HOMEWORK 4

Full name	Nguyễn Thị Minh Ngọc
Student ID	11219280
Class	DSEB 63

1.

We have:

- The a data set of observations: $\mathbf{x} = (x_1, x_2 \dots x_N)^T$, representing N observations of the scalar variable x
- The corresponding target values of $\mathbf{x}$: $\mathbf{t} = (x_1, x_2 \dots x_N)^T$

Based on those data, we need to find a model which can help us to make predictions for some new value of the input variable x.

Suppose that the observations are drawn independently from a Gaussian distribution.

$$t = y(x, \mathbf{w}) + \text{noise}$$

With noise represents factors that affect the output but cannot be evaluated easily. Suppose that noise $\sim N(0, \beta^{-1}) \rightarrow\ t = y(x, w) + \text{noise} \sim N(y(x, \mathbf{w}), \beta^{-1})$ where $\beta = \frac{1}{\sigma^2}$. Then:

$$p(t|x, \mathbf{w}, \beta^{-1}) = \ N(t|x, \mathbf{w}, \beta^{-1})$$

To find the best model, we need to determine the unknown parameter **w** to maximize $p(t|x, \mathbf{w}, \beta^{-1})$. The likelihood function:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1}) = \prod_{n=1}^{N} N(t|x, \mathbf{w}, \beta^{-1})$$

Because $0 < p(t|x, \mathbf{w}, \beta^{-1}) < 1$, so the product of N p(t) will come to almost 0. Hence, it is convenient to maximize the logarithm of the likelihood function:

$$\log\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1})\right) = \sum_{n=1} \log\left(\frac{1}{\sqrt{2\pi^{\beta^{-1}}}} e^{\frac{-(t-y(x,\mathbf{w}))^2\beta}{2}}\right)$$

$$= N\log\frac{1}{\sqrt{2\pi^{\beta^{-1}}}} + \sum_{n=1} \log\left(e^{\frac{-(t-y(x,\mathbf{w}))^2\beta}{2}}\right)$$

$$= N\log\frac{1}{\sqrt{2\pi^{\beta^{-1}}}} + \sum_{n=1} \log\left(e^{\frac{-(t-y(x,\mathbf{w}))^2\beta}{2}}\right)$$

$$= N\log\frac{1}{\sqrt{2\pi^{\beta^{-1}}}} - \frac{\beta}{2}\sum_{n=1} (y(x, \mathbf{w}) - t)^2$$

To maximize $\log\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1})\right)$ and find **w**, we need to minimize $\frac{\beta}{2}\sum_{n=1}(y(x, w) - t)^2$ or minimize $\sum_{n=1}(y(x, \mathbf{w}) - t)^2$.

Let P = $\sum_{n=1}(y(x, \mathbf{w}) - t)^2$ where $y = w_1x + w_0$

Suppose that:

$$X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}; \qquad \mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}; \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} w_1x_1 + w_0 \\ w_2x_2 + w_0 \\ \vdots \\ w_nx_n + w_0 \end{bmatrix} = X \cdot \mathbf{w}$$

Then $P = \|y(x_1\mathbf{w}) - t\|_2^2 = \|X.\mathbf{w} - t\|_2^2$

We have:

$$\frac{\partial P}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial P}{\partial w_0} \\ \frac{\partial P}{\partial w_1} \end{bmatrix} = \begin{bmatrix} 2(X\mathbf{w} - t) \\ 2X(X\mathbf{w} - t) \end{bmatrix} = 2X^T(X\mathbf{w} - t) = 0$$

$$\rightarrow 2X^T\mathbf{w}X - 2X^Tt = 0$$

$$\rightarrow X^T\mathbf{w}X = X^Tt$$

$$\rightarrow \mathbf{w} = (X\text{^}TX)^{-1}X^Tt$$

Extra: Prove that X^TX is invertible when X is full of rank

…