LFD Problem Set 6

John Cohen

October 21, 2024

Exercise 3.4

Consider a noisy target $y = w^{*T}x + \epsilon$ for generating the data, where ϵ is a noise term with zero mean and σ^2 variance, independently generated for ever example (x,y). The expected error of the best possible linear fit to this target is thus σ^2.
For the data $D = \{(x_1, y_1,)...(x_N, y_N)\}$, denote the noise in y_n as ϵ_n and let $\epsilon = [\epsilon_1, ..., \epsilon_N]^T$; assume that X^TX is invertible. By following the steps below, show that the expected in-sample error of linear regression with respect to D is given by

$$E_D[E_{in}(w_{lin})] = \sigma^2(1 - \frac{d+1}{N}).$$

(a) Show that the in-sample estimate of y is given by $y^- = Xw^* + H\epsilon$.

With $y = w^{*T}x + \epsilon$ and $H = X(X^TX)^{-1}X^T$ and $y^- = Hy$:

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

$$\begin{aligned} y^- = Hy = X(X^TX)^{-1}X^T(Xw^* + \epsilon) &= X(X^TX)^{-1}X^TXw^* + X(X^TX)^{-1}X^T\epsilon \\ &= Xw^* + X(X^TX)^{-1}X^T\epsilon = Xw^* + H\epsilon \end{aligned}$$

(b) Show the the in-sample error vector $y^- - y$ can be expressed by a matrix times ϵ. What is the matrix?

$$\begin{aligned} y^- - y = Xw^* + H\epsilon - (w^{*T}x + \epsilon) &= Xw^* + H\epsilon - Xw^* - \epsilon \\ &= H\epsilon - \epsilon = (H - I)\epsilon \end{aligned}$$

This shows the matrix is $(H - I)$.

(c) Express $E_{in}(w_{lin})$ in terms of ϵ using (b), and simplify the expression using Exercise 3.3(c).

$$\begin{aligned}
E_{in}(w_{lin}) &= \frac{1}{N}||y^{-} - y||^2 \\
&= \frac{1}{N}||(H-I)\epsilon||^2 \\
&= \frac{1}{N}\epsilon^T||(H-I)||^2\epsilon \\
&= \frac{1}{N}\epsilon^T(I-H)\epsilon
\end{aligned}$$

(d) Prove the $E_D[E_{in}(w_{lin}) = \sigma^2(1 + \frac{d+1}{N})$ using (c) and the independence of $\epsilon_1, ..., \epsilon_N$.

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

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

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

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

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

$$E_D[E_{in}(w_{in})] = \frac{1}{N}[N\sigma^2 - (trace(H))\sigma^2]$$

and with trace(H) = d+1 from 3.3:

$$E_D[E_{in}(w_{in})] = \frac{1}{N}[N\sigma^2 - (d+1)\sigma^2]$$

$$E_D[E_{in}(w_{in})] = \sigma^2 - \frac{(d+1)\sigma^2}{N}$$

$$E_D[E_{in}(w_{in})] = \sigma^2(1 - \frac{(d+1)}{N})$$

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

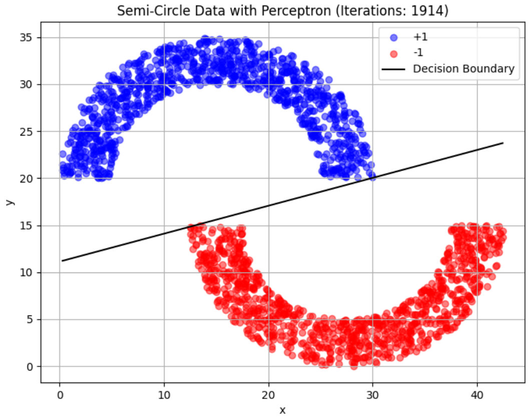
$$y^{-} = Xw^{*} + H\epsilon$$
$$E_{test}(w_{lin}) = \frac{1}{N}||y^{-} - y||^2$$
$$E_{test}(w_{lin}) = \frac{1}{N}||Xw^{*} + H\epsilon - (Xw^{*} + \epsilon^{'}||^2$$
$$E_{test}(w_{lin}) = \frac{1}{N}||H\epsilon - \epsilon^{'}||^2$$
$$E_{test}(w_{lin}) = \frac{1}{N}(H\epsilon - \epsilon^{'})^T(H\epsilon - \epsilon^{'})$$
$$E_{test}(w_{lin}) = \frac{1}{N}(\epsilon^T H - \epsilon^{'T})(H\epsilon - \epsilon^{'})$$
$$E_{test}(w_{lin}) = \frac{1}{N}(\epsilon^T HH\epsilon - 2\epsilon^{'T}H\epsilon + \epsilon^{'T}\epsilon^{'})$$
$$E_{test}(w_{lin}) = \frac{1}{N}(\epsilon^T H\epsilon - 2\epsilon^{'T}H\epsilon + \epsilon^{'T}\epsilon^{'})$$
$$E_{D,\epsilon'}[E_{test}(w_{lin})] = E_{D,\epsilon'}[\frac{1}{N}(\epsilon^T H\epsilon - 2\epsilon^{'T}H\epsilon + \epsilon^{'T}\epsilon^{'})]$$
$$E_{D,\epsilon'}[E_{test}(w_{lin})] = \frac{1}{N}[E_{D,\epsilon'}(\epsilon^T H\epsilon) - 2E_{D,\epsilon'}(\epsilon^{'T}H\epsilon) + E_{D,\epsilon'}(\epsilon^{'T}\epsilon^{'})]$$
$$E_D[\epsilon^T\epsilon] = N\sigma^2$$
$$E_D[\epsilon^T H\epsilon] = (d+1)\sigma^2$$
$$E_{D,\epsilon'}[E_{test}(w_{lin})] = \sigma^2(1 + \frac{d+1}{N}) - \frac{2}{N}E_{D,\epsilon'}(\epsilon^{'T}H\epsilon)$$
$$E_{D,\epsilon'}(\epsilon^{'T}H\epsilon) = E_{D,\epsilon'}(trace(\epsilon^{'T}H\epsilon))$$
$$E_{D,\epsilon'}(\epsilon^{'T}H\epsilon) = E_{D,\epsilon'}(\sum_{i=1}^{N}\epsilon_i^{'}H_{ii}\epsilon_i)$$
$$E_{D,\epsilon'}(\epsilon^{'T}H\epsilon) = \sum_{i=1}^{N}(E(\epsilon_i^{'})H_{ii}E(\epsilon_i))$$
$$E_{D,\epsilon'}(\epsilon^{'T}H\epsilon) = 0$$

$$E_{D,\epsilon'}[E_{test}(w_{lin})] = \sigma^2(1 + \frac{d+1}{N}) - \frac{2}{N}(0)$$
$$E_{D,\epsilon'}[E_{test}(w_{lin})] = \sigma^2(1 + \frac{d+1}{N})$$

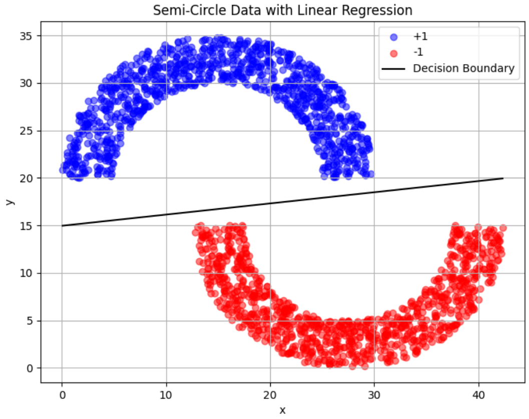
Problem 3.1

Consider the double semi-circle "toy" learning task below. There are two semi-circles of width thk with inner radius rad, separated by sep as shown (red is -1 and blue is +1). The center of the top semi-circle is aligned with the middle of the dege of the bottom semi-circle. The task is linearly separable when $sep \geq 0$, and not so for $sep < 0$. Set rad = 10, thk - 5 and sep = 5. Then, generate 2,00 example uniformly, which means you will have approximately 1,000 example for each class.

(a) Run the PLA starting from w = 0 until it converges. Plot the data and the final hypothesis.



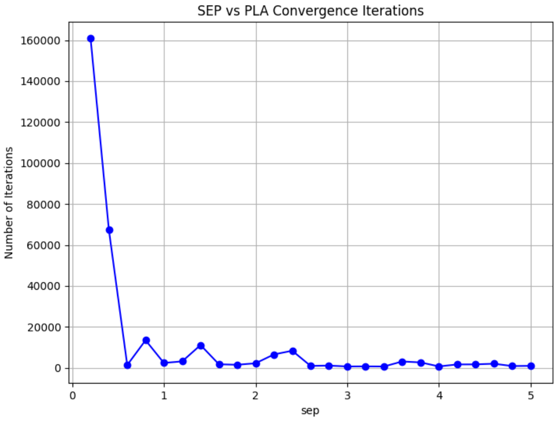
(b) Repeat part (a) using the linear regression (for classification) to obtain w. Explain your observations.



I observed the the final hypothesis given by linear regression seemed to be more "in the middle” of the two regions. This makes sense with the difference in how it is computed versus the PLA. Linear regression minimizes the total squared error across all data points resulting in a more balanced (”middle”) hyperplane (line in this case). PLA searches fro a hyperplane that perfectly separates the data but its result depends on the initial weights and more importantly in this case, the order that the data points are presented. The PLA will produce more varying results than linear regression if we were to repeat the experiment multiple times and especially with the same data points in a different order.

Problem 3.2

For the double-semi-circle task in Problem 3.1, vary sep in the range {0.2,0.4,...,5}. Generate 2,000 examples and run the PLA starting with w = 0. Record the number of iterations PLA takes to converge.
Plot sep versus the number of iterations taken for PLA to converge. Explain your observations. [Hint: Problem 1.3.]



As the separation increases the amount of iterations the PLA takes to converge decreases. Conceptually this makes a lot of sense. With a larger separation value, there are more acceptable solutions the PLA can compute, decreasing the chance that it will over-correct during a given iteration (the chance it will need another iteration) and thus decreasing the expected number of iterations required before reaching an acceptable line. Problem 1.3 gives us a bound on the number of iterations the PLA requires. This bound is $K \leq \frac{R^{@}||w^*||^2}{\rho^2}$ where K is the number of iterations that PLA takes to converge, w^* is the optimal weight vector or line, $R = max_{1 \leq n \leq N}||x_n||$ is the maximum norm of input vectors, and $\rho = min_{1 \leq n \leq N} y_n(w^* x_n)$ is smallest ”distance” between any data point and the decision boundary given by w^*. Increasing the sep value is akin to increasing the all distances between any data point and w^*, including the smallest. Sep is directly related to rho. So, increasing sep, increases ρ, decreasing the upper bound on iterations K in $K \leq \frac{R^{@}||w^*||^2}{\rho^2}$. A higher sep value will decrease the upper bound on iterations explaining why we see a decrease in iterations with an increase in sep values in the experiment!

Problem 3.8

For linear regression, the out-of-sample error is

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

Show that among all hypotheses, the one that minimizes E_{out} is given by

$$h^*(x) = E[y|x].$$

The function h^* can be treated as a deterministic target function, in which case we can write $y = h^*(x) + \epsilon(x)$ where $\epsilon(x)$ is an (input dependent) noise variable. Show that $\epsilon(x)$ has expected value zero.

Prove $h^*(x) = E[y|x]$ minimizes E_{out}:
The expected error $E_{out}(h) = E[(h(x) - y)^2]$ can be written as an expectation over x and y:

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

Let $\Delta x = h(x) - E[y|x]$. Now,

$$(h(x)-y)^2 = (h(x)+(-E[y|x]+E[y|x])-y)^2 = ((h(x)-E[y|x])+E[y|x]-y)^2 = (\Delta(x)+E[y|x]-y)^2$$

$$= \Delta(x)^2 + 2\Delta(x)(E[y|x] - y) + (E[y|x] - y)^2$$

$$E_{y|x}[(h(x) - y)^2] = \Delta(x)^2 + 2\Delta(x)E_{y|x}[E[y|x] - y] + E_{y|x}[(E[y|x] - y)^2]$$

Since $E_{y|x}[E[y|x] - y] = E[y|x] - E_{y|x}[y] = 0$, then clearly $E_{y|x}[(h(x) - y)^2] = \Delta(x)^2 + 0 + 0$. The expected error minimizes when $\Delta(x) = 0$. With this and $\Delta x = h(x) - E[y|x]$, $0 = h^*(x) - E[y|x]$. $h^*(x) = E[y|x]$.

Show $\epsilon(x)$ has zero mean:
With $y = h^*(x) + \epsilon(x)$,

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

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

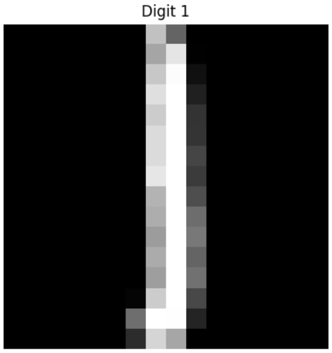
Therefore $\epsilon(x)$ has an expected value of zero given x, meaning it represents a noise term centered around zero.

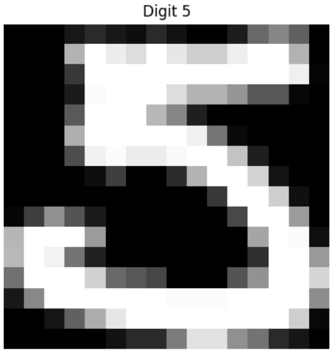
Problem 3.6

Handwritten Digits

You can download the two data files with handwritten digits data: training data (ZipDigits.train) and test data (ZipDigits.test). Each row is a data example. The first entry is the digit, and the next 256 are grayscale values between -1 and 1. 256 pixels corresponds to a 16 × 16 image. For this problem, we will only use the 1 and 5 digits, so remove the other digits from your training and test examples.

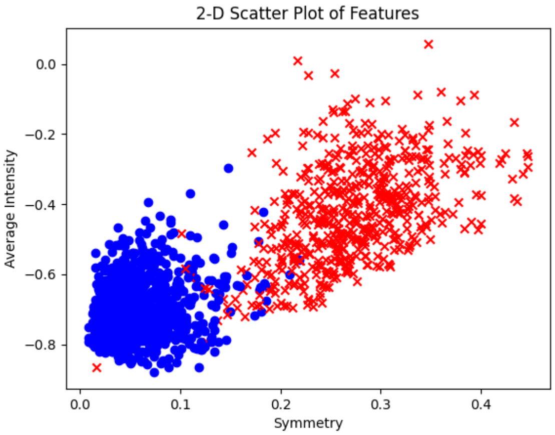
(a) Familiarize yourself with the data by giving a plot of two of the digit images.


Digit 1


Digit 5

(b) Develop two features to measure properties of the image that would be useful in distinguishing between 1 and 5. You may use symmetry and average intensity (as discussed in class) or anything else you think will work better. Give the mathematical definition of your two features.
I will be using symmetry and average intensity to distinguish between 1 and 5.
Synmnetry: $\frac{1}{256}\sum_{i=1}^{16}\sum_{j=1}^{8}|p_{ij} - p_{i(16-j+1)})|$
Average Intensity: $\frac{1}{256}\sum_{i=0}^{16}\sum_{j=0}^{16} p_{ij}$
where p_{ij} represents the pixel value, 1 if black and 0 if white, of the pixel at row i column j in the given image.

(c) As in the text, give a 2-D scatter plot of your features: for each data example, plot the two features with a red '×' if it is a 5 and a blue '○' if it is a 1.


2-D Scatter Plot of Features
Average Intensity
0.0
−0.2
−0.4
−0.6
−0.8
Symmetry
0.0
0.1
0.2
0.3
0.4