Exercise 3.4

a)

We have $\widehat{y} = X(X^TX)^{-1}X^Ty$

$H = X(X^TX)^{-1}X^T$ (3.6)

$\widehat{y} = Hy$

From the problem, we have

$y = w^{*T}x + \epsilon$

$\widehat{y} = H(w^{*T}x + \epsilon)$

$\widehat{y} = H(Xw^* + \epsilon)$

$\widehat{y} = X(X^TX)^{-1}X^T(Xw^* + \epsilon)$

$\widehat{y} = X(X^TX)^{-1}X^TXw^* + X(X^TX)^{-1}X^T\epsilon$

$\widehat{y} = Xw^* + X(X^TX)^{-1}X^T\epsilon$

$\widehat{y} = Xw^* + H\epsilon$

b)

$\widehat{y} - y = Xw^* + H\epsilon - Xw^* - \epsilon$

$= H\epsilon - \epsilon$

$= (H - I)\epsilon$

Where I is an identity matrix.

$H - I$ is the matrix

c)

$E_{in}(w_{lin}) = \frac{1}{N}(\widehat{y} - y)^2$

$= \frac{1}{N}||(H - I)\epsilon||^2$

$= \frac{1}{N}\epsilon^T(H - I)^T(H - I)\epsilon$

Since H is symmetric.

$= \frac{1}{N}\epsilon^T(I - H)^2\epsilon$

From exercise 3.3, we know that $(I - H)^K = I - H$

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

d)

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

$= \frac{1}{N}E_D[\epsilon^T I\epsilon - \epsilon^T H\epsilon]$

$= \frac{1}{N}(E_D[\sum_{n=1}^{N} \epsilon_n^2] - E_D[\sum_{i=1}^{N}\sum_{j=1}^{N} \epsilon_i H_{ij}\epsilon_j])$

$= \frac{1}{N}(\sum_{n=1}^{N} E_D[\epsilon_n^2] - \sum_{i=1}^{N}\sum_{j=1}^{N} E_D[\epsilon_i H_{ij}\epsilon_j])$

Since ϵ has a zero mean, this means its expected value of 0. Also, it has a variance σ^2.

$= \frac{1}{N}(N\sigma^2 - \sum_{i=1}^{N} E_D[\epsilon_i^2 H_{ii}])$

$= \frac{1}{N}(N\sigma^2 - \sum_{i=1}^{N} H_{ii}E_D[\epsilon_i^2])$

$= \frac{1}{N}(N\sigma^2 - tr(H)\times E_D[\epsilon_i^2])$

$= \frac{1}{N}(N\sigma^2 - (d + 1)\sigma^2)$

$= \sigma^2(1 - \frac{d+1}{N})$

e)

We got $y‘ = w^{*T}x + \epsilon' = Xw^* + \epsilon'$

$\hat{y} = Xw^* + H\epsilon'$

$\hat{y} - y = Xw^* + H\epsilon - Xw^* - \epsilon' = H\epsilon - \epsilon'$

$E_{test}(w_{lin}) = \frac{1}{N}(\hat{y} - y)^2 = \frac{1}{N}||H\epsilon - \epsilon'||^2$

$E_{D,\epsilon'}[E_{test}(w_{lin})] = E_{D,\epsilon'}[\frac{1}{N}(H\epsilon - \epsilon')^T(H\epsilon - \epsilon')]$

$= E_{D,\epsilon'}[\frac{1}{N}(\epsilon^T H - \epsilon'^T)(H\epsilon - \epsilon')]$

$= E_{D,\epsilon'}[\frac{1}{N}(\epsilon^T H^2\epsilon - 2\epsilon'^T H\epsilon + \epsilon'^T\epsilon')]$

Since H = H^K

$= \frac{1}{N}E_{D,\epsilon'}[\epsilon^T H\epsilon - 2\epsilon'^T H\epsilon + \epsilon'^T\epsilon']$

From d) we know that $E_{D,\epsilon'}[\epsilon^{T} H\epsilon] = (d + 1)\sigma^{2}$ and $E_{D,\epsilon'}[\epsilon'^{T}\epsilon'] = N\sigma^{2}$

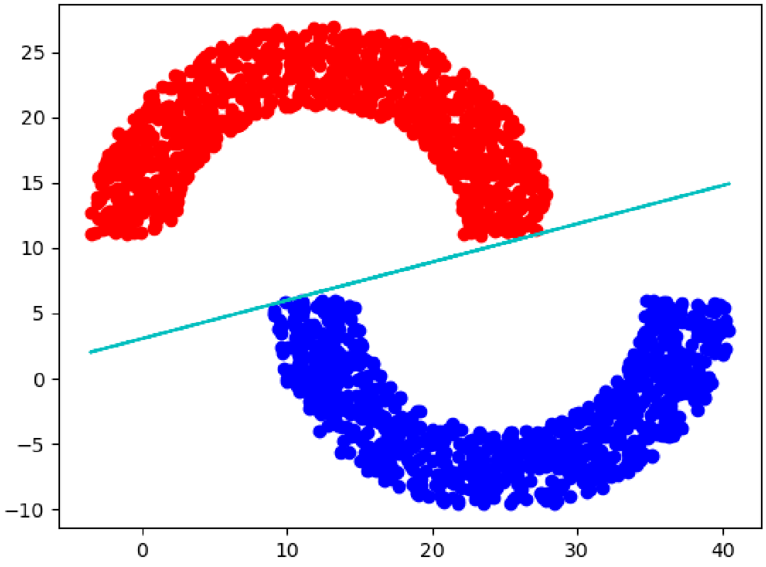
$E_{D,\epsilon'}[2\epsilon'^{T} H\epsilon] = 2E_{D,\epsilon'}[\sum_{i}^{N}\sum_{j}^{N}\epsilon_{i}'^{T} H\epsilon_{j}] = 2\sum_{i}^{N}\sum_{j}^{N} E_{D,\epsilon'}[\epsilon_{i}'^{T} H\epsilon_{j}]$, since the expected value of ϵ is 0, $E_{D,\epsilon'}[2\epsilon'^{T} H\epsilon] = 0$

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

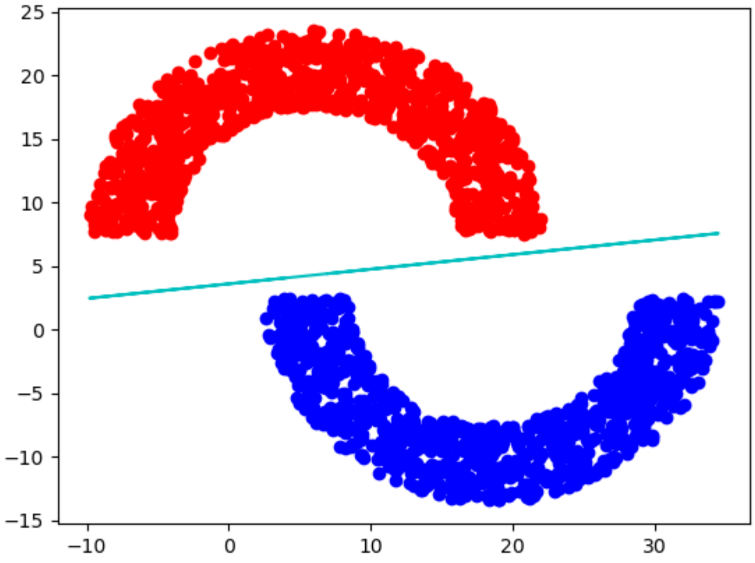
$= \sigma^{2}(1 + \frac{d+1}{N})$

Problem 3.1

a) PLA

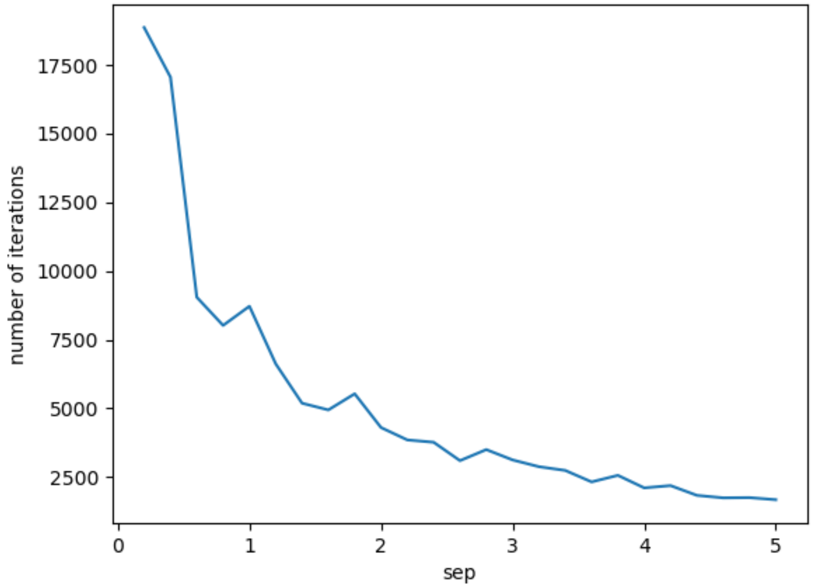


b) Linear Classifier



When using PLA, we stop when no points are misclassified, which means the line will stop right after the last points are correctly classified. This means the line will be close to some data point. However, when using linear regression, the w_{lin} is gained by using first derivative to minimize the average square error between y and $\widehat{y}$. The line will stay right in the place where the sum of the square error is minimized.

Problem 3.2



From the graph, we observe that as the more the data is separated, the faster the PLA converges. According to the equation $t \le \frac{R^2||w^*||^2}{\rho^2}$ from problem 1.3, $R = max_{1\le n\le N}||x_n||$ will be constant since we got a certain circle. As for $||w^*||^2$ and ρ^2, where $\rho = min_{1\le n\le N}y_n(w^{*T}x_n)$, the $\frac{min_{1\le n\le N}(w^{*T}x_n)}{||w||}$ is the minimum distance between the plane $w^{*T}x$ and the point x_n and the $||w^*||/\rho$ is the inverse of it. Hence we can conclude that the larger the inverse of the distance(smaller sep), the larger the iteration number; also, smaller the inverse of the distance(larger sep), the smaller the iteration number. This corresponds to the observation that sep and iteration have an inverse relationship.

Problem 3.8

$E_{out}(h) = E_{x,y}[(h(x) - y)^2]$

In order to get the minimum, we need to take the first derivative and let it equal to 0, so we have:

$\frac{dE_{out}(h)}{dh} = E_{x,y}[2(h(x) - y)]$

$= 2E_{x,y}[(h(x) - y)]$

$= 2E_x E_{y|x}[(h(x) - y)]$

$E_{y|x}[h(x)] - E_{y|x}[y] = 0$

$E_{y|x}[y] = E_{y|x}[h(x)]$

$h(x) = E[y \mid x]$

Hence, we could conclude that $h^*(x) = E[y \mid x]$

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

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

$E[y] = E[E[y \mid x]] + E[\epsilon(x)]$

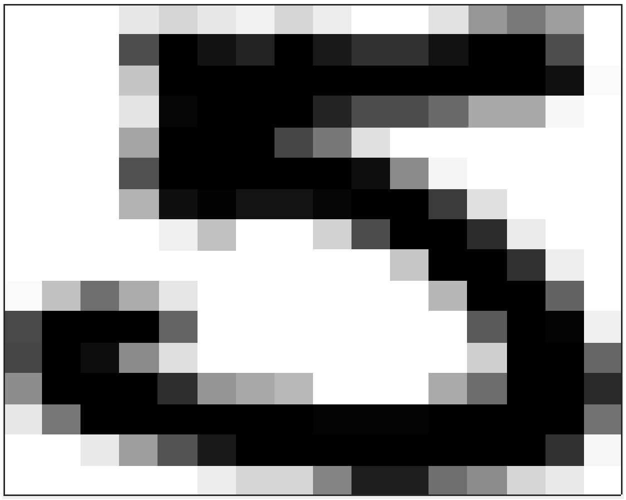
$E[\epsilon(x)] = E[y] - E[E[y \mid x]]$

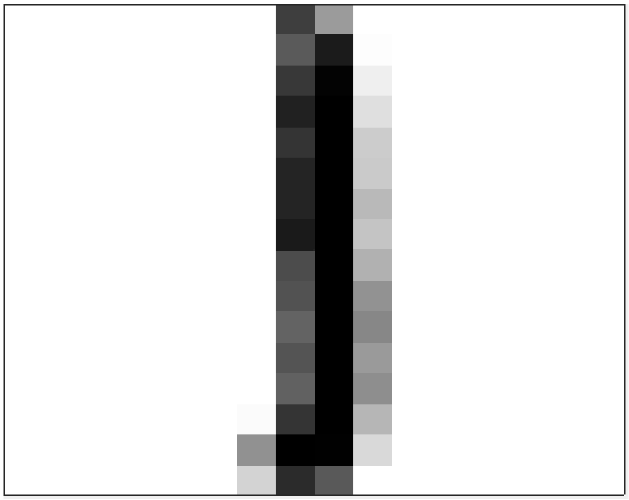
$E[\epsilon(x)] = E[y] - E[y]$

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

Handwritten Digits Data - Obtaining Features

a)





b)

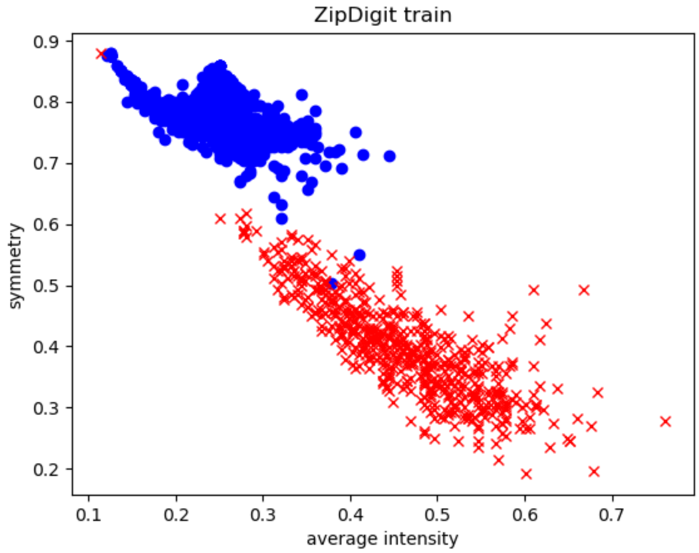
Average intensity:

I would like define it as the number of non-white (!= -1) pixels divided by the number of total pixels which is 256. Mathematically, it is the occupancy of black pixels in the 16x16 grid, which is num_of_black / 256.

Symmetry:

I would like to define it as the similarity between the original image and its flipped version. I will flip the matrix horizontally and vertically and compare it with the original matrix and compute the similarity. Mathematically, it is average of horizontal similarity and vertical similarity.

c)
Training set:


ZipDigit train
symmetry
average intensity