NYU CS-GY 6923: Machine Learning Fall 2025

Homework 1

Name: Ethan Nie

Problem 1

Collaborators: None

(a) For arbitrary β_j:

$$\begin{aligned}\frac{\partial L_{WSS}}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}[w_i \cdot (\langle \beta, x_i \rangle - y_i)^2] \\ &= \sum_{i=1}^{n} w_i \cdot 2(\langle \beta, x_i \rangle - y_i) \cdot \frac{\partial}{\partial \beta_j}(\langle \beta, x_i \rangle - y_i)\end{aligned}$$

Since

$$\langle \beta, x_i \rangle = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_j x_{ij} + \beta_m x_{im}$$

and the only term related to β_j is $\beta_j x_{ij}$,
thus

$$\frac{\partial}{\partial \beta_j}(\langle \beta, x_i \rangle - y_i) = x_{ij}$$

So,

$$\frac{\partial L_{WSS}}{\partial \beta_j} = 2\sum_{i=1}^{n} w_i \cdot (\langle \beta, x_i \rangle - y_i) \cdot x_{ij}$$

(b) Let $X = [x_1^T; ...; x_n^T], y = (y_1, ..., y_n)^T, W = diag(w_1, ..., w_n)$,

$$\begin{aligned}L(\beta) &= \sum_{i=1}^{n} w_i \cdot (\langle \beta, x_i \rangle - y_i)^2 \\ &= \sum_{i=1}^{n} w_i \cdot (\beta^T x_i - y_i)^2 \\ &= (y - X\beta)^T W (y - X\beta)\end{aligned}$$

$$\nabla_\beta L(\beta) = -2X^T W y + 2X^T W X \beta$$

Let $\nabla_\beta L(\beta) = 0$, $\beta^* = (X^T W X)^{-1} X^T W y$
(c)When $r_i \geq 0$,

$$c_+ r_i = (c_+ + c_-) \cdot \tau r_i$$

When $r_i \leq 0$,

$$-c_- r_i = (c_+ + c_-) \cdot (\tau - 1) r_i$$

Thus,

$$\tau = \frac{c_+}{c_+ + c_-}$$

τ is the proportion of the total cost that is attributed to overestimation.
When $c_+ >> c_-$, $\tau \approx 1$. The model will consistently underestimate to avoid the huge penalty from overestimation.
When $c_- >> c_+$, $\tau \approx 0$. The model will consistently overestimate to avoid the huge penalty of underestimation.
When $c_+ = c_-$, $\tau = 0.5$. The model will tend to predict the median due to the same penalty for overestimation and underestimation.

Problem 2

Collaborators: None

(a)

$$\frac{\partial L(m)}{\partial m} = \sum_{i=1}^{n} 2(m - y_i)$$
$$= 2(mn - \sum_{i=1}^{n} y_i)$$

Let $\frac{\partial L(m)}{\partial m} = 0$,

$$m = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

(b)

$$L(m) = max_i|y_i - m|$$
$$= max(m - min_i y_i, max_i y_i - m)$$

To minimize L(m), let $m - min_i y_i = max_i y_i - m$, $m = \frac{max_i y_i + min_i y_i}{2}$

(c)Let $N_<(m)$ be the number of $y_i < m$, $N_>(m)$ be the number of $y_i > m$. The total change in L(m) when increasing m by δ is

$$\Delta L = (N_>(m) - N_<(m)) \cdot \delta$$

Thus, to minimize L(m), we should let $N_<(m) = N_>(m)$. That is, set m to the median of the data.

(d)To minimize the loss, $m = \tau$. When $\tau = 0.5$, setting m to the median minimizes the loss as proven. When $\tau > 0.5$, the penalty for overestimation is heavier, leading to a larger m to minimize the loss. When $\tau < 0.5$, the penalty for underestimation is heavier, leading to a smaller m to minimize the loss.

Problem 3

Collaborators: None

(a) Because

$$a_1 + s_1\lambda = a_2 + s_2\lambda$$

So

$$a_2 = a_1 + s_1\lambda - s_2\lambda$$

Substitute the a2 in the model:

$$f(x_i) = \begin{cases} a_1 + s_1 x_i & x_i < \lambda \\ a_1 + s_1\lambda - s_2\lambda + s_2 x_i & x_i \geq \lambda \end{cases}$$

(b)For all i, define $z_{i1} = 1, z_{i2} = x_i, z_{i3} = max\{x_i - \lambda, 0\}$.
Let $f(x_i) = \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3}$, $X_i = [z_{i1}, z_{i2}, z_{i3}]$,
then

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

if $x_i < \lambda$, $f(x_i) = \beta_1 + \beta_2 x_i$, thus $\beta_1 = a_1, \beta_2 = s_1$;
if $x_i \geq \lambda$, $f(x_i) = \beta_1 + \beta_2 x_i + \beta_3(x_i - \lambda)$, thus $\beta_3 = s_2 - s_1$;
So,

$$a_1 = \beta_1, s_1 = \beta_2, s_2 = \beta_2 + \beta_3$$

(c)

Problem 4

Collaborators: None

(a) Let original model be $\hat{y} = \beta_0 + \beta_1 x$.
Let mean centering model be

$$\begin{aligned}\hat{y}' &= \beta_0' + \beta_1'(x - \bar{x}) \\ &= (\beta_0' - \beta_1'\bar{x}) + \beta_1' x\end{aligned}$$

It is obvious that $\hat{y}$ and $\hat{y}'$ are actually equivalent, because the equation for the new model can be algebraically rearranged into the exact same functional form as the original model. In this case, $\beta_0' = \beta_0 + \beta_1\bar{x}, \beta_1' = \beta_1$.

(b) Let standard deviation model be

$$\begin{aligned}\hat{y}'' &= \beta_0'' + \beta_1''\frac{x}{\sigma} \\ &= \beta_0'' + \frac{\beta_1''}{\sigma}x\end{aligned}$$

It is obvious that $\hat{y}$ and $\hat{y}''$ are actually equivalent, because the equation for the new model can be algebraically rearranged into the exact same functional form as the original model. In this case, $\beta_0'' = \beta_0, \beta_1'' = \sigma\beta_1$.

(c) My answers will not change for l_1 loss and l_∞ loss. Both mean-centering and normalization do not alter the column space of the data matrix. The column space represents the entire set of possible prediction vectors the model can generate.