Machine Learning from Data Assignment 6

Greg Stewart

October 18, 2018

Exercise 3.4

Consider noisy target $y = w^{*T}x + \epsilon$ for generating data, where ϵ is a noise term with zero mean and σ^2 variance, independently generated for every example $\mathbf{x}, y$. Expected error of best possible linear fit to the target is σ^2. For the data D denote the noise in y_n as ϵ_n and let $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^T$, and assume X^TX is invertible. Follow the steps to show the expected in sample error of linear regression w.r.t. D is given by

$$\mathbb{E}_D[E_{in}(\mathbf{w}_{lin})] = \sigma^2(1 - \frac{d+1}{N}).$$

(a) *Show that the in sample estimate of* $\mathbf{y}$ *is given by* $\hat{\mathbf{y}} = X\mathbf{w}^* + H\epsilon$.

Since $\hat{\mathbf{y}} = H\mathbf{y}$, we can use the definition of H and of $\mathbf{y}$ to arrive at this result.

$$\begin{aligned}\hat{\mathbf{y}} &= [X(X^TX)^{-1}X^T](X\mathbf{w}^* + \epsilon) \\ &= X[(x^TX)^{-1}(X^TX)]\mathbf{w}^* + H\epsilon \\ &= X\mathbf{w}^* + H\epsilon\end{aligned}$$

(b) *Show the in sample error vector* $\hat{\mathbf{y}} - \mathbf{y}$ *can be expressed by a matrix times* ϵ. *What is the matrix?*

$$\begin{aligned}\hat{\mathbf{y}} - \mathbf{y} &= X\mathbf{w}^* + H\epsilon - (X\mathbf{w}^* + \epsilon) \\ &= H\epsilon - \epsilon = (H - I)\epsilon\end{aligned}$$

So obviously the matrix in question is $(H - I)$.

(c) *Express* $E_{in}(\mathbf{w}_{lin})$ *in terms of* ϵ *using (b), and simplify the expression using Exercise 3.3(c).*

$$\begin{aligned}E_{in}(\mathbf{w}_{lin}) &= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2 \\ &= \frac{1}{N}[\epsilon^T(H - I)^T][(H - I)\epsilon] \\ &= \frac{1}{N}\epsilon^T(H - I)^2\epsilon \\ &= \frac{1}{N}\epsilon^T(I - H)\epsilon \\ &= \frac{1}{N}(\epsilon^T\epsilon - \epsilon^TH\epsilon)\end{aligned}$$

where we have used the fact that $(I - H)^k = (I - H)$.

(d) *Prove the original equality we sought.*

$$\begin{aligned}
\mathbb{E}_D[E_{in}] &= \mathbb{E}_D[\frac{1}{N}\epsilon^T\epsilon] - \mathbb{E}_D[\frac{1}{N}\epsilon^T H\epsilon] \\
&= \frac{N\sigma^2}{N} - \frac{1}{N}\mathbb{E}_D[\sum_i H_{ii}\epsilon_i^2] - \frac{1}{N}\mathbb{E}_D[\sum_{i,j} H_{ij}\epsilon_i\epsilon_j] \\
&= \sigma^2 - \frac{1}{N}\operatorname{Tr} H\epsilon_i^2 - \frac{1}{N}(0) \\
&= \sigma^2 - \frac{1}{N}(d+1) \\
&= \sigma^2(1 - \frac{d+1}{N})
\end{aligned}$$

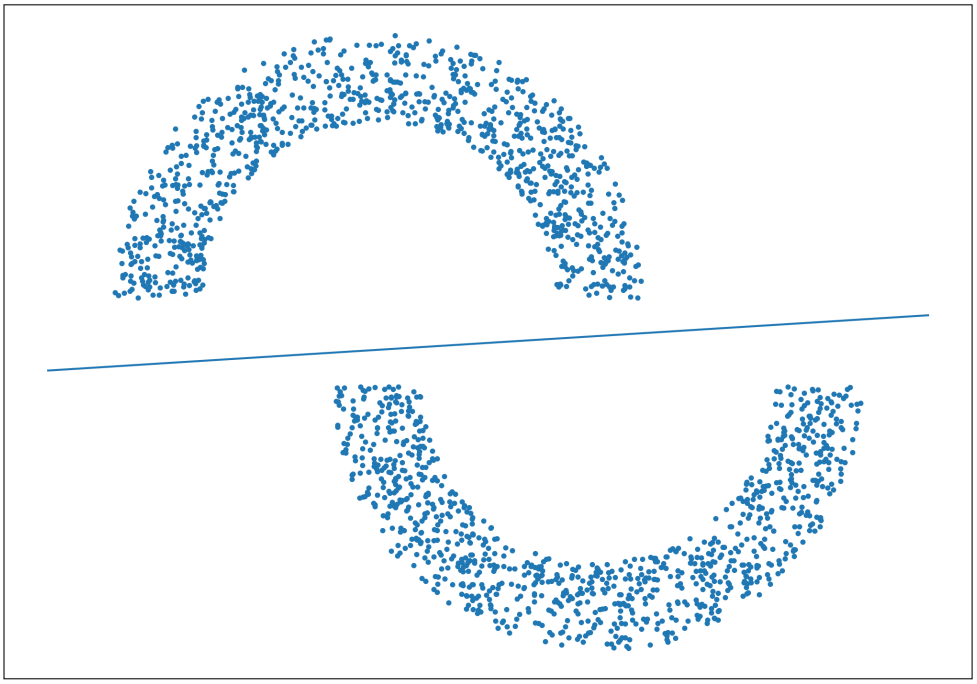
(e) As the noise term is a random variable and likely follows a normal distribution, the expectation of the noise term is again 0. Since the rest of the test data comes from the same target function, we end up with the same expected error. That is,

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

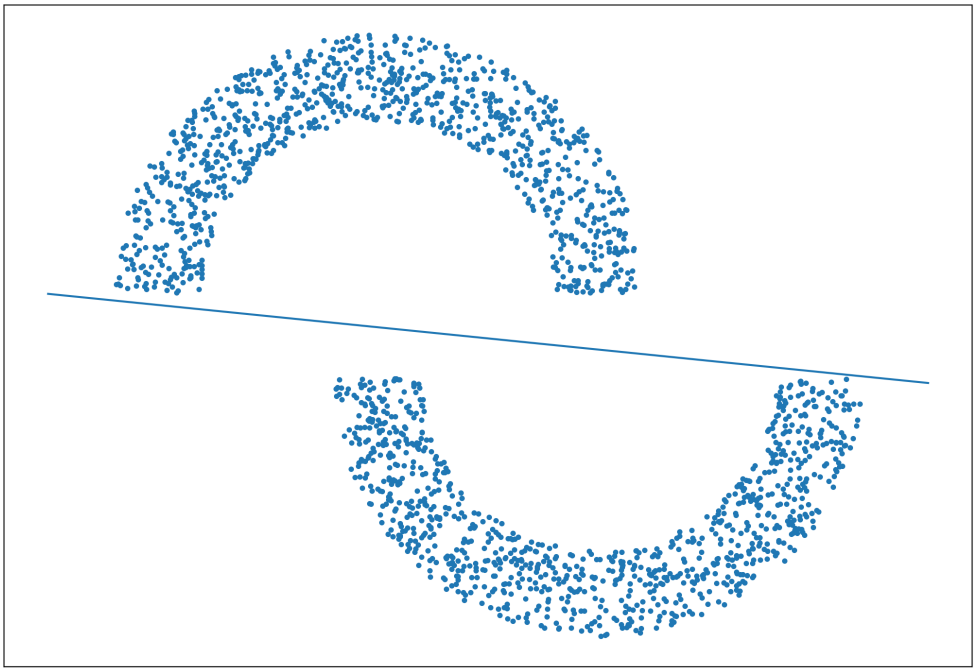
Problem 3.1

Use 2000 uniformly generated samples for the two semicircle region for this problem.

(a) *Run the PLA starting from* $\mathbf{w} = \mathbf{0}$ *until convergence, and plot.*



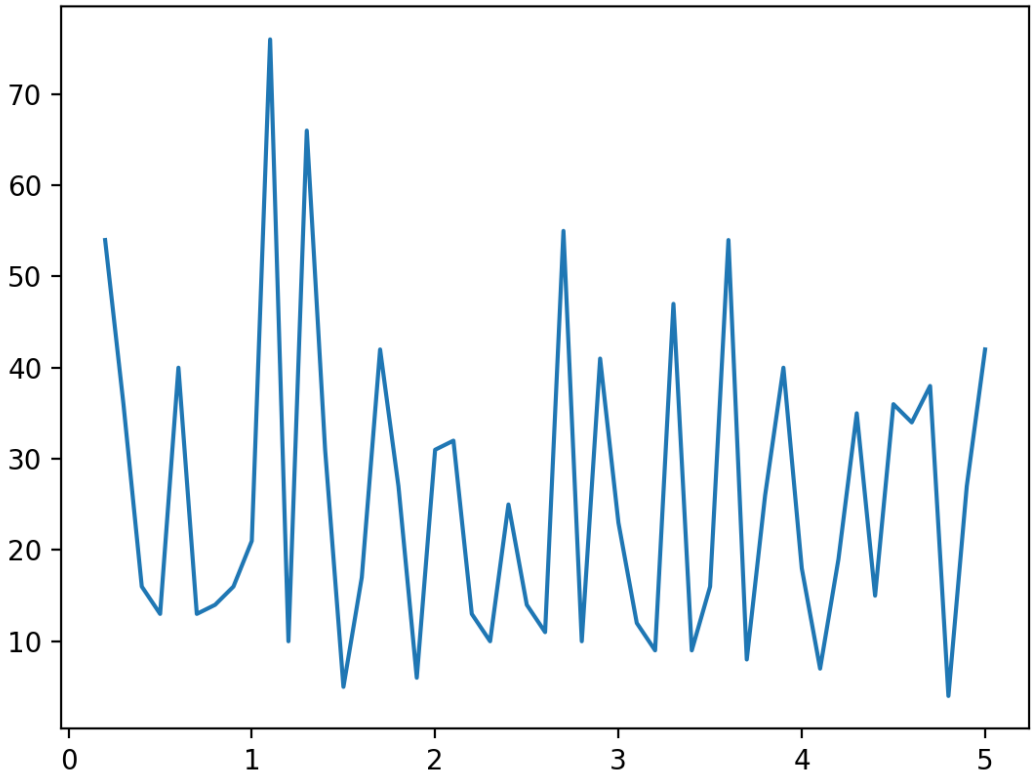
(b) *Repeat (a) using linear regression and explain.*



Since the data is linearly separable, linear regression easily separates the data, just as PLA was able to do. However, this is closer to an "optimum" linear hypothesis, which explains the difference in slope. This regression produced a negative slope, corresponding to the densities of data points. PLA was not able to do this because it simply adjusts based on what data is misclassified on each iteration.

Problem 3.2

Vary the separation and plot sep vs. the number of iterations of PLA. Explain.



Convergence of the PLA is not really dependent on the separation so much as the magnitude of $\mathbf{x_n} \in X$. So it is unsurprising that there is little variation in the number of iterations required for convergence. This test would be better if it was done for each sep value several times, but it is evident nonetheless that there is not much difference in each sep, save for the small separations, which are slightly larger in value.

Problem 3.8

Show that among all hypotheses, the one that minimizes the least squares E_{out} is

$$h^*(x) = \mathbb{E}[y|x]$$

We can rewrite the error function and expand from there as follows:

$$\begin{aligned} E_{out}(h) &= \mathbb{E}[(h(\mathbf{x}) - y)^2] \\ &= \mathbb{E}[((h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]) + (\mathbb{E}[y|\mathbf{x}] - y))^2] \\ &= \mathbb{E}[(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])^2 + 2(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y) + (\mathbb{E}[y|\mathbf{x}] - y)^2] \end{aligned}$$

Now, $\mathbb{E}[2(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])(\mathbb{E}[y|\mathbf{x}] - y)] = 0$ and $\mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] = 0$, so we finally have

$$E_{out}(h) = \mathbb{E}[(h(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}])^2]$$

Which is obviously minimized when $h(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ since in this case $E_{out} = 0$, the answer we desired for $h^*(\mathbf{x})$.

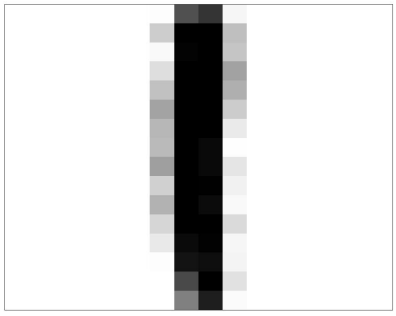
Now, we can write y as $y = h^*(\mathbf{x}) + \epsilon(\mathbf{x})$ which can be rewritten as

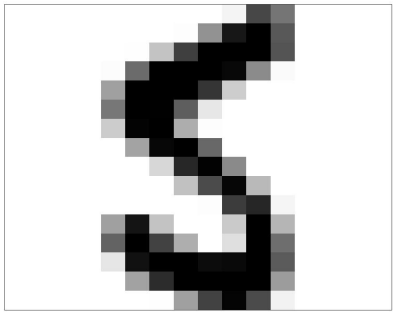
$$y = \mathbb{E}[y|\mathbf{x}] + \epsilon(\mathbf{x}).$$

So in order to keep our desired E_{out}, we must have $\mathbb{E}[\epsilon(\mathbf{x})] = 0$.

Handwritten Digits Data

(a) *Plot two of the digit images*





(b) *Develop two features to measure properties of teh image that would be useful in distinguishing between 1 and 5, e.g. symmetry and average intensity. Give the mathematical definition of the two features.*

Average intensity and symmetry over the center vertical axis will be the two features used.

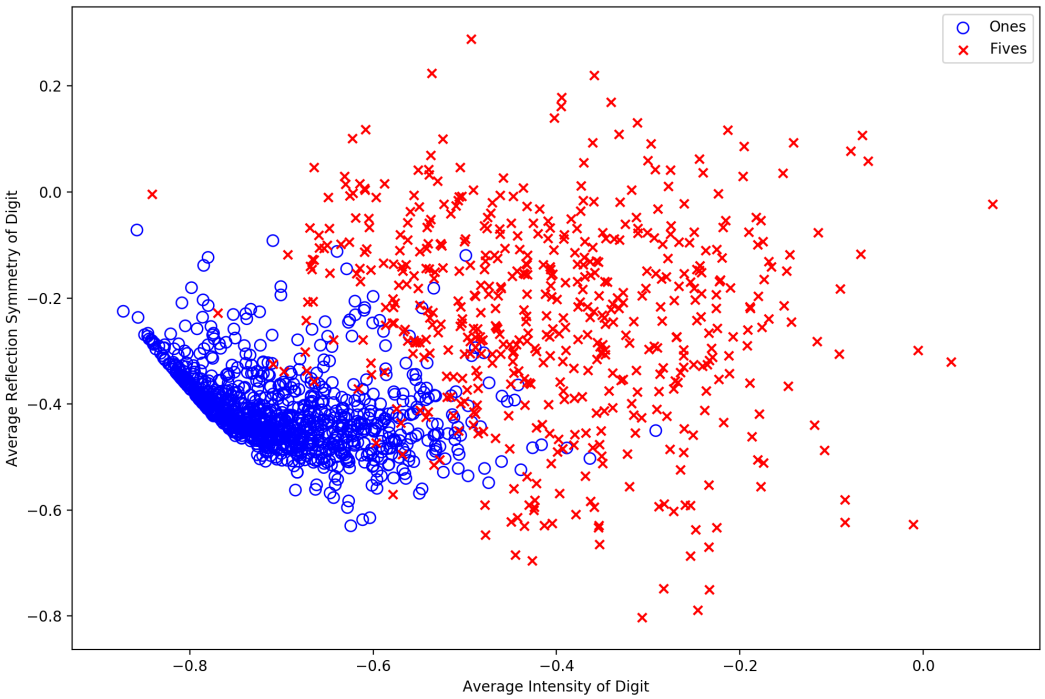
Average intensity is given by the average value of 256 pixels. Defining pixels as p_i for $i = 0, \ldots, 255$, we have for the definition

$$Intensity = \frac{1}{256} \sum_{i=0}^{255} p_i$$

For symmetry we will instead define the digit image as a 16×16 matrix $P_{16\times16}$ where the rows of the matrix correspond to the pixel values of the rows of the image. Thus, we can define the reflection symmetry as an average after subtracting the right half of the image from the left half:

$$Symmetry = \frac{1}{128} \sum_{i=0,j=0}^{i=15,j=7} (P_{ij} - P_{i,15-j})$$

(c) *As in the text, give a 2D scatter plot of the features. For each data example, plot the two features with a red 'x' if it is a 5 and a blue 'o' if it is a 1.*


Ones
Fives
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Average Reflection Symmetry of Digit
−0.8
−0.6
−0.4
−0.2
0.0
Average Intensity of Digit