CSCI 4100 Assignment 7

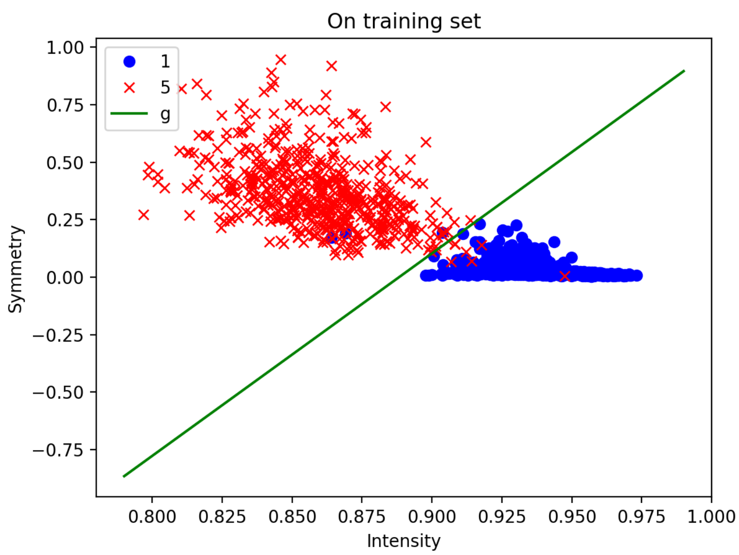
Boliang Yang 661541863

1. Classifying Handwritten Digits: 1 vs. 5

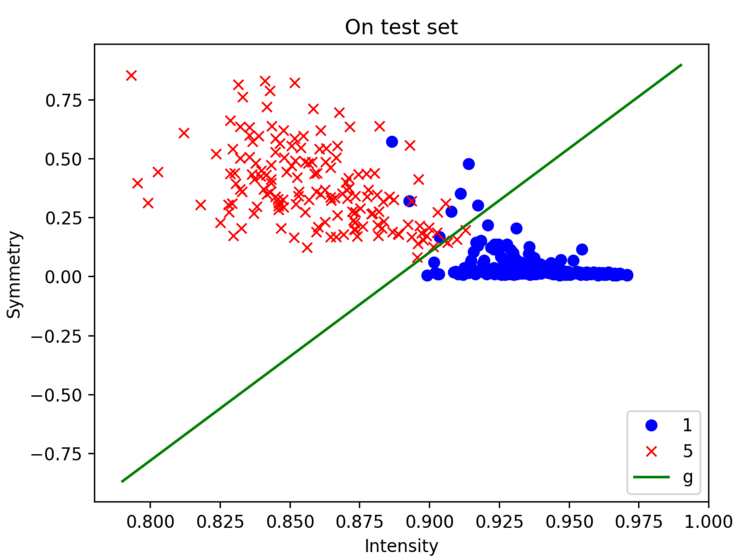
   I pick Linear Regression for classification followed by pocket for improvement.

   (a) With $\hat{y} = X(X^T X)^{-1} X^T y$, we compute initial weight and then update with pocket algorithm. The final hypothesis is $y = 8.805882x - 7.821971$

   Plot of the training data:



   Plot of the test data:



(b) $E_{in} = 0.00640614990391,\ E_{test} = 0.0235849056604$

(c) Error bound based on E_{in}:

Since for a linear perceptron in two dimensions, $d_{vc} = 3$. Since $N = 1561$, then

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4\left((2N)^{d_{vc}}+1\right)}{\delta}\right)} \leq 0.006406149903 + 0.38231711 \leq 0.3887232599$$

Error bound based on E_{test}:

Since there is only one hypothesis, we can simply use Hoeffding bound for the error bar from the test data set: $P[|E_{out}(g) - E_{test}(g)| \geq \epsilon] = 0.05 \leq 2e^{-2\epsilon^2 N}$

Since $N = 424$, we have $\epsilon^2 \leq \frac{\ln\frac{2}{0.05}}{2N} = 0.004350094$, then $\epsilon \leq 0.06595524$.

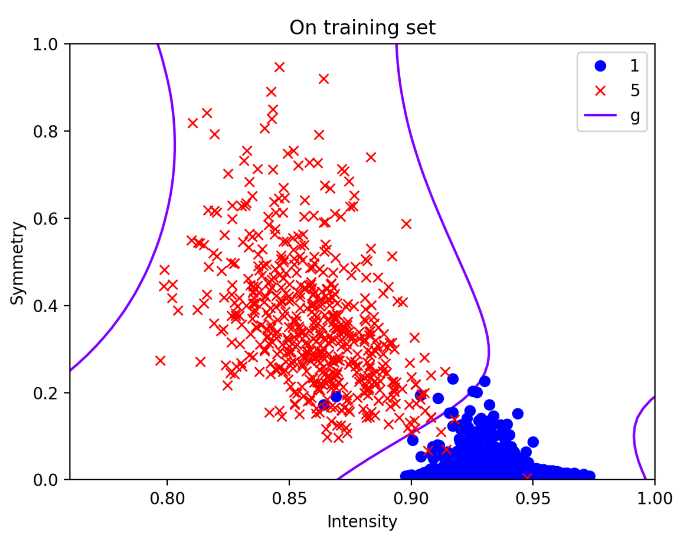
$$E_{out}(g) \leq E_{test}(g) + \epsilon = 0.0235849056604 + 0.06595524 = 0.0895401457$$

Thus, the error bound based on E_{test} is the better bound, and this is because that it has only one hypothesis rather than all linear hypotheses.

(d) Now we use a 3rd order polynomial transform, we replace $[1, x_1, x_2]$ with $[1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3]$. The final hypothesis is

$$\begin{aligned} y = 586.5385 - 2140.3126x_1 + 585.6888x_2 + 2565.9608x_1^2 \\ - 1296.62878x_1x_2 - 53.9275x_2^2 - 1012.3479x_1^3 \\ + 707.9874x_1^2x_2 + 74.9614x_1x_2^2 - 6.0069x_2^3 \end{aligned}$$

Plot of the training data:

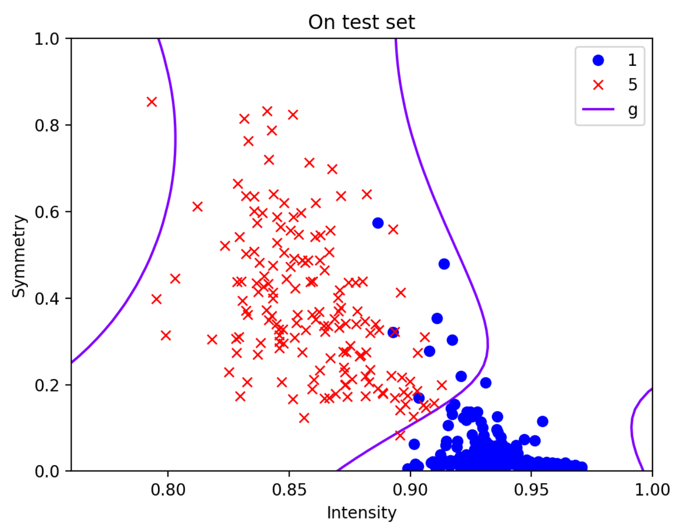


$E_{in} = 0.00640614990391$

Error bound based on E_{in}:

Since from problem we have $d_{vc} = d + 1 = 11$, $N = 1561$, then $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{vc}}+1)}{\delta}\right)} \leq 0.00640614990391 + 0.68996856 \leq 0.696374709$

Plot of the test data:



$E_{test} = 0.0235849056604$

Error bound based on E_{test}:

$$P[|E_{out}(g) - E_{test}(g)| \geq \epsilon] = 0.05 \leq 2e^{-2\epsilon^2 N}$$

Since $N = 424$, we have $\epsilon^2 \leq \frac{\ln\frac{2}{0.05}}{2N} = 0.004350094$, then $\epsilon \leq 0.06595524$.

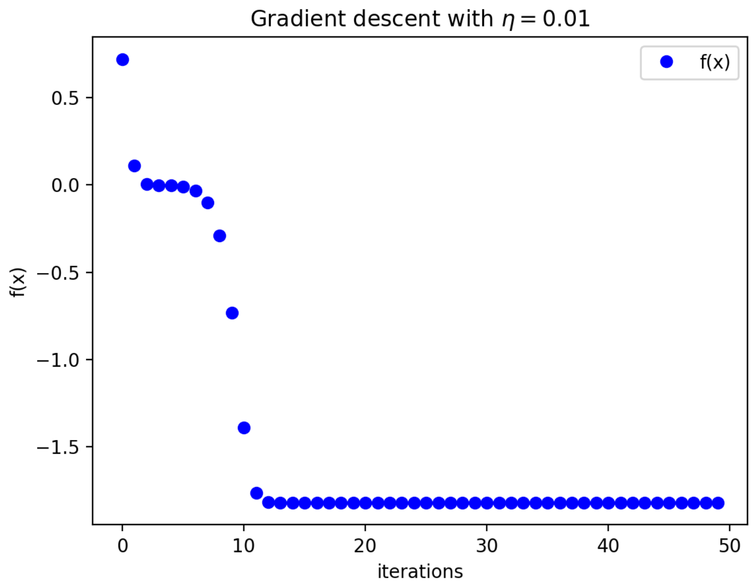
$E_{out}(g) \leq E_{test}(g) + \epsilon = 0.0235849056604 + 0.06595524 = 0.0895401457$

Thus, the error bound based on E_{test} is the better bound, and this is because that it has only one hypothesis rather than all linear hypotheses.

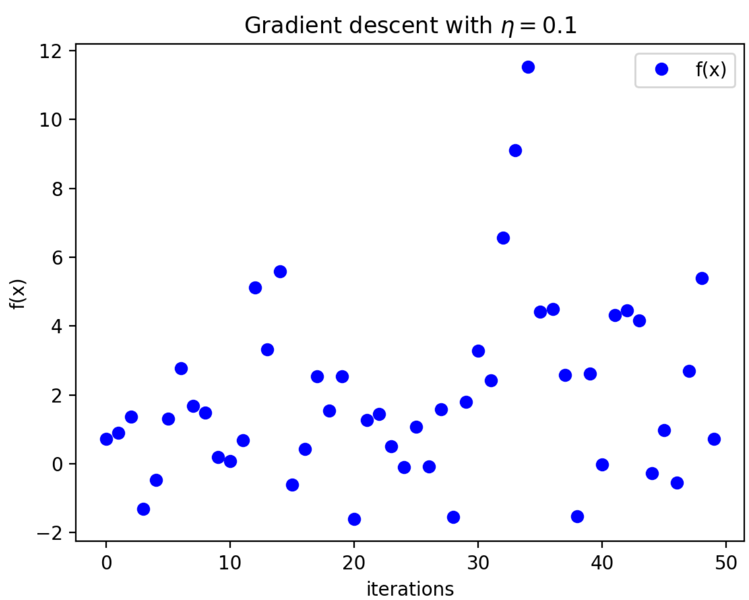
(e) The linear model without the 3rd order transform is better. The two models produce the same $E_{out}(g)$, then by Occam's razor principle, we should choose the hypothesis as simple as possible: the simplest model that fits the data is also the most plausible.

2. Gradient Descent on a "Simple" Function

   (a) The plot of gradient descent with $\eta = 0.01$:



The plot of gradient descent with $\eta = 0.1$:



With $\eta = 0.01$, the algorithm quickly finds the minimum in 50 iterations; with $\eta = 0.1$, the minimization cannot occur because gradient descent keeps overstepping the minimum.

(b)

Initial point	$\eta = 0.01$		$\eta = 0.1$	
	Location	Minimum value	Location	Minimum value
(0.1,0.1)	(0.2438,-0.2379)	-1.8201	(0.2036,-0.1742)	-1.6005
(1,1)	(1.2181,0.7128)	0.5933	(-0.6838,-0.1101)	-0.6753
(-0.5,-0.5)	(-0.7314,-0.2379)	-1.3325	(-0.2547,0.2840)	-1.7276
(-1,-1)	(-1.2181,-0.7128)	0.5932	(0.6838,0.1101)	-0.6753

Finding the "true" global minimum of an arbitrary function is a hard problem because the output from the gradient descent could change depending on the initial positions and step distance.

3. Problem 3.16 in LFD

(a) Since the cost only exist when wrong classification occurs, then $\text{cost(accept)} = P[y = -1|x]c_a = (1 - P[y = +1|x])c_a = \big(1 - g(x)\big)c_a$; and $\text{cost(reject)} = P[y = +1|x]c_r = g(x)c_r$

(b) Since we accept if $g(x) \geq k$, then $\text{cost(accept)} \leq \text{cost(reject)}$. And it follows that $\big(1 - g(x)\big)c_a \leq g(x)c_r$, which is $g(x) \geq \frac{c_a}{c_a+c_r}$. Thus, $k = \frac{c_a}{c_a+c_r}$

(c) For the Supermarket: $c_r = 10$, $c_a = 1$, then $k = \frac{c_a}{c_a+c_r} = \frac{1}{1+10} = \frac{1}{11}$

For CIA: $c_r = 1$, $c_a = 1000$, then $k = \frac{c_a}{c_a+c_r} = \frac{1000}{1000+1} = \frac{1000}{1001}$

Intuition: If c_a is larger and c_r is smaller, then k will be larger.