17CS10058 DL Assignment 1

Gurjot Singh Suri

Ans: 1.1) $v = g(\theta) = 1 - \theta$

$$P(x|v) = \theta^{x}(1-\theta)^{1-x}$$

$$= (1-v)^{x} v^{1-x} \quad \left[\begin{matrix} v = 1-\theta \\ \theta = 1 - v \end{matrix}\right]$$

$$\therefore P(x|v) = (1-v)^{x} v^{1-x} \quad x \in \{0, 1\}$$

Joint probability

$$P(x_1, x_2, \ldots, x_n | v)$$

$$= \prod_{i=1}^{n} P(x_i | v)$$

$$= \prod_{i=1}^{n} (1-v)^{x_i} v^{1-x_i}$$

$$= (1-v)^{\sum_{i=1}^{n} x_i} v^{n - \sum_{i=1}^{n} x_i}$$

Log likelihood

$$\mathcal{L}(v) = \log P(x_1, x_2, \ldots x_n | v)$$

$$= \log (1-v)^{\sum_{i=1}^{n} x_i} v^{n - \sum_{i=1}^{n} x_i}$$

$$\mathcal{L}(v) = \left(\sum_{i=1}^{n} x_i\right)(\log(1-v)) + v\left(n - \sum_{i=1}^{n} x_i\right)$$

For maximising $\mathcal{L}(v)$,

$$\frac{\partial \mathcal{L}(v)}{\partial v} = 0$$

$$\therefore \left(\sum_{i=1}^{n} x_i\right) \times \frac{1}{1-v} \times (-1) + \frac{1}{v} \times \left(n - \sum_{i=1}^{n} x_i\right) = 0$$

$$\therefore \frac{\sum_{i=1}^{n} x_i}{1-v} = \frac{n - \sum_{i=1}^{n} x_i}{v}$$

$$\therefore v \sum_{i=1}^{n} x_i = n - nv + \left(\sum_{i=1}^{n} x_i\right) v - \sum_{i=1}^{n} x_i$$

$$\therefore nv = n - \sum_{i=1}^{n} x_i$$

$$\therefore v = 1 - \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{v_{MLE} = 1 - \frac{\sum_{i=1}^{n} x_i}{n}}$$

This is same as

$$v_{MLE} = g(\theta_{MLE}) = 1 - \theta_{MLE}$$
$$= 1 - \frac{\sum_{i=1}^{n} x_i}{n}$$

Ans: 1.2) Joint distribution

$$P(x_1, x_2, \ldots, x_{2n} \mid \theta)$$

$$= \prod_{i=1}^{2n} P(x_i \mid \theta)$$

$$= \prod_{i=1}^{2n} \left(\frac{1}{2} e^{-|x_i - \theta|}\right)$$

$$= \frac{1}{2^{2n}} e^{-\sum_{i=1}^{2n} |x_i - \theta|}$$

Log-likelihood

$$\mathcal{L}(\theta) = \log P(x_1, x_2, \cdots x_{2n} \mid \theta)$$

$$= \log \frac{1}{2^{2n}} e^{-\sum_{i=1}^{2n} |x_i - \theta|}$$

$$\mathcal{L}(\theta) = -\log 2^{2n} - \sum_{i=1}^{2n} |x_i - \theta|$$

For maximising $L(\theta)$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$$

$$-\sum_{i=1}^{2n} \frac{\partial |x_i - \theta|}{\partial \theta} = 0$$

$$\therefore -\sum_{i=1}^{2n} \operatorname{sign}(x_i - \theta) \times (-1) = 0$$

$$\therefore \sum_{i=1}^{2n} \operatorname{sign}(x_i - \theta) = 0$$

$$\left[\begin{array}{l} \frac{\partial |x|}{\partial x} \\ = \operatorname{sign}(x) \\ = 1 \quad x > 0 \\ = 0 \quad x = 0 \\ = -1 \quad x < 0 \end{array} \right]$$

~~This wi~~ Each

As the sign function takes only +1, -1 and 0 values,

no of terms with value -1 = no of terms with value 1

That is, θ is greater than ~~exactly~~ half of $x_1, x_2, \cdots, x_{2n}$ and less than half of $x_1, x_2, ---, x_{2n}$ Sorted

Let $\boxed{x_1', x_2', -----, x_{2n}'}$ be the

∴ sequence obtained on sorting

$x_1, x_2, ------. x_{2n}$

So, $\sum_{i=1}^{2n} \text{sign}(x_i' - \theta) = 0$

$\theta >$

So, θ is greater than $x_1', x_2', ---x_n'$

$\theta <$

and θ is less than $x_{n+1}', x_{n+2}', ----, x_{2n}'$

So, θ_{MLE} is any value in the interval $[x_n', x_{n+1}']$

One particular value for θ_{MLE} is the $\boxed{\text{median}}$ of $x_1, x_2, ---, x_{2n}$.

One value of θ_{MLE} $= \boxed{\frac{x_n' + x_{n+1}'}{2}}$ where $x_1', x_2', --, x_{2n}'$ are in $\boxed{\text{sorted}}$ order

Also, θ_{MLE} can be anything in the interval $\boxed{[x_n', x_{n+1}']}$

θ_{MLE} Io any $\boxed{\text{median}}$ of $x_1, x_2, ---, x_{2n}$.

Ans. 2)

2.1)

Likelihood of a single sample, i

$$= P(y_i | x_i, \theta, d_1, d_2)$$

$$= Z(d_1, d_2) \frac{e^{d_1(y_i - \theta^T x_i)}}{\left(d_1 e^{2(y_i - \theta^T x_i)} + d_2\right)^{\frac{d_1+d_2}{2}}}$$

Log likelihood of single sample i

$$= \mathcal{L}_i(\theta) = \log(P(y_i | x_i, \theta, d_1, d_2)))$$

$$= \log\left(Z(d_1,d_2) \frac{e^{d_1(y_i - \theta^T x_i)}}{\left(d_1 e^{2(y_i - \theta^T x_i)} + d_2\right)^{\frac{d_1+d_2}{2}}}\right)$$

$$\mathcal{L}_i(\theta) = \log Z(d_1, d_2) + d_1(y_i - \theta^T x_i)$$
$$- * \frac{d_1+d_2}{2} \times \log\left(d_1 e^{2(y_i - \theta^T x_i)} + d_2\right)$$

$$\therefore \frac{\partial \mathcal{L}_i(\theta)}{\partial \theta} = 0 + d_1 \times (-x_i)$$
$$- * \frac{d_1+d_2}{2} \times \frac{1 \times d_1 \times e^{2(y_i - \theta^T x_i)}}{d_1 e^{2(y_i - \theta^T x_i)} + d_2} \times (-2x_i)$$

$$= -d_1 x_i + \frac{(d_1+d_2)\, d_1 x_i\, e^{2(y_i - \theta^T x_i)}}{d_1 e^{2(y_i - \theta^T x_i)} + d_2}$$

$$\therefore \frac{\partial l_i(\theta)}{\partial \theta} = -d_1 x_i + \frac{(d_1+d_2)\, d_1 x_i}{d_1 + d_2\, e^{-2(y_i - \theta^T x_i)}}$$

$$\therefore \boxed{\frac{\partial l_i(\theta)}{\partial \theta} = -d_1 x_i \left(1 - \frac{d_1 + d_2}{d_1 + d_2 \cdot e^{-2(y_i - \theta^T x_i)}}\right)}$$

Ans

Ans: 2.2) From the plots it is seen that the learning rate of ~~7e-2~~ 7e-2 in the first case is high which causes drastic updates at each step and causes divergent behavior and the loss never converges. In the second case, the learning rate of 1e-3 is the best as the loss decreases exponentially and converges swiftly. In the third case, the learning rate of 1e-2 is too low as the loss decreases linearly and the convergence is very slow and the final value of loss is much longer than the second case for the same number of updates. Thus, a learning rate of $1e^{-3}$ is optimal, $7e^{-2}$ is high and $1e^{-6}$ is low.

~~Ans: 2.3)~~ 3)

Ans: 3.1)

Cumulative distribution function $F_{logistic}$

$$= \frac{1}{1 + e^{-(x-\mu)/s}}$$, where μ is mean and s is standard deviation

For our case,

$$\mu = 0, \quad s = \sigma_\epsilon$$

So, $F_{\text{logistic}} = \dfrac{1}{1+e^{-\frac{x}{\sigma_\epsilon}}} = F_{\epsilon_i}$

$$P(y_i = 1 \mid \theta, x_i)$$

$$= P(\theta^T x_i + \epsilon_i \geq 0)$$

$$= 1 - P(\theta^T x_i + \epsilon_i \leq 0)$$

$$= 1 - P(\epsilon_i \leq -\theta^T x_i)$$

$$= 1 - F_{\text{logistic}}(-\theta^T x_i) \quad \left[\begin{array}{c}\text{Definition} \\ \text{of cdf}\end{array}\right]$$

$$= 1 - \frac{1}{1+e^{-\frac{(-\theta^T x_i)}{\sigma_\epsilon}}}$$

$$= 1 - \frac{1}{1+e^{\frac{\theta^T x_i}{\sigma_\epsilon}}}$$

$$= \frac{e^{\frac{\theta^T x_i}{\sigma_\epsilon}}}{1+e^{\frac{\theta^T x_i}{\sigma_\epsilon}}}$$

$$= \frac{1}{1+e^{-\frac{\theta^T x_i}{\sigma_\epsilon}}}$$

$$\therefore \boxed{P(y_i = 1, \theta, x_i) = \text{logistic}\left(\frac{\theta^T x_i}{\sigma_\epsilon}\right)}$$

Ans :- 3.2) Given, $\sigma_\epsilon = 1$

Now,

$$P(y_i = 0 \mid \theta, x_i)$$

$$= 1 - P(y_i = 1 \mid \theta, x_i)$$

$$\boxed{\cancel{\sigma_\epsilon = 1}} \quad = 1 - \text{logistic}\left(\frac{\theta^T x_i}{\cancel{\sigma_\epsilon}\, 1}\right) = 1 - \text{logistic}(\theta^T x_i)$$

y_i takes only binary values, 0 or 1

$$y_i \in [0, 1]$$

Case 1 $\quad y_i = 1$

$$\text{RHS} = \left(\text{logistic}(\theta^T x_i)\right)^1 \times \left(1 - \text{logistic}(\theta^T x_i)\right)^{1-1}$$

$$= \text{logistic}(\theta^T x_i) \times \left(1 - \text{logistic}(\theta^T x_i)\right)^0$$

$$= \text{logistic}(\theta^T x_i)$$

$$= P(y_i = 1 \mid \theta, x_i)$$

$$= \text{LHS}$$

Case 2 $\quad y_i = 0$

$$\text{RHS} = \left(\text{logistic}(\theta^T x_i)\right)^0 \times \left(1 - \text{logistic}(\theta^T x_i)\right)^1$$

$$= 1 - \text{logistic}(\theta^T x_i)$$

$$= P(y_i = 0 | \theta, x_i)$$

$$= LHS$$

For all cases, RHS = LHS

Hence, $P(y_i | \theta, x_i) = \left(\text{logistic}(\theta^T x_i)\right)^{y_i} \left(1 - \text{logistic}(\theta^T x_i)\right)^{(1-y_i)}$

Ans 3.3)

$$\log P(y_i | \theta, x_i)$$

$$= \cancel{y_i} \log \left(\left(\text{logistic}(\theta^T x_i)\right)^{y_i} \left(1 - \text{logistic}(\theta^T x_i)\right)^{1-y_i} \right)$$

$$= y_i \log\left(\text{logistic}(\theta^T x_i)\right) + (1-y_i) \log\left(1 - \text{logistic}(\theta^T x_i)\right)$$

$$= y_i \log\left(\frac{1}{1+\exp(-\theta^T x_i)}\right) + (1-y_i) \log\left(1 - \frac{1}{1+\exp(-\theta^T x_i)}\right)$$

$$= y_i \log\left(\frac{\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\right) + (1-y_i) \log\left(1 - \frac{\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\right)$$

$$= y_i \log\left(\frac{\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\right) + (1-y_i) \log\left(\frac{1}{1+\exp(\theta^T x_i)}\right)$$

$$= y_i \theta^T x_i - y_i \log(1+\exp(\theta^T x_i)) + (1-y_i) \times -\log(1+\exp(\theta^T x_i))$$

$$= y_i \theta^T x_i - y_i \log(1+\exp(\theta^T x_i)) - \log(1+\exp(\theta^T x_i)) + y_i \log(1+\exp(\theta^T x_i))$$

$$\boxed{\log P(y_i \mid \theta, x_i) = y_i \theta^T x_i - \log(1 + \exp(\theta^T x_i))}$$

Ans: 3.4)

$$\ell_{MLE}(\theta) = \log P(y_1, \dots, y_n \mid \theta, x_1, x_2, \dots, x_n)$$

$$= \log P(y \mid X, \theta) = \log \prod_{i=1}^{n} P(y_i \mid \theta, x_i)$$

$$= \sum_{i=1}^{n} \log P(y_i \mid \theta, x_i)$$

$$\ell_{MLE}(\theta) = \sum_{i=1}^{n} \left(y_i \theta^T x_i - \log(1 \mp \exp(\theta^T x_i))\right) \quad ①$$

Now,

$$\theta^T x_i = \sum_{j=1}^{d} \theta_j x_{ij}$$

$$= \sum_{j=1}^{d} x_{ij} \theta_j$$

$$\boxed{\theta^T x_i = (X\theta)_i} \quad \left[\begin{array}{l}\text{Multiplication} \\ \text{definition}\end{array}\right.$$

From ①

$$\therefore \ell_{MLE} = \sum_{i=1}^{n} y_i (X\theta)_i - \sum_{i=1}^{n} \log(1 + \exp((X\theta)_i))$$

$$= \sum_{i=1}^{n} y_i (X\theta)_i - \sum_{i=1}^{n} \log(1_{n \times 1 \, i} + \exp(X\theta)_i)$$

$$= y^T X\theta - \sum_{i=1}^{n} \log(1_{n \times 1} + \exp(X\theta))_i$$

$$\left[x^T y = \sum_{i=1}^{n} x_i y_i\right]$$

$$= y^T x\theta - \sum_{i=1}^{n} (1_{n\times1})_i \log(1_{n\times1} + \exp(x\theta))_i$$

$$\left[\begin{array}{l} (1_{n\times1})_i = 1 \\ \text{multiplication by } 1 \end{array} \right]$$

$$\therefore \quad = y^T X\theta - 1_{n\times1}^T \cdot \log(1_{n\times1} + \exp(X\theta))$$

Hence

$$\boxed{\ell_{MLE}(\theta) = y^T X\theta - 1_{n\times1}^T \cdot \log(1_{n\times1} + \exp(X\theta))}$$

Ans: 3.5) We have,

$$\theta^T x_i = \sum_{k=1}^{d} \theta_k X_{ik}$$

$$\therefore \frac{\partial \theta^T x_i}{\partial \theta_j} = \sum_{k=1}^{d} \frac{\partial(\theta_k X_{ik})}{\partial \theta_j}$$

$$\frac{\partial(\theta^T x_i)}{\partial \theta_j} = X_{ij} \quad \text{——} \quad ①$$

Also, $\theta^T x_i = \sum_{k=1}^{d} \theta_k x_{ik}$

$$= \sum_{k=1}^{d} x_{ik} \times \theta_k$$

$$\theta^T x_i = (X\theta)_i \quad \text{——} \quad ② \quad \left[\begin{array}{l} \text{Definition of} \\ \text{Matrix} \\ \text{multiplication} \end{array} \right]$$

$$\ell_{MLE}(\theta) = \sum_{i=1}^{n} \log p(y_i \mid \theta x_i)$$

$$\ell_{MLE}(\theta) = \sum_{i=1}^{n} \left(y_i \theta^T x_i - \log(1 + \exp(\theta^T x_i)) \right)$$

$$\therefore \frac{\partial \mathcal{L}_{MLE}(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left(\frac{\partial (y_i \theta^T x_i)}{\partial \theta_j} - \frac{\partial \log(1+\exp(\theta^T x_i))}{\partial \theta_j} \right)$$

$$= \sum_{i=1}^{n} \left(y_i \frac{\partial \theta^T x_i}{\partial \theta_j} - \frac{1 \times \exp(\theta^T x_i)}{1+\exp(\theta^T x_i)} \times \frac{\partial \theta^T x_i}{\partial \theta_j} \right)$$

$$= \sum_{i=1}^{n} \left(y_i x_{ij} - \frac{1}{1+\exp(-\theta^T x_i)} \times x_{ij} \right)$$

[Using ①]

$$= \sum_{i=1}^{n} \left(y_i x_{ij} - \text{logistic}(\theta^T x_i) \times x_{ij} \right)$$

$$= \sum_{i=1}^{n} \left(y_i x_{ij} - \text{logistic}\left((X\theta)_i \right) \times x_{ij} \right)$$

[Using ②]

$$= \sum_{i=1}^{n} x_{ij} \left(y_i - (\text{logistic}(X\theta))_i \right)$$

$$= \sum_{i=1}^{n} (X^T)_{ji} \times \left(y - \text{logistic}(X\theta) \right)_i$$

$$\frac{\partial \mathcal{L}_{MLE}(\theta)}{\partial \theta_j} = \left(X^T (y - \text{logistic}(X\theta)) \right)_j$$

[Definition of Matrix multiplication]

$$\therefore \left(\frac{\partial \mathcal{L}_{MLE}(\theta)}{\partial \theta} \right)_j = \left(X^T (y - \text{logistic}(X\theta)) \right)_j$$

$$\therefore \boxed{\frac{\partial \mathcal{L}_{MLE}(\theta)}{\partial \theta} = X^T (y - \text{logistic}(X\theta))}$$

Ans: 4.1.2) SGD is slower than GD because n-samples number of python loop iterations of theta updates in case of SGD are replaced by a single theta update in case of GD which uses faster numpy ~~operations~~ vectorized operation. Since numpy vector operations in case of GD such as matrix multiplications and subtractions are faster than python loops in SGD, SGD is slower than GD and takes more time for the same number of epochs.