# Assignment 6

Alberto Mejia
RIN: 661514960

CSCI 4100 - Machine Learning from Data

October 14, 2019

1. (200) LFD Exercise 3.4

(a)

$y = w^{*T}x + \epsilon$
$\hat{y} = Xw^{*} + H\epsilon$
$D = \{(x_1,y_1),\ldots,(x_N,y_N)\}$
$\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2,\ldots, \epsilon_N]^T$
$H = X(X^TX)^{-1}X^T$

$\hat{y} = Xw_{lin}$ -> $X(X^TX)^{-1}X^Ty$
$\hat{y} = Hy$ -> $H(w^{*T}x + \epsilon)$
$= X(X^TX)^{-1}X^T(w^{*T}x + \epsilon)$
$= Xw^{*} + X(X^TX)^{-1}X^T\epsilon$
$= Xw^{*} + H\epsilon$

(b) $\hat{y} - y = (Xw^{*} + H\epsilon) - (Xw^{*} + \epsilon) = (H - I)\epsilon$

(c) $E_{in}(w) = \frac{1}{N}||Xw - y||^2$ (3.3)
$E_{in}(w_{lin}) = \frac{1}{N}||\hat{y} - y||^2$
$= \frac{1}{N}||(H - I)\epsilon||^2$
$= \frac{1}{N}[(H - I)\epsilon]^T[(H - I)\epsilon]$
$= \frac{1}{N}\epsilon^T[(H - I)]^T[(H - I)\epsilon]$
$= \frac{1}{N}\epsilon^T(H - I)^2\epsilon$
H – I is symmetric, thus $(H - I)^T = H - I$
Therefore, $(I - H)^T = I - H$ Exercise 3.3 part (c)
$= \frac{1}{N}\epsilon^T(I - H)^2\epsilon$
$= \frac{1}{N}\epsilon^T(I - H)\epsilon$

(d) Prove $E_D[E_{in}(w_{lin})] = \sigma^2(1 - \frac{d+1}{N})$
$trace(H) = d + 1$ Exercise 3.3 part (d)
$E_D[\epsilon] = \sigma^2$
$E_D[\epsilon^T\epsilon] = N\sigma^2$
$E_D[\frac{1}{N}\epsilon^T\epsilon] = \frac{1}{N}N\sigma^2$ -> $\frac{N\sigma^2}{N}$ -> $\sigma^2$
$E_D[\epsilon^TH\epsilon] = trace(H)\sigma^2$ $E_D[\epsilon^TH\epsilon]$ is a diagonal matrix

$E_D[E_{in}(w_{lin})] = E_D[\frac{1}{N}\epsilon^T(I - H)\epsilon]$
$= E_D[\frac{1}{N}\epsilon^T(I\epsilon - H\epsilon)]$
$= E_D[\frac{1}{N}\epsilon^T\epsilon - \frac{1}{N}\epsilon^TH\epsilon]$

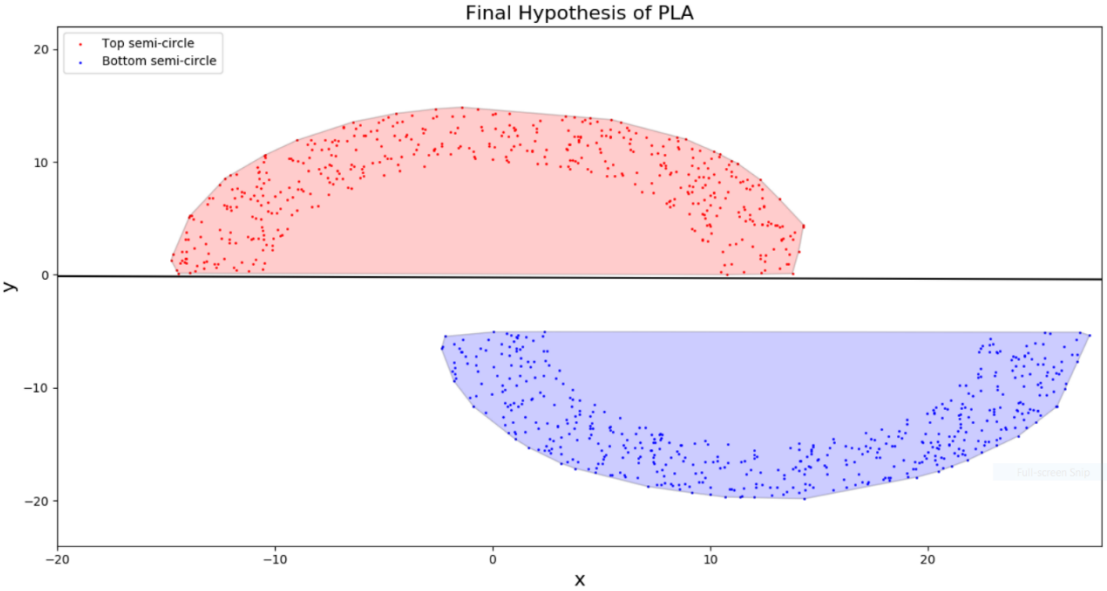
$$
\begin{aligned}
&= \frac{1}{N}E_D[\epsilon^T\epsilon] - \frac{1}{N}E_D[\epsilon^T H\epsilon] \\
&= \frac{1}{N}(N\sigma^2 - E_D[\epsilon^T H\epsilon]) \\
&= \frac{1}{N}(N\sigma^2 - \textit{trace(H)}\sigma^2) \\
&= \frac{1}{N}(N\sigma^2 - (d+1)\sigma^2) \\
&= \sigma^2 - \frac{1}{N}\sigma^2(d+1) \\
&= \sigma^2 - \sigma^2\frac{(d+1)}{N} \\
&= \sigma^2(1 - \frac{(d+1)}{N})
\end{aligned}
$$

(e) Prove $E_{D,\epsilon'}[E_{test}(w_{lin})] = \sigma^2(1 + \frac{d+1}{N})$

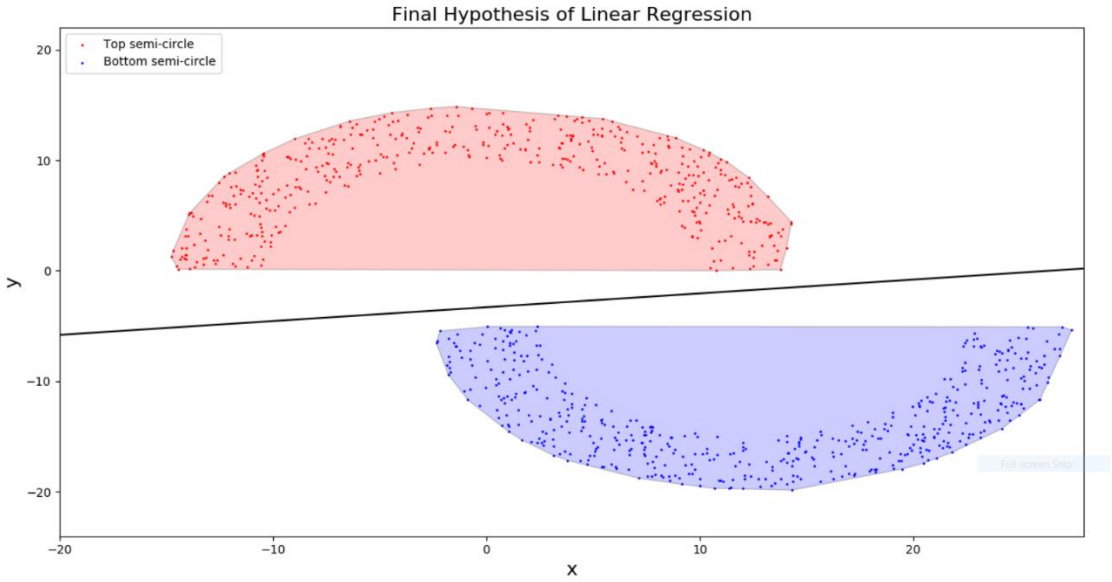
$$
\begin{aligned}
E_{D,\epsilon'}[E_{test}(w_{lin})] &= E_{D,\epsilon'}[\frac{1}{N}||Xw - y||^2] \\
&= \frac{1}{N}[||(H\epsilon - \epsilon')||^2] \\
&= \frac{1}{N}[||(H\epsilon - \epsilon')^T(H\epsilon - \epsilon')||^2] \\
&= \frac{1}{N}[||(H^T\epsilon^T - \epsilon'^T)(H\epsilon - \epsilon')||^2] \\
&= \frac{1}{N}[||H^T\epsilon^T H\epsilon - H^T\epsilon^T\epsilon' - \epsilon'^T H\epsilon + \epsilon'^T\epsilon'||^2] \\
&= \frac{1}{N}[||H^T\epsilon^T H\epsilon - 2\epsilon^T H^T\epsilon' + \epsilon'^T\epsilon'||^2] \\
&= \frac{1}{N}[||H^T\epsilon^T H\epsilon - 2\epsilon^T H^T\epsilon' + \epsilon'^T\epsilon'||^2] \\
&= E_{D,\epsilon'}[\frac{1}{N}[||H^T\epsilon^T H\epsilon - 2\epsilon^T H^T\epsilon' + \epsilon'^T\epsilon'||^2]] \\
&= E_{D,\epsilon'}\frac{1}{N}[||H^T\epsilon^T H\epsilon - 2\epsilon^T H^T\epsilon' + \epsilon'^T\epsilon'||^2] \\
&= \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) \\
&= \sigma^2(1 + \frac{d+1}{N})
\end{aligned}
$$

2. (200) LFD Problem 3.1

(a)

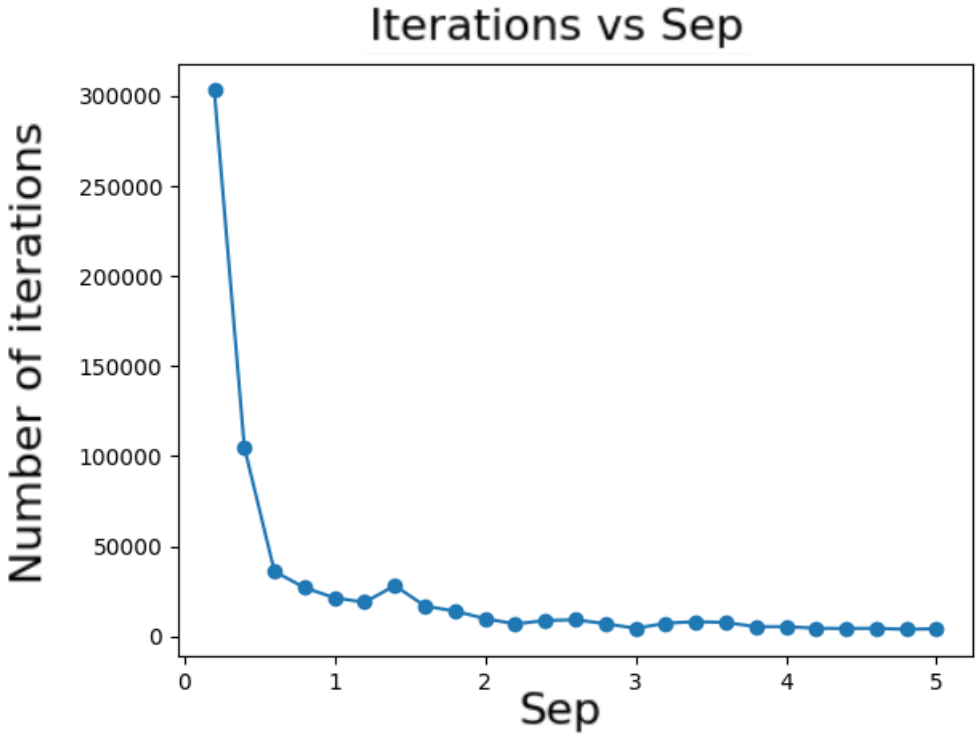


(b)




As we can observe, using linear regression we get a smaller in sample error as it produces a more precise result.

3. (200) LFD Problem 3.2




We can observe a couple of things from the graph. One of them is that as the Sep increases, the number of iterations decrease, AKA the number of steps for convergence. Thus, the points also shrink in distance as Sep increases.

4. (200) LFD Problem 3.8

For Linear Regression: $E_{out}(h) = E[(h(x) - y)^2]$

$y = h^*(x) + \epsilon(x)$

$\epsilon(x) = y - h^*(x)$

$(a+b)^2 = a^2 + 2ab + b^2$ (1) Algebraic formula

Show that $E_{out}(h)$ is minimized by: $h^*(x) = E[y|x]$ and show $E[\epsilon(x)] = 0$

$$
\begin{aligned}
E_{out}(h) &= E[(h(x) - y)^2] \\
&= E[(h(x) - (h^*(x) + \epsilon(x)))^2] \\
&= E([h(x) - (h^*(x) + y - h^*(x)))^2] \\
&= E[((h(x) - h^*(x)) - y + h^*(x))^2] \\
&= E[((h(x) - h^*(x)) + (h^*(x) - y))^2] \\
&= E[(h(x) - h^*(x))^2 + (h^*(x) - y)^2] \\
&= E[(h(x)-h^*(x))^2 + 2(h(x)-h^*(x))(h^*(x)-y) + (h^*(x)-y)^2] \quad \text{Apply (1)} \\
&= E[(h(x) - h^*(x))^2] + 2E[(h(x) - h^*(x))(h^*(x) - y)] + E[(h^*(x) - y)^2]
\end{aligned}
$$

Assume minimization of $E_{out}(h)$ by $h^*(x) = E[y|x]$

$E[(h(x) - h^*(x))(h^*(x) - y)]$

$\nabla E(h^*) = E[(h^*(x) - y)] = 0$

$$
\begin{aligned}
E[y|x] &= E[h^*(x) + \epsilon(x)|x] \\
&= E[h^*(x)|x] + E[\epsilon(x)|x]
\end{aligned}
$$

$E[\epsilon(x)|x] = 0$

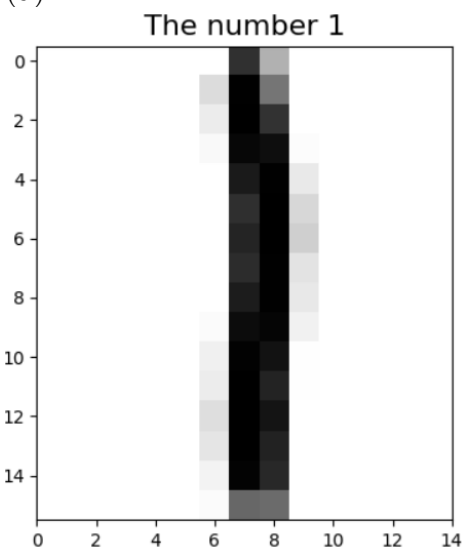
$E[(h(x) - h^*(x))(h^*(x) - y)] = 0$

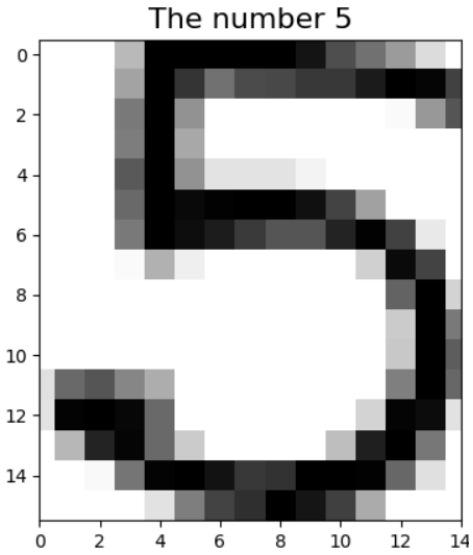
$\therefore E_{out}(h)$ is minimized when $h^*(x) = E[y|x]$

5. (200) LFD Problem 3.6 (6xxx Level Only)

6. (200) Handwritten Digits Data - Obtaining Features

(a)



(b) We choose intensity and symmetry as our two features to measure the properties of the images that will help us distingiush between numbers like 1 and 5. These features are chosen not only because Professor Malik told us too, but also because the data is preprocessed already using Computer Vision to fit all numbers into the same aspect ratio(16×16) and then extra noise is filtered out (background is stricly one color and the numbers are stricly grayscale).
Let D be a matrix for a digit.

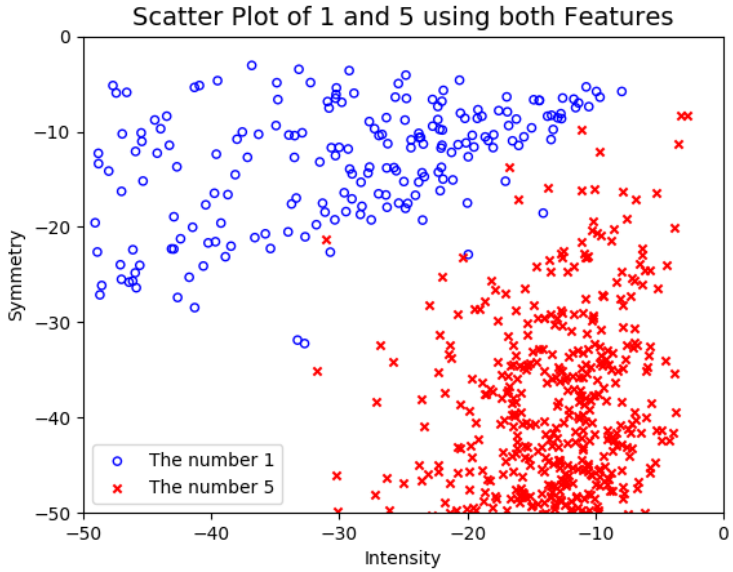
For intensity, we just do the following for each digit: Take the sum of all the grayscale values to determine the intensity of black it has per pixel. Thus, we get the following formula

$$Intensity = \sum_{i=0}^{15}(D[0][i] + D[1][j] + \ldots + D[n][i])$$

For Symmetry, the logic is as follows: Calculate the horizontal symmetry of the image, which is defined as the difference between the image and its opposite-axis inverted version. Mathematically, we'll define it as
$(D[i][j] - D[k][l]) = (r)$ if $(D[i][j] - D[k][l] > 0)$
$(D[i][j] - D[k][l]) = -1*(r)$ if $(D[i][j] - D[k][l] < 0)$
We can see that this is the piecewise function for absolute value.

$$Symmetry = \sum_{i=0}^{15}\sum_{0}^{7}|D[i][j] - D[i][15-j]|$$

(c)




As we can see, an increase in symmetry shows an increase of ones where as fives are more situated in the places with less symmetry. Fives and ones also have enough intensity difference to show, fives have more black pixels and thus more 'average'

intensity. Therefore they will be found more along increasing intensity. Being that we are just comparing pixels for symmetry, the fives will start to dip into the symmetry side of the graph but not as much as the ones. Less symmetrical ones will also dip into the intensity side since some ones will have “five” like attributes (like those fancy ones with an angled top and wide bottom)