\partial E_{in}(g)}{\partial b} = -1\sum_{i=1}^{2}(x_i^3 - ax_i^2 - bx_i)$$

let $\frac{\partial E_{in}(g)}{\partial b}$ and $\frac{\partial E_{in}(g)}{\partial a} = 0$.

$$0 = -\sum_{i=1}^{2}(x_i^3 - ax_i^2 - bx_i)$$

$$= -(x_1^3 - ax_1^2 - bx_1 + x_2^3 - ax_2 - bx_2) = 0$$

$$\boxed{x_1^3 - ax_1^2 - bx_1 + x_2^3 - ax_2^2 - bx_2 = 0} \quad — Eq^n ③$$

Expanding eq ②.

$$0 = -(x_1^2 - ax_1 - b + x_2^2 - ax_2 - b)$$

$$x_1^2 - ax_1 - b + x_2^2 - ax_2 - b = 0 \quad \text{—— ④}$$

lets multiply this eq^n by x_1 we get.

$$x_1^3 - ax_1^2 - bx_1 + x_2^2 x_1 - ax_1x_2 - bx_1 = 0$$

lets subtract this eq^n from eq ③

$$\cancel{x_1^3} - \cancel{ax_1^2} - \cancel{bx_1} + x_2^3 - ax_2^2 - bx_2 = 0$$

$$\cancel{x_1^3} - \cancel{ax_1^2} - \cancel{bx_1} + x_2^2x_1 - ax_1x_2 - bx_1 = 0$$

$$- \quad + \quad + \quad - \quad + \quad +$$

$$x_2^3 - x_2^2x_1 - ax_2^2 + ax_1x_2 - bx_2 + bx_1$$

$$x_2^2(x_2 - x_1) - ax_2(x_2 - x_1) + b(x_1 - x_2) = 0$$

$$x_2^2(x_2 - x_1) + ax_2(x_1 - x_2) + b(x_1 - x_2) = 0$$

$$-x_2^2(x_1 - x_2) + ax_2(x_1 - x_2) + b(x_1 - x_2) = 0$$

taking $(x_1 - x_2)$ common we get.

$$(x_1 - x_2)\left[-x_2^2 + ax_2 + b\right] = 0$$

$$(x_1 - x_2)(-x_2^2 + ax_2 + b) = 0$$

We get these two Eq^n after solving it.

$$x_1 = x_2 \quad \&$$

$$x_2^2 - ax_2 - b = 0 \quad \text{—— (5)}$$

Now lets multiply Eq (4) by x_2.

$$x_1^2 x_2 - ax_1x_2 - bx_2 + x_2^3 - ax_2^2 - bx_2 = 0$$

Subtract this eq^n by (3)

$$x_1^3 - ax_1^2 - bx_1 + \cancel{x_2^3} - \cancel{ax_2^2} - \cancel{bx_2} = 0$$

$$\underset{-}{x_1^2x_2} \underset{+}{-} ax_1x_2 \underset{+}{-} bx_2 \underset{-}{+} \cancel{x_2^3} \underset{+}{-} \cancel{ax_2^2} \underset{+}{-} \cancel{bx_2} = 0$$

$$x_1^2(x_1 - x_2) + ax_1(-x_1 + x_2) + b(-x_1 + x_2) = 0$$

$$x_1^2(x_1 - x_2) + ax_1(x_2 - x_1) + b(x_2 - x_1) = 0$$

$$x_1^2(x_1 - x_2) - ax_1(x_1 - x_2) - b(x_1 - x_2) = 0$$

taking $(x_1 - x_2)$ as common.

$$(x_1 - x_2)(x_1^2 - ax_1 - b) = 0$$

$$x_1^2 - ax_1 - b = 0$$

$$\boxed{x_1^2 - ax_1 - b = 0} \quad \text{—— (6)}$$

$$x_1^2 - ax_1 = b$$

lets substitute $b = x_1^2 - ax_1$ is eq^n (5)

$$x_2^2 - a x_2 - x_1^2 + a x_1 = 0$$

$$x_2^2 - x_1^2 + a(-x_2 + x_1) = 0$$

$$x_2^2 - x_1^2 - a(x_2 - x_1) = 0$$

$$\cancel{(x_2 - x_1)}(x_2 + x_1) = a\cancel{(x_2 - x_1)}$$

$$\boxed{a = x_2 + x_1}$$

lets put $a = x_2 + x_1$ in eq 5

$$x_2^2 - (x_2 + x_1)x_2 - b = 0$$

$$\cancel{x_2^2} - \cancel{x_2^2} - x_1 x_2 - b = 0$$

$$\boxed{b = -x_1 x_2}$$

lets put a & b in our hypothesis eq^n.

$$h(x) = a x + b$$

$$= (x_2 + x_1) x + (-x_1 x_2)$$

This is our $g^D(x)$

$$\boxed{g^D(x) = (x_2 + x_1) x - x_1 x_2} \quad (7)$$

We know that :

$$\mathbb{E}_D[g^D(x)] = \bar{g}(x).$$

$$\therefore \bar{g}(x) = \mathbb{E}_D[(x_2+x_1)x - x_1x_2]$$

$$= \mathbb{E}_D[x_2x + x_1x - x_1x_2]$$

$$\boxed{\bar{g}(x) = \mathbb{E}_D[x_2]x + \mathbb{E}_D(x_1)x - \mathbb{E}_D(x_1)\mathbb{E}_D(x_2)}$$

This is our analytic expression for the average function $\bar{g}(x)$.

lets take mean of x_1 & x_2.

As x_1 & x_2 are uniformly distributed over the range of $[-1, 1]$

The mean of variables x_1 & x_2 will be 0.

$\therefore x_1 = 0$ & $x_2 = 0$

Substituting in the above eq^n we get.

$$\bar{g}(x) = 0.x + 0.x - 0.0$$

$$\boxed{\bar{g}(x) = 0} \quad \text{--- (8)}$$

We know

$$\text{Variance} = \mathbb{E}_D\left[\left(g^D(x) - \bar{g}(x)\right)^2\right]$$

from eq^n 7 & 8 we get.

$$= E_D\left[\left(g^D(x) - 0\right)^2\right]$$

$$= E_D\left[\left((x_1+x_2)x - x_1x_2\right)^2\right]$$

$$\left[(a-b)^2 = a^2 - 2ab + b^2\right]$$

$$= E_D\left[(x_1+x_2)^2x^2 - 2x(x_1+x_2)(x_1x_2) + (x_1x_2)^2\right]$$

$$= E_D(x_1+x_2)^2x^2 - E_D 2x(x_1+x_2)(x_1x_2) + E_D x_1^2x_2^2$$

$$\left[\because \text{Variance} = \frac{(b-a)^2}{12} = \frac{(1-(-1))^2}{12}\right.$$

where $b \Rightarrow$ upper bound

$a \Rightarrow$ lower bound

$12 \Rightarrow$ is used for uniform distribution)

$$\left[\therefore \text{Variance} = \frac{2^2}{12} = \frac{4}{12} = \frac{1}{3}\right]$$

This variance value is for x_1^2 & x_2^2.

$$\text{Variance} = E_D\,(x_1+x_2)^2 x^2 - 2x_1x_2(x_1+x_2)x + x_1^2 x_2^2$$

[Substituting the values of mean & variance i.e $x^2 = x_1^2 = x_2^2 = \frac{1}{3}$ and $x_1 = x_2 = 0$]

$$= E_D\,(x_1^2 + x_2^2 + 2x_1x_2)x^2 - 2x_1x_2(x_1+x_2)x + x_1^2 x_2^2$$

$$= \left(\frac{1}{3} + \frac{1}{3} + 2\times 0\times 0\right)\frac{1}{3} - 2\times 0\times 0\,(0+0)\,0 + \frac{1}{3}\times\frac{1}{3}$$

$$= \left(\frac{2}{3}\right)\frac{1}{3} - 0 + \frac{1}{9}$$

$$= \frac{2}{9} + \frac{1}{9}$$

$$= \frac{3}{9} = \frac{1}{3}$$

$$\boxed{\text{Variance}\,(x) = \frac{1}{3}} \quad \text{———} \quad (9)$$

Similarly we know that:

$$\text{bias}(x) = \mathbb{E}_x\left(\bar{g}(x) - f(x)\right)^2$$

$(\because \bar{g}(x) = 0$ from eq^n 8$)$

$$= \mathbb{E}_x\left(0 - x^2\right)^2$$

$$\text{bias}(x) = \mathbb{E}_x(x^4)$$

[We know that

$$\mathbb{E}_x(x^4) = \int_{-1}^{1} x^4 \underbrace{f(x)}_{\downarrow} \, dx.$$

probablity density function. having const value as $\frac{1}{2}$ for range [-1, 1]

$$= \int_{-1}^{1} x^4 \times \frac{1}{2} \, dx$$

taking Integration we get.

$$= \frac{1}{2}\left[\frac{x^5}{5}\right]_{-1}^{1}$$

$$= \frac{1}{2}\left(\frac{1}{5} - \frac{(-1)^5}{5}\right)$$

$$= \frac{1}{\cancel{2}}\left[\frac{\cancel{2}}{5}\right] = \frac{1}{5}$$

$$\because E_x(x^4) = \frac{1}{5}$$

$$\boxed{\text{bias}(x) = \frac{1}{5}} \text{ ——— } (10)$$

$E_{out}(g) = \text{bias} + \text{Variance}$

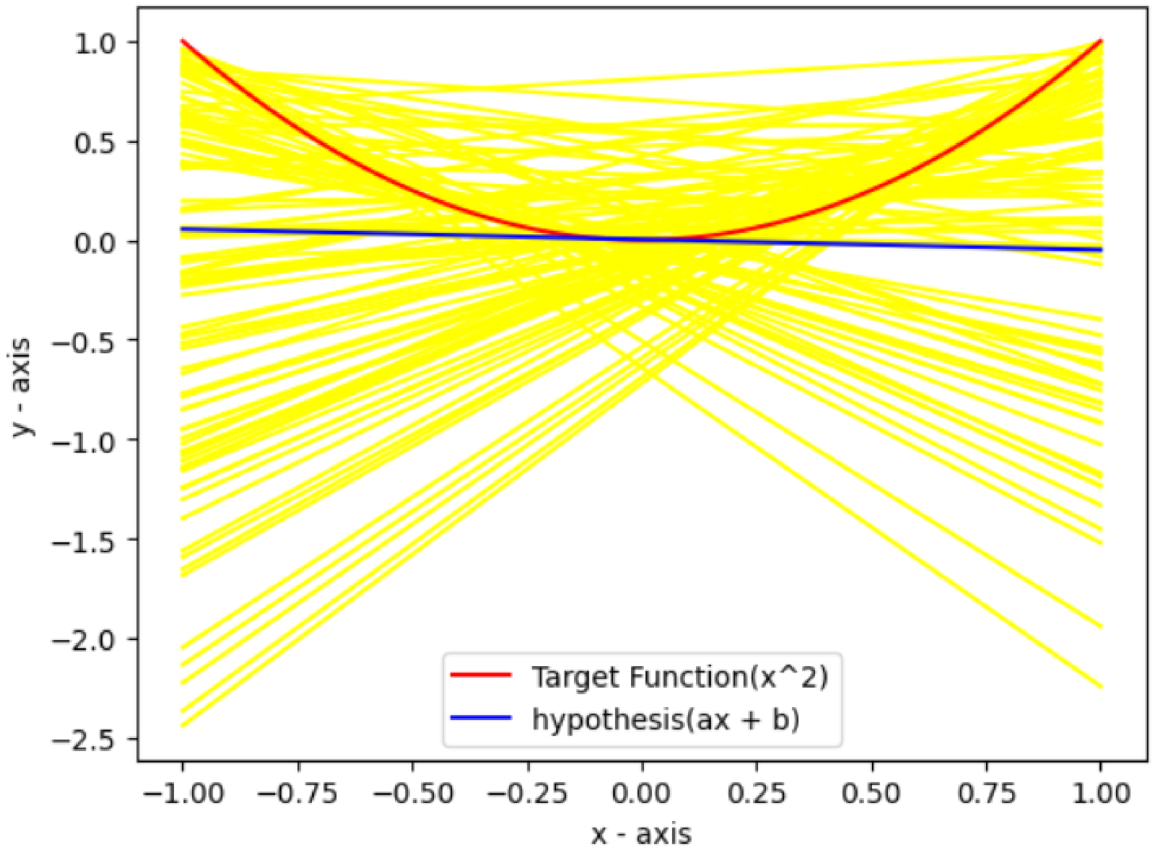
from eq^n (9) & (10)

$$= \frac{1}{5} + \frac{1}{3}$$

$$= \frac{3+5}{15}$$

$$\boxed{E_{out}(g) = \frac{8}{15}}$$

2.24 (c)




```
The Variance is : 0.2088145102341011
The Bias is     : 0.5062529745866432
The Eout is     :  0.5062529745866429
The Bias + Variance is : 0.7150674848207443
```

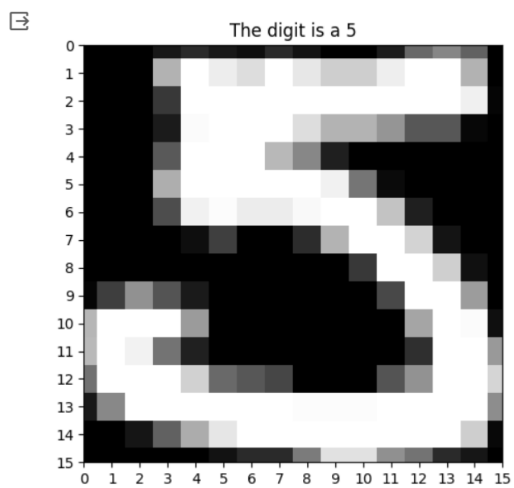
(5) Computing Features with the Handwritten Digits Data.

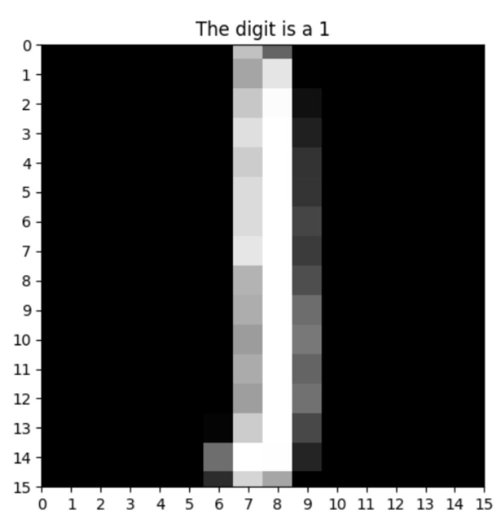
This week, you will familiarize yourself with the MNIST handwritten digits dataset. You may download the following files from Brightspace that accompany this homework: ZipDigits.train and ZipDigits.test contain training and test datasets respectively. The content and format of these files is documented in ZipDigits.info.

Each row of ZipDigits.train (or ZipDigits.test) corresponds to an example of a handwritten digit. Entries in a row are space separated. The first entry in a row correctly identifies the digit (0-9), and the next 256 entries are grayscale values between -1 (white) and 1 (black) corresponding to a 16×16 image read row after row.

For this problem, we will use only the digits 1 and 5, so you must first remove the other digits from the training and test datasets. Then, perform the following tasks:

A. Familiarize yourself with the dataset by giving a plot of the first two digits in ZipDigits.train. Hint: If you are using the Python programming language, you may use matplotlib.pyplot.imshow which takes a 2-D array as input. You may refer to the documentation on how to display a grayscale image.



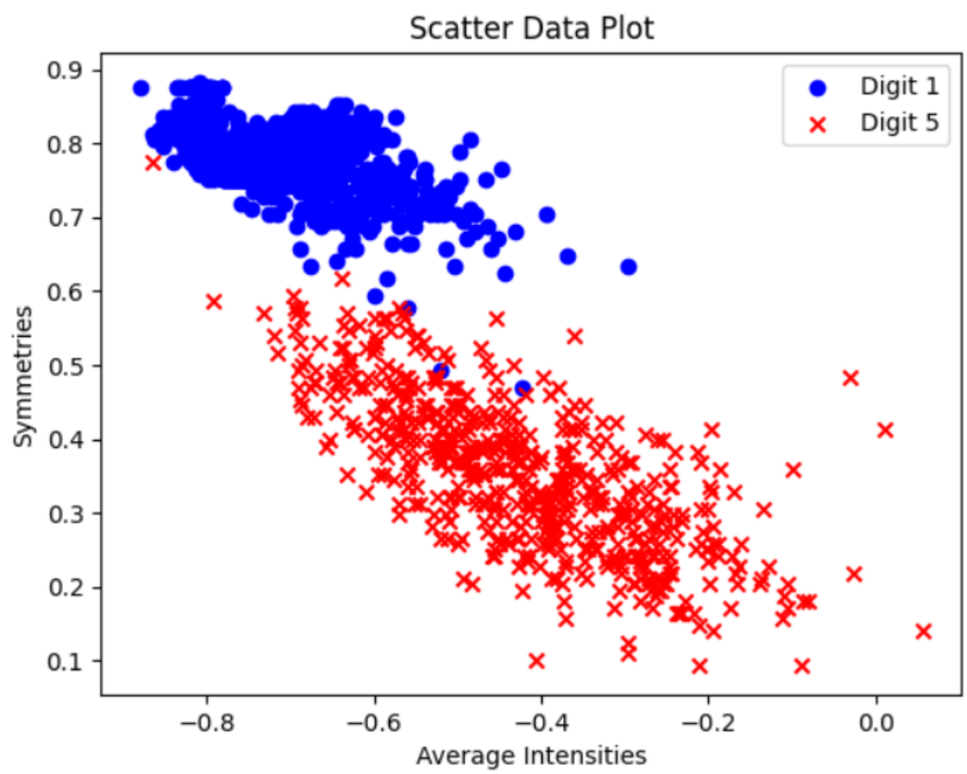


B. Develop two features to measure properties of the image that would be useful in distinguishing between the digits 1 and 5. You may use average intensity and symmetry (defined in LFD Example 3.1) as your two features, or define and compute any other features you think are better suited to help distinguish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

Solved inside google colab

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.

```
<matplotlib.legend.Legend at 0x7c20366ec9d0>
```



(6) Classifying Handwritten Digits: 1 vs. 5

Pick one of the following two classification algorithms for non-separable data:

(i) Linear regression for classification followed by a pocket algorithm.

(ii) Logistic regression for classification using gradient descent. Use your classification algorithm of choice to find the best separator using the training data only (using ZipDigits.train). For each example, use the two features you computed in HW2 as the inputs; and the output is +1 if the handwritten digit is 1 and −1 if the handwritten digit is 5. Once you have found a separator using your classification algorithm:

A. Give separate plots of the training data (ZipDigits.train) and test data (ZipDigits.test) which display the data points using the two features you computed in HW2, together with the separator
B. Compute Ein on your training data (ZipDigits.train) and Etest, the error of your separator on the test data (ZipDigits.test)
C. Obtain a bound on the true out-of-sample error (Eout). You should get two bounds, one based on Ein and another based on Etest. Use a tolerance of $\delta = 0.05$. Which is the better bound?
D. Repeat parts (a)-(c) using a 3-rd order polynomial transform
E. Which model would you deliver to the USPS, the linear model with the 3-rd order polynomial transform or the one without? Explain

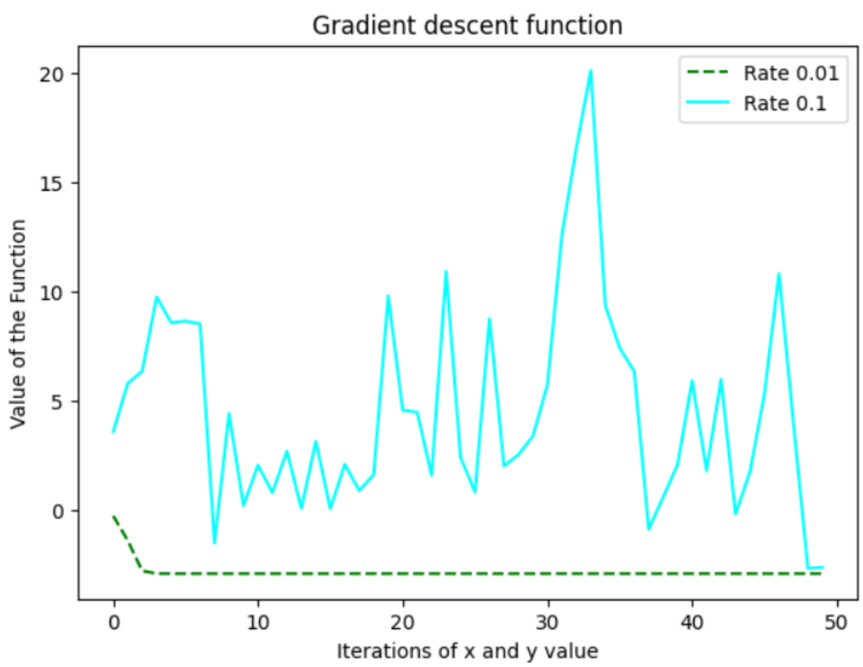
(7) Gradient Descent on a Simple Function. Consider the function

$$f(x, y) = 2x^2 + y^2 + 3\sin(2\pi x)\cos(2\pi y)$$

A. Implement gradient descent to minimize this function. Run gradient descent starting from the point ($x = 0.1$, $y = 0.1$). Set the learning rate to $\eta = 0.01$ and the number of iterations to 50. Give a plot that displays how the function value drops through successive iterations of gradient descent. Repeat this with a learning rate of $\eta = 0.1$ and provide a plot of the function value with each iteration. What do you observe?

B. Obtain the "minimum" value and location of the minimum value of the function you get using gradient descent with the same learning rate $\eta = 0.01$ and number of iterations (50) as part (a), from the following starting points: (i) (0.1, 0.1), (ii) (1, 1), (iii) (0.5, 0.5), (iv) (0.0, 0.5), (v) (−0.5, −0.5), (vi) (−1, 1). Write down the minimum value obtained using gradient descent and the location of the minimum value for each of these starting points. As you may appreciate, finding the "true" global minimum value of an arbitrary function is a hard problem.

```
The minimum value for (0.1, 0.1) is :  -2.8790846587644263  at location ( -0.24182894588827974 , -6.848884023658433e-35 )
The minimum value for (1, 1) is     :  -0.9286447086312597  at location ( 0.7252678803578847 , 0.9831635693235358 )
The minimum value for (0.5, 0.5) is :  -2.633242590975637  at location ( 0.24181813075114938 , 0.4916822590684804 )
The minimum value for (0.0, 0.5) is :  -2.633242590975637  at location ( 0.2418181307511494 , 0.491682259068480З )
The minimum value for (-0.5, -0.5) is :  -1.6660267055389741  at location ( -0.7253691676028068 , -0.4915941934004666 )
The minimum value for (-1, 1) is :  1.0051615759793315  at location ( -1.2084759495263577 , 0.9827889176425988 )
```



Access Link

Please refer to the link below for the google colab

Harsimran_Singh_Dhillon_ML_HW2