Yared Taye

Yash Yadati

Lokesh Gangaramaney

Deep Learning CS – 541

HOMEWORK 2

$$\nabla_{w_1} J(\theta) = \frac{n}{4}[4w_1 + 2w_2 + 4b - 2] = 0$$

$$\hookrightarrow 4w_1 + 2w_2 = 2 - 4b \quad \text{——} \ (1)$$

$$\nabla_{w_2} J(\theta) = \frac{n}{4}[4w_2 + 2w_1 + 4b - 2] = 0$$

$$\hookrightarrow 4w_2 + 2w_1 = 2 - 4b \quad \text{——} \ (2)$$

solving (1) & (2)

$$\begin{aligned} \hookrightarrow \quad 4w_1 + 2w_2 &= 2 - \cancel{4b} \\ -\quad 2w_1 + 4w_2 &= \cancel{2 - 4b} \\ \hline 2w_1 - 2w_2 &= 0 \end{aligned}$$

$$\Rightarrow 2w_1 = 2w_2 \Rightarrow \boxed{w_1 = w_2} \ \text{—} (*)$$

Plug in w_1 for w_2 (use $*$) : $4w_1 + 2w_1 = 2 - 4b$

$$\hookrightarrow 6w_1 + 4b - 2 = 0$$

$$\hookrightarrow \text{root} : \boxed{w_1 = \pm \frac{1}{3}(1 - 2b) = w_2} \ \text{—} (A)$$

$$\nabla_b J(\theta) = \frac{n}{\cancel{4}}[\cancel{8}^{2}b - \cancel{4}^{1} + \cancel{4}^{1}(w_1 + w_2)] = 0$$

$$\Rightarrow 2b - 1 + w_1 + w_2 = 0$$

$$\Rightarrow 2b - 1 + 2w_1 = 0 \quad \left(\text{use } +\frac{1}{3}(1-2b)\right) \text{ from } (A)$$

$\left(-\frac{1}{3}(1-2b)\right)$ will also give same result

$$\Rightarrow 2b - 1 + \frac{2}{3}(1 - 2b) = 0$$

$$\Rightarrow -\frac{1}{3} + \left(\frac{-4+6}{3}\right)b = 0$$

$$\Rightarrow 2b = 1 \quad \therefore \boxed{b = 1/2}$$

plug into (A)

$$\text{then } w_1 = \pm \frac{1}{3}\left(1 - \cancel{2} \cdot \frac{1}{\cancel{2}}\right) = 0$$

$$\therefore \boxed{w_1 = w_2 = 0}$$

PROBLEM 2

For problem 2, we used the following values to optimize the hyperparameters using a grid search. As per instruction, we used 4 sets of values for each hyperparameter to tune the model.

Learning Rate = [0.0001, 0.001, 0.005, 0.01]

Mini-Batch Sizes = [10, 50, 100, 200]

L2 Regularization Strength = [0.02, 0.05, 0.07, 0.1]

Number of Epochs = [5, 10, 50, 100]

Best Results Hyperparameters

Learning Rate = 0.001

Batches = 10

Epochs = 100

L-2 Regularization Strength = 0.02

MSE on Training Set = 71.98839593074759

MSE on Testing Set = 87.44495379155147

$\boxed{\#3}$ Regularization Cost $= \frac{\alpha}{2} \omega^T S \omega$

$\hookrightarrow$ let $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$

$\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}$

$$J_r = \frac{\alpha}{2} \underbrace{[\omega_1 \ \omega_2]}_{} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} \quad (1\times 2,\ 2\times 2,\ 2\times 1)$$

$$= \frac{\alpha}{2}\left([\omega_1 \ \omega_2] \begin{bmatrix} S_{11}\omega_1 + S_{12}\omega_2 \\ S_{21}\omega_1 + S_{22}\omega_2 \end{bmatrix} \right) \quad (1\times 2,\ 2\times 1 = 1\times 1)$$

$$= \frac{\alpha}{2}\left(S_{11}\omega_1^2 + S_{12}\omega_1\omega_2 + S_{21}\omega_1\omega_2 + S_{22}\omega_2^2 \right)$$

In order to discourage weights from becoming too asymmetric, we want $\frac{\alpha}{2}(\omega_1 - \omega_2)^2 = 0 \leftarrow$ Will maximize cost of the difference in ω_1 & ω_2.

$$\therefore \cancel{\frac{\alpha}{2}}(\omega_1 - \omega_2)^2 = \cancel{\frac{\alpha}{2}}\left(S_{11}\omega_1^2 + S_{22}\omega_2^2 + \omega_1\omega_2(S_{12} + S_{21})\right)$$

$$\omega_1^2 + \omega_2^2 - 2\omega_1\omega_2 = S_{11}\omega_1^2 + S_{22}\omega_2^2 + \omega_1\omega_2(S_{12} + S_{21})$$

$\hookrightarrow$ $S_{11} = 1$, $S_{22} = 1$ & $S_{12} + S_{21} = -2$ (Assuming whole numbers that are not zero)

$$S_{12} = S_{21} = -1$$

$$\therefore S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

(Q4) We are given that:

(a) $P(x_t | x_1, \ldots\ldots, x_{t-1}) = P(x_t | x_{t-1})$

(b) $P(y_t | x_t, y_1, \ldots, y_{t-1}) = P(y_t | x_t)$

Bayes theorem,

$$P(a|b,c) = \frac{P(b|a,c)\, P(a|c)}{P(b|c)}$$

Or

$$P(a|b,c) \propto P(b|a,c)\, P(a|c)$$

Using $a = x_t$, $b = y_t$, $c = y_1 \ldots\ldots y_{t-1}$

$$P(x_t | y_1, \ldots y_t) \propto P(y_t | x_t, y_1 \ldots y_{t-1})^* P(x_t | y_1 \ldots y_{t-1})$$

$$P(x_t | y_t, y_1 \ldots y_{t-1}) = P(y_t | x_t)^* P(x_t | y_1 \ldots y_{t-1})$$

Using (a), $P(x_t | x_1, \ldots, x_{t-1}) = P(x_t | x_{t-1})$, we get:

$$P(x_t | y_1, \ldots, y_t) = P(y_t | x_{t-1}) \sum_{x_{t-1}} P(x_t | x_{t-1}) P(x_{t-1} | y_1, \ldots, y_{t-1})$$

$\boxed{\#5}$ $P(y|x, w, \sigma^2) = \mathcal{N}(y, x^T w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - x^T w)^2}{2\sigma^2}\right)$

Using collected $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, MLE for w & σ^2?

$P(D|w, \sigma^2) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}, w, \sigma^2)$ → Conditional Independence

taking log

$$\log P(D|w, \sigma^2) = \log \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log P(y^{(i)}|x^{(i)}, w, \sigma^2)$$

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - x^T w)^2}{2\sigma^2}\right)\right]$$

→ $\log(XY)$

$$= \sum_{i=1}^{n} \left[\log\left(\exp\left(\frac{-(y - x^T w)^2}{2\sigma^2}\right)\right) + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right]$$

$(2\pi\sigma^2)^{-1/2}$

$$= \sum_{i=1}^{n} \left[\frac{-(y - x^T w)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

$$= \sum_{i=1}^{n} \left[\frac{-(y - x^T w)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2)\right]$$

$$\therefore \boxed{\log P(D|w, \sigma^2) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y - x^T w)^2 - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)}$$

$$\rightarrow \frac{d}{d\sigma^2} \log P(D|w,\sigma^2) = \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y - x^T w)^2 - \frac{n}{2(\sigma^2)} = 0$$

$$= \cancel{\frac{-n}{2(\sigma^2)^2}} \left[-\frac{1}{n} \sum_{i=1}^{n} (y - x^T w)^2 + \sigma^2 \right] = 0$$

$$\sigma^2 = +\frac{1}{n} \sum_{i=1}^{n} (y - x^T w)^2$$

$\hookrightarrow (a-b)^2 = (b-a)^2$

$$\therefore \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x^T w - y)^2$$

$$\boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)T} w - y^{(i)}\right)^2}$$

expanded $(y - x^T w)^2$

$$\rightarrow \frac{d}{dw} \log P(D|w,\sigma^2) = \cancel{\frac{-1}{2(\sigma^2)}} \sum_{i=1}^{n} (y^2 + (x^T w)^2 - 2x^T w y) = 0$$

$$= \sum_{i=1}^{n} 2(x^T w) x^T - 2x^T y = 0$$

$$= \cancel{2 \sum_{i=1}^{n} x^T} \sum_{i=1}^{n} (w x^T - y) = 0$$

$$\Rightarrow w \sum_{i=1}^{n} x^T = \sum_{i=1}^{n} y$$

$$w = \frac{\sum_{i=1}^{n} y}{\sum_{i=1}^{n} x^T} \times \frac{\sum_{i=1}^{n} x}{\sum_{i=1}^{n} x} \Rightarrow \boxed{w = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} xx^T} = \left(\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right)^{-1} \left(\sum_{i=1}^{n} x^{(i)} y^{(i)}\right)}$$