Q.1.

⇒ Let $X = \{x_1, x_2, x_3 \cdots x_n\}$ be a set of i.i.d (independent & identically distributed) samples drawn from a normal distribution s.t.

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad (1)$$

μ = mean of the distribution
σ = variance of the distribution

The probability density function of a normal distribution is:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Estimate $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$

$$L(\theta) = f(X_1, X_2, \cdots X_n \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \quad (3)$$

The MLE of θ is the value that maximizes $L(\theta)$. i.e $\theta_{MLE} = \arg\max_{\theta} L(\theta)$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} f(x_i \mid \mu, \sigma^2) \quad (4)$$

Substituting the Gaussian p.d.f:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

Taking the log-likeli ood to make differentiation easier,

$$\log L(\mu, \sigma^2) = \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right]$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \quad (6)$$

Taking the derivative w.r.t μ

$$\frac{\partial}{\partial\mu} \log L = \frac{\partial}{\partial\mu} \left(-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right.$$

$$= \frac{-1}{2\sigma^2} \sum_{i=1}^{n} \frac{\partial}{\partial\mu} (x_i - \mu)^2$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)$$

$$= -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \quad (7)$$

Setting this derivative to zero.

$$\frac{\partial}{\partial \mu} \log L = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0.$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - \mu) = 0.$$

$$\boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i} \qquad (8)$$

Taking the derivative w.r.t σ^2:

$$\frac{\partial}{\partial \sigma^2} \log L = \frac{\partial}{\partial \sigma^2} \left(-\frac{n}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad (9)$$

Setting this derivative to zero.

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\Rightarrow \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{n}{2\sigma^2}$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2} \qquad (10).$$

$\therefore$ (8) and (10) make up the MLE for univariate Gaussian.

Q.2.

=> Given a linear model

$$y(x, w) = w_0 + w_1 x_1 \quad (1)$$

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - y(x_n, w)\}^2 \quad (2)$$

Substituting (1) into (2)

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - (w_0 + w_1 x_n)\}^2 \quad (3)$$

To minimize $E(w)$, take partial derivatives of $E(w)$ w.r.t w_0 & w_1 and set to 0.

A) $$\frac{\partial E(w)}{\partial w_0} = \sum_{n=1}^{N}\{t_n - (w_0 + w_1 x_n)\}(-1) = 0 \quad (4)$$

$$= \sum_{n=1}^{N}\{t_n - w_0 - w_1 x_n\} = 0$$

Simplifying,

$$\sum_{n=1}^{N} t_n - w_0 N - w_1 \sum_{n=1}^{N} x_n = 0.$$

$$\Rightarrow \boxed{\sum_{n=1}^{N} t_n = w_0 N + w_1 \sum_{n=1}^{N} x_n} \quad (5)$$

B) $$\frac{\partial E(w)}{\partial w_1} = \sum_{n=1}^{N}\{t_n - (w_0 + w_1 x_n)\}(-x_n) = 0.$$

$$= \sum_{n=1}^{N}\{t_n - w_0 - w_1 x_n\} x_n = 0$$

Simplyfing,

$$\sum_{n=1}^{N} t_n x_n - w_0 \sum_{n=1}^{N} x_n - w_1 \sum_{n=1}^{N} x_n^2 = 0.$$

$$\Rightarrow \boxed{\sum_{n=1}^{N} t_n x_n = w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{N} x_n^2} \quad (6)$$

The two equations are:

$$w_0 N + w_1 \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} t_n$$

$$w_0 \sum_{n=1}^{N} x_n + w_1 \sum_{n=1}^{n} x_n^2 = \sum_{n=1}^{n} t_n x_n$$

In matrix form,

$$\begin{bmatrix} N & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} t_n \\ \sum_{n=1}^{N} x_n t_n \end{bmatrix} \quad (7)$$

Comparing (9) w.r.t $\sum_{j=0}^{1} A_{ij} w_j = T_i$

$$A_{00} = N \qquad A_{01} = \sum_{n=1}^{N} x_n$$

$$A_{10} = \sum_{n=1}^{N} x_n \qquad A_{11} = \sum_{n=1}^{N} x_n^2.$$

$$T_0 = \sum_{n=1}^{N} t_n \qquad T_1 = \sum_{n=1}^{N} t_n x_n$$

$\therefore$ The normal equations for the least squares solutions are:

$\sum_{j=0}^{1} A_{ij} w_j = T_i$ where

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}$$

$$T_i = \sum_{n=1}^{N} (x_n)^i t_n.$$

Q. 3.

=> For a binary classification problem, cross-entropy loss function is:

$$L_{CE} = -[y \log \hat{y} + (1-y) \log (1-\hat{y})] ; \quad (1)$$

$y \in \{0,1\}$ is true label
$\hat{y}$ is predicted probability.

$$\frac{\partial L_{CE}}{\partial \hat{y}} = -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)$$

$$= \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \quad (2)$$

The least square error function is defined as:

$$L_{LSE} = \frac{1}{2}(y-\hat{y})^2 ; \quad (3)$$

y is true label
$\hat{y}$ is the predicted probability.

$$\frac{\partial L_{LSE}}{\partial \hat{y}} = \hat{y} - y \quad (4)$$

The sigmoid function is defined as:

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (5)$$

$$\frac{\partial}{\partial z}\sigma(z) = \sigma(z)[1-\sigma(z)] = \hat{y}(1-\hat{y}) \quad (6)$$

Differentiating (2) w.r.t z,

$$\frac{\partial L_{CE}}{\partial z} = \frac{\partial L_{CE}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z}$$

$$= \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \cdot \hat{y}(1-\hat{y}).$$

$$= \hat{y} - y \quad (7)$$

This shows that when using sigmoid function $\hat{y} = \sigma(z)$, the derivative of CE Loss and least squares error are the same.

Q. 4.

=> for a discrete random variable X that can take k different categories, the multinomial p.m.f is:

$$P(x_1, x_2, \ldots x_k \mid n, p) = \frac{n!}{x_1! \, x_2! \cdots x_k!} \prod_{k=1}^{N} p_k^{x_k} \qquad (1)$$

; where

x_k : number of times category k is observed
p_k : probability of category k
n : $\sum_{k=1}^{N} x_k$ is total number of trials

p : $(p_1, p_2 \cdots p_k)$ vector of probabilities

When $n = 1$, the distribution becomes a categorical distribution

For a single observation X, which has k categories, $x = (x_1, x_2 \cdots x_k)$ is a one-hot vector where only one of the x_i is 1 and rest are 0.

Substituting $n = 1$ into (1)

$$P(X = x \mid p) = \frac{1!}{x_1! \, x_2! \cdots x_k!} \, p_1^{x_1} \, p_2^{x_2} \cdots p_k^{x_k}. \qquad (2)$$

Since, only one $x_i = 1$ and rest are 0,

$$P(X = x \mid p) = p_i \quad ; \text{where } x_i = 1. \qquad (3)$$

The likelihood function is given by:

$$L(p \mid X_1, X_2 \cdots X_m) = \prod_{j=1}^{m} p_{i_j}$$; i_j is the observed category for sample j.

$$L(p \mid \mathbf{X}) = \prod_{j=1}^{m} \prod_{i=1}^{k} p_i^{x_{ji}}$$; summing over all samples gives total count of each category i as $x_i = \sum_{j=1}^{m} x_{ji}$

$$\boxed{L(p \mid \mathbf{X}) = \prod_{i=1}^{k} p_i^{x_i}}$$

Q.5.

5.

=> Given,

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Let $y = \sigma(x) = \frac{1}{1+e^{-x}}$. (1)

Solving for x,

$$y(1+e^{-x}) = 1$$

or $y + ye^{-x} = 1$

or $ye^{-x} = 1-y$

Taking natural log on both sides,

$$\ln(ye^{-x}) = \ln(1-y)$$

or $\ln y + \ln e^{-x} = \ln(1-y)$

or $\ln y + (-x) = \ln(1-y)$

or $x = \ln y - \ln(1-y)$

or $\boxed{x = \ln\left(\frac{y}{1-y}\right)}$ (2)

$$\text{logit}(y) = \ln\left(\frac{y}{1-y}\right)$$

From (2), $x = \ln\left(\frac{y}{1-y}\right) = \text{logit}(y)$

$\therefore$ The inverse of the sigmoid function is the logit function.

References

1.2 - Maximum Likelihood Estimation | STAT 415

20_mle_annotated, Jerry Cain

Gaussian Distribution and Maximum Likelihood Estimate Method (Step-by-Step) | by Anel Music | The Startup | Medium

Lecture 6: The Method of Maximum Likelihood for Simple Linear Regression, CMU

Lecture 8: Properties of Maximum Likelihood Estimation (MLE), Purdue

Maximum Likelihood Estimation for Gaussian Distributions - Programmathically

Maximum likelihood estimation for the univariate Gaussian | The Book of Statistical Proofs