Homework 6

Runmin Lu

October 17, 2021

Exercise 3.4

(a)

$$\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\mathbf{w}_{\text{lin}} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}) \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}^* + \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} \\
&= \mathbf{X}\mathbf{w}^* + \mathbf{H}\boldsymbol{\epsilon}
\end{aligned}$$

(b)

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

(c)

$$
\begin{aligned}
E_{\text{in}}(\mathbf{w}_{\text{lin}}) &= \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \\
&= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2 \\
&= \frac{1}{N}||(\mathbf{H} - \mathbf{I})\boldsymbol{\epsilon}||^2 \\
&= \frac{1}{N}||(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}||^2 \\
&= \frac{1}{N}\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon} \\
&= \frac{1}{N}\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H})^2\boldsymbol{\epsilon} \\
&= \frac{1}{N}\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}
\end{aligned}
$$

(d)

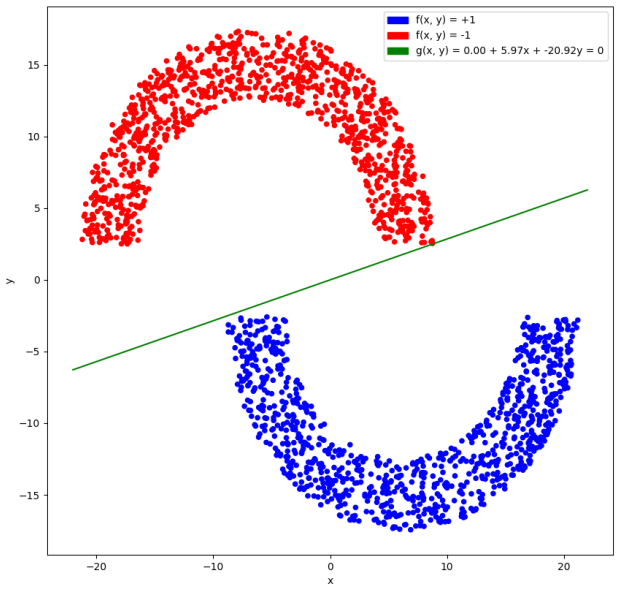
$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] &= \mathbb{E}_{\mathcal{D}}[\frac{1}{N}\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon}] \\
&= \frac{1}{N}(\mathbb{E}_{\mathcal{D}}[||\boldsymbol{\epsilon}||^2] - \mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon}]) \\
\mathbb{E}_{\mathcal{D}}[||\boldsymbol{\epsilon}||^2] &= \sum_{i=1}^{N}\mathbb{E}_{\mathcal{D}}[\epsilon_i^2] \\
&= N\sigma^2 \\
\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon} &= \sum_{i=1}^{N}\sum_{j=1}^{N}H_{ij}\epsilon_i\epsilon_j \\
\mathbb{E}_{\mathcal{D}}[\epsilon_i\epsilon_j] &= \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases} \qquad (\epsilon_i, \epsilon_j \text{ are independent}) \\
\mathbb{E}_{\mathcal{D}}[\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon}] &= \sum_{i=1}^{N}H_{ii}\sigma^2 \\
&= \sigma^2\text{trace}(\mathbf{H}) \\
&= \sigma^2(d+1) \\
\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{lin}})] &= \frac{1}{N}(N\sigma^2 - \sigma^2(d+1)) \\
&= \sigma^2(1 - \frac{d+1}{N})
\end{aligned}
$$

(e)

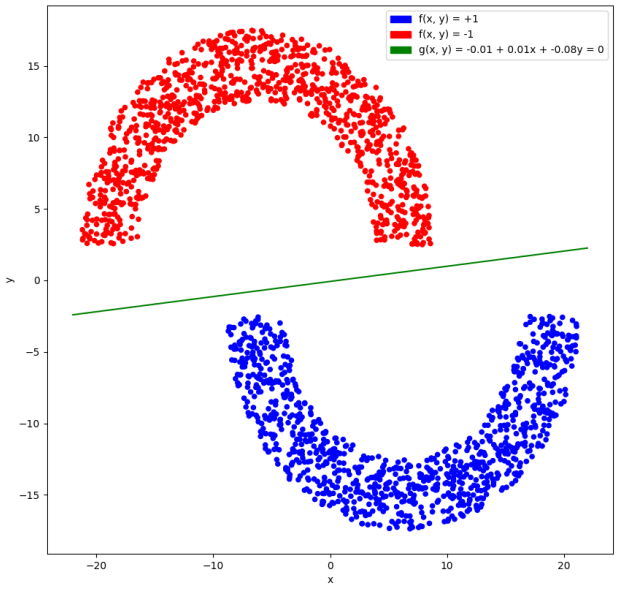
$$
\begin{aligned}
\mathbf{y}' &= \mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}' \\
\hat{\mathbf{y}} - \mathbf{y}' &= \mathbf{X}\mathbf{w}^* + \mathbf{H}\boldsymbol{\epsilon} - (\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}') \\
&= \mathbf{H}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}' \\
E_{\text{test}}(\mathbf{w}_{\text{lin}}) &= \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}'||^2 \\
&= \frac{1}{N}||\mathbf{H}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'||^2 \\
&= \frac{1}{N}(\boldsymbol{\epsilon}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}'^T\mathbf{H}\boldsymbol{\epsilon} + ||\boldsymbol{\epsilon}'||^2) \\
&= \frac{1}{N}(\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}'^T\mathbf{H}\boldsymbol{\epsilon} + ||\boldsymbol{\epsilon}'||^2) \qquad \text{(from Exercise 3.3 (b))} \\
\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[E_{\text{test}}(\mathbf{w}_{\text{lin}})] &= \frac{1}{N}(\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon}] - \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[2\boldsymbol{\epsilon}'^T\mathbf{H}\boldsymbol{\epsilon}] + \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[||\boldsymbol{\epsilon}'||^2]) \\
\boldsymbol{\epsilon}'^T\mathbf{H}\boldsymbol{\epsilon} &= \sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}\epsilon_i'\epsilon_j \\
\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\epsilon_i'\epsilon_j] &= 0 \qquad (\epsilon_i', \epsilon_j \text{ are independent}) \\
\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}'^T\mathbf{H}\boldsymbol{\epsilon}] &= 0 \\
\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[E_{\text{test}}(\mathbf{w}_{\text{lin}})] &= \frac{1}{N}(\mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[\boldsymbol{\epsilon}^T\mathbf{H}\boldsymbol{\epsilon}] + \mathbb{E}_{\mathcal{D},\boldsymbol{\epsilon}'}[||\boldsymbol{\epsilon}'||^2] \\
&= \frac{1}{N}(\sigma^2(d+1) + N\sigma^2) \\
&= \sigma^2(1 + \frac{d+1}{N})
\end{aligned}
$$

Problem 3.1

(a)



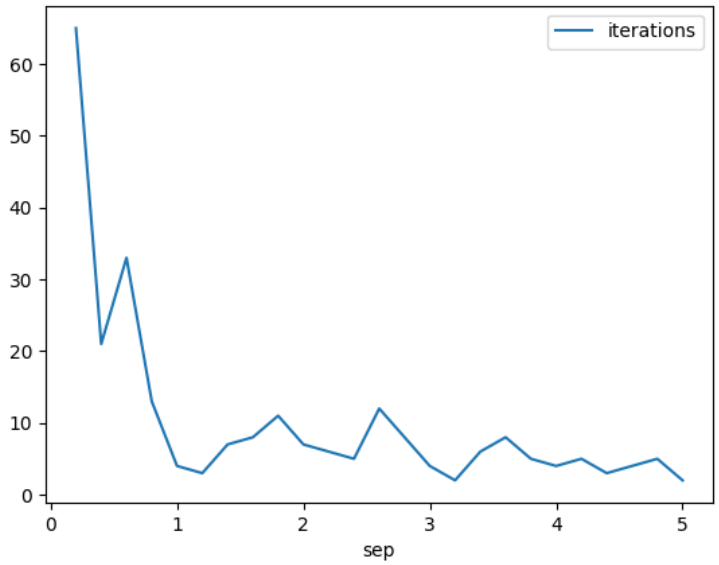
(b)



If you run this perceptron and regression multiple times, the weight vector generated by regression is a lot

more consistent than the one for perceptron. The reason behind this is that regression is deterministic: $\mathbf{w}_{\text{lin}}$ only depends on the data points. On the other hand, $\mathbf{w}$ generated by preceptron also depends on the order we update it. In each iteration, we can have multiple options of misclassified points to update the weight with, which can result in many possible subsequent weight.

Problem 3.2



As *sep* increases, the number of iterations it takes to converge generally decreases. From Problem 1.3, we know that $t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}$ where $\rho = \min_{1 \leq n \leq N} y_n(\mathbf{w}^{*T}\mathbf{x}_n)$. As *sep* increases, all the $\mathbf{x}_n$'s are farther from the origin, which means that their norms increase, $\mathbf{w}^{*T}\mathbf{x}_n$ also increases in absolute value, *rho* increases, and the upper bound for t decreases since ρ^2 is in the denominator.

Problem 3.8

Suppose for contradiction that there's some $h'(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] + \epsilon(\mathbf{x})$ that gives an even smaller E_{out}, where $\epsilon(\mathbf{x})$ is some aribtrary deterministic function of $\mathbf{x}$.

Then

$$\begin{aligned}
E_{\text{out}}(h') - E_{\text{out}}(h^*) &= \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] + \epsilon(\mathbf{x}) - y)^2] - \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] \\
&= \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2 + 2\epsilon(\mathbf{x})(\mathbb{E}[y|\mathbf{x}] - y) + \epsilon(\mathbf{x})^2]\mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] \\
&= \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] + \mathbb{E}[2\epsilon(\mathbf{x})(\mathbb{E}[y|\mathbf{x}] - y)] + \mathbb{E}[\epsilon(\mathbf{x})^2] - \mathbb{E}[(\mathbb{E}[y|\mathbf{x}] - y)^2] \\
&= \mathbb{E}[2\epsilon(\mathbf{x})(\mathbb{E}[y|\mathbf{x}] - y)] + \mathbb{E}[\epsilon(\mathbf{x})^2] \\
&= 2\epsilon(\mathbf{x})\mathbb{E}[\mathbb{E}[y|\mathbf{x}] - y] + \mathbb{E}[\epsilon(\mathbf{x})^2] \qquad \text{because } \epsilon(\mathbf{x}) \text{ is deterministic} \\
&= \mathbb{E}[\epsilon(\mathbf{x})^2] \qquad \text{because } y \text{ is expected to equal } E[y|\mathbf{x}] \text{ so } \mathbb{E}[\mathbb{E}[y|\mathbf{x}] - y] = 0 \\
&\geq 0
\end{aligned}$$

$E_{\text{out}}(h')$ is always at least $E_{\text{out}}(h^*)$ so h^* gives the minimum E_{out}.

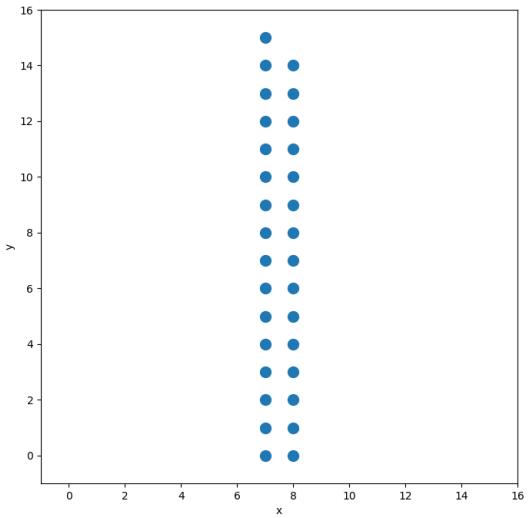
$$\begin{aligned}
E_{\text{out}}(h^*) &= \mathbb{E}[(h^*(\mathbf{x}) - y)^2] \\
&= \mathbb{E}[(h^*(\mathbf{x}) - (h^*(\mathbf{x}) + \epsilon(\mathbf{x})))^2] \\
&= \mathbb{E}[\epsilon(\mathbf{x})^2]
\end{aligned}$$

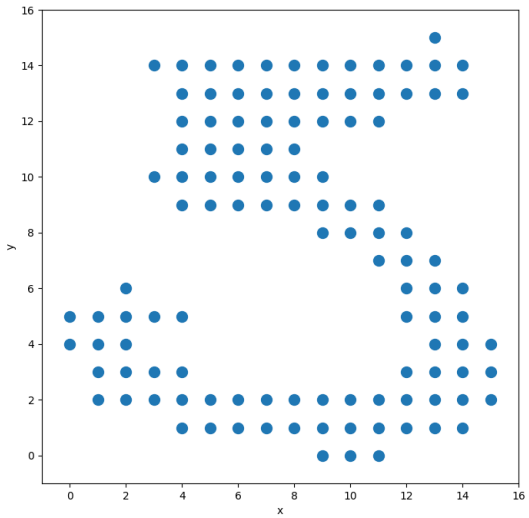
To minimize E_{out}, we need

$$\begin{aligned}
\frac{dE_{\text{out}}}{d\epsilon} &= 0 \\
\mathbb{E}[2\epsilon(\mathbf{x})] &= 0 \\
E[\epsilon(\mathbf{x})] &= 0
\end{aligned}$$

6

(a)





(b)

The first feature is bounding box area A: the area of the smallest axis rectangle that contains all the non-white pixels.

Suppose the set of non-white pixels is $\{(x_1, y_1), ..., (x_n, y_n)\}$.

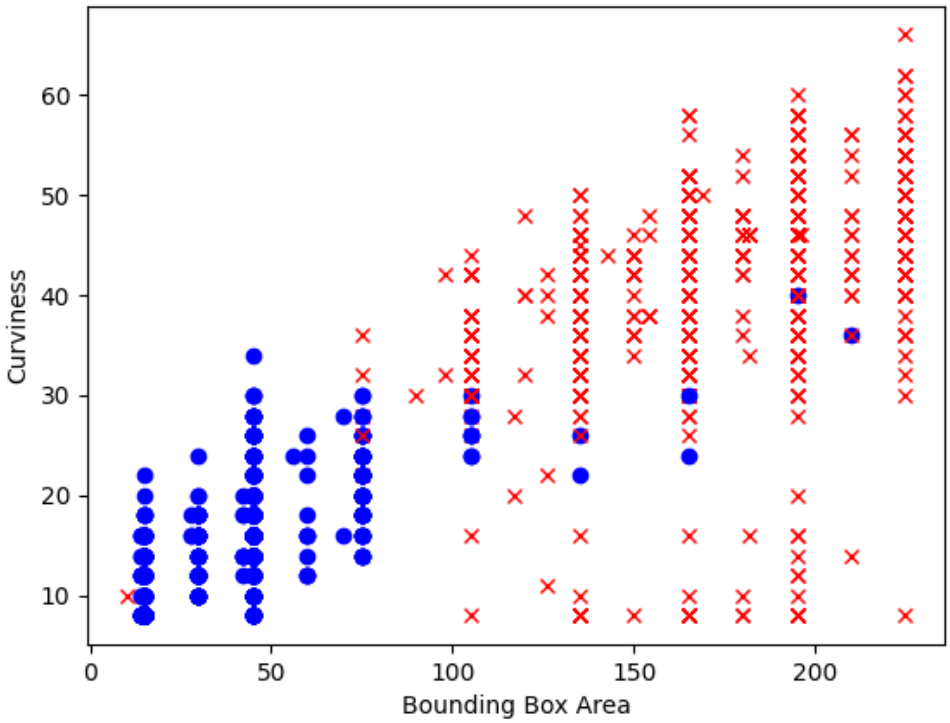
Then $A = (\max_i x_i - \min_i x_i)(\max_i y_i - \min_i y_i)$

The second feature is curviness, which is measured by the following algorithm.

Loop over all the boundary non-white pixels and compute the angle formed by each triple (prev, curr, next) and take the absolute value of difference of that and π (straight means no curve). In practice, we take $\frac{\pi}{4}$ as 1 unit of angle and it doesn't affect the result of learning because scaling is linear.

Sum up all of them to generate the curviness.

(c)



Note: I don't know why but when I give a marker of 'o' it gets filled.