DATE: / /

→ **Linear Regression :-** $D : \{(x_1, y_1) \ldots (x_N, y_N)\}$

$n = 1, 2, \ldots N$

For likelihood function,

$$P(y|x) = \mathcal{N}(y \mid f(x), \sigma^2)$$

where, $x_n \in R^D$ are inputs

$y_n \in R$ are noisy function values.

$$y = f(x) + \epsilon. \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

i.i.d. gaussian noise.

Now, consider the parameters are linear in model.

$$P(y \mid x, \theta) = \mathcal{N}(y \mid x^T\theta, \sigma^2)$$

$$\Leftrightarrow y = x^T\theta + \epsilon$$

$\theta \in R^D \leftarrow$ parameters.

Using Probabilistic graphical model,

$$P(Y \mid X, \theta) = P(y_1, \ldots y_N \mid x_1, \ldots x_N, \theta)$$

$$= \prod_{n=1}^{N} P(y_n \mid x_n, \theta)$$

$$= \prod_{n=1}^{N} \mathcal{N}(y_n \mid x_n^T\theta, \sigma^2)$$

where $X = \{x_1, \ldots, x_N\}$

$Y = \{y_1, \ldots, y_N\}$

For maximum likelihood of parameters,

$$\theta_{ML} = \arg\max_{\theta} p(Y \mid X, \theta)$$

take negative log-likelihood,

$$L(\theta) := -\log p(Y \mid X, \theta) = -\log \prod_{n=1}^{N} P(y_n \mid x_n, \theta)$$

$$= -\sum_{n=1}^{N} \log P(y_n \mid x_n, \theta)$$

we know that $y = x^T\theta + \epsilon$, $\epsilon \sim N(0, 6^2)$

$$L(\theta) = -\sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi 6^2}} e^{-\frac{(y_n - x_n^T\theta)^2}{26^2}}\right)$$

$$= \frac{1}{26^2} \sum_{n=1}^{N} (y_n - x_n^T\theta)^2$$

$$= \frac{1}{26^2} (y - X\theta)^T (y - X\theta)$$

$$= \frac{1}{26^2} \| y - X\theta \|^2$$

$X := [x_1, \ldots x_N]^T \in R^{N \times D}$

$y := [y_1, \ldots y_N]^T \in R^N$

differentiate w.r.t. θ,

$$\frac{dL}{d\theta} = \frac{d}{d\theta}\left(\frac{1}{26^2} (y - X\theta)^T (y - X\theta)\right)$$

$$= \cancel{\frac{1}{26^2} \frac{d}{d\theta} (y^T y) - }$$

$$= \frac{d}{d\theta}\left(\frac{1}{26^2} (y^T - \theta^T X^T)(y - X\theta)\right)$$

PAGE NO.:
DATE: / /

$$\frac{dL}{d\theta} = \cancel{\phi}\frac{1}{2\sigma^2}\frac{d}{d\theta}\left(y^Ty - y^TX\theta - \theta^TX^Ty + \theta^TX^TX\theta\right)$$

$$\frac{dL}{d\theta} = \frac{1}{2\sigma^2}\frac{d}{d\theta}\left(y^Ty - 2y^TX\theta + \theta^TX^TX\theta\right)$$

$$(\because (\theta^TX^Ty)^T = y^TX\theta)$$

$$\frac{dL}{d\theta} = \frac{1}{2\sigma^2}\left(-2y^TX + 2\theta^TX^TX\right)$$

$$\frac{dL}{d\theta} = \frac{1}{\sigma^2}\left(-y^TX + \theta^TX^TX\right) = 0^T$$

$$-y^TX + \theta_{ML}^TX^TX = 0^T$$

$$\theta_{ML}^TX^TX = y^TX$$

$$\theta_{ML}^T = y^TX(X^TX)^{-1}$$

$$\theta_{ML} = \left(y^TX(X^TX)^{-1}\right)^T$$

$$\theta_{ML} = (X^TX)^{-1}X^Ty$$

Now, if inputs are in non-linear transformation,

$$P(y \mid x, \theta) = \mathcal{N}(y \mid \phi^T(x)\theta, \sigma^2)$$

$$\Leftrightarrow \quad y = \phi^T(x)\theta + \epsilon$$

$$= \sum_{k=0}^{K-1} \theta_k \phi_k(x) + \epsilon$$

where $\phi : R^D \rightarrow R^K$ non linear transformation of input X.

$\phi_k : R^D \rightarrow R$ k^{th} component of feature vector ϕ.

$$\Phi = \begin{bmatrix} \phi^T(x_1) \\ \vdots \\ \phi^T(x_N) \end{bmatrix} = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{K-1}(x_1) \\ \vdots & & \\ \phi_0(x_N) & \cdots & \phi_{K-1}(x_N) \end{bmatrix}_{N \times K}$$

Taking negative -log likelihood

$$-\log P(Y|x,\theta) = \frac{1}{2\sigma^2}(y-\Phi\theta)^T(y-\Phi\theta) + c$$

Maximum likelihood estimation for parameter,

$$\theta_{ML} = (\Phi^T\Phi)^{-1}\Phi^T y$$