(3) The method of ordinary least squares assumes that there is constant variance in the errors.

Let the data observed $D = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$

Let us assume that the targets are generated by $y(x, w)$

i.e, $t = y(x, w) + \varepsilon$ $\qquad \varepsilon \sim N(\varepsilon \mid 0, \sigma^2)$

where ε follows a normal distribution with varying variance.

a) likelihood for a heteroscedastic setting of a single data point =

$$p(t_n \mid x_n, w, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{1}{2\sigma_n^2}(t_i - y(x_i; w))^2\right)$$

Now, we are going to model the prior distribution also with normal distribution.

Prior =

$$p(w \mid \alpha) = N(w_n \mid 0, \alpha^{-1}) \qquad \alpha = \text{precision parameter}$$

$$= \left(\frac{\alpha}{2\pi}\right)^{1/2} \cdot \exp\left(-\frac{\alpha}{2} w_n \cdot w_n\right)$$

b) Maximum likelihood estimation:

$$\text{likelihood} = P(t_1 \dots t_n \mid x_1 \dots x_n, \omega, \sigma^2)$$

$$= p(t_1 \mid x_1, \omega, \sigma_1^2) \cdot p(t_2 \mid x_2, \omega, \sigma_2^2) \dots p(t_n \mid x_n, \omega, \sigma_n^2)$$

$$= \prod_{i=1}^{n} p(t_i \mid x_i, \omega, \sigma_i^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(t_i - y(x_i,\omega))^2}{2\sigma_i^2}}$$

Instead of working with likelihood, it will be easier to work with loglikelihood.

$$\log(\text{likelihood}) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}(t_i - y(x_i,\omega))^2$$

This is the objective function that will be considered for the ML estimate of parameters.

$$\Rightarrow \omega_{ML} = \arg\max_{\omega} \left[\underbrace{\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right)}_{\text{this does not depend on } \omega} - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}(t_i - y(x_i,\omega))^2 \right]$$

$$\Rightarrow \omega_{ML} = \arg\min_{\omega} \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(t_i - y(x_i,\omega))^2$$

4a

MAP estimation :

Given a prior, the posterior distribution is given by:

$$p(w \mid x, t, \beta, \alpha) = \frac{p(t \mid x, w, \beta) \cdot p(w \mid \alpha)}{p(t \mid x, \beta, \alpha)}$$

we have to maximize this

this does not depend on w

$$p(w \mid \alpha) = \prod_{i=1}^{N} \mathcal{N}(w_i \mid 0, \alpha^{-1}) = \prod_{i=1}^{N} \left(\frac{\alpha}{2\pi}\right)^{1/2} e^{-\frac{\alpha}{2} w_i w_i}$$

$$= \left(\frac{\alpha}{2\pi}\right)^{N/2} \prod_{i=1}^{N} e^{-\frac{\alpha}{2} w_i w_i}$$

$$\log P(w \mid \alpha) = \log\left(\frac{\alpha}{2\pi}\right)^{N/2} - \frac{\alpha}{2} w^T w$$

$$P(t \mid x, w, \beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot e^{\frac{-(t_i - y(x_i, w))^2}{2\sigma_i^2}}$$

$$\log P(t \mid x, w, \beta) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} (t_i - y(x_i, w))^2$$

$$\log P(w \mid x, t, \beta, \alpha) = \log P(t \mid x, w, \beta) + \log P(w, \alpha) - \log P(t \mid x, \beta, \alpha)$$

does not depend on w

$$-\log P(w \mid x, t, \beta, \alpha) = -\log P(t \mid x, w, \beta) - \log P(w, \alpha) + \log P(t \mid x, \beta, \alpha)$$

$$w_{MAP} = \arg\min_{w} -\log P(w \mid x, t, \beta, \alpha)$$

$$= \arg\min_{w} \left[\frac{1}{2} \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (t_i - y(x_i, w))^2 + \frac{\alpha}{2} w^T w \right]$$

c) from maximum likelihood, objective function is given by

$$\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(t_i - y(x_i,w)\right)^2$$

Let $r_i = \frac{1}{\sigma_i^2}$ $(r_i > 0)$

$$\Rightarrow E_D(w) = \frac{1}{2}\sum_{i=1}^{N} r_i\left(t_i - y(x_i,w)\right)^2$$

sum of squares error function with weighting factor.

Let $y(x,w) = w^T\phi(x)$ be linear in terms of basis functions $\phi_i(x)$

$$\Rightarrow E_D(w) = \frac{1}{2}\sum_{i=1}^{N} r_i\left(t_i - w^T\phi(x_i)\right)^2$$

To minimize $E_D(w)$, $\frac{\partial}{\partial w}E_D(w) = 0$

$$\Rightarrow \frac{\partial}{\partial w}\frac{1}{2}\sum_{i=1}^{N} r_i\left(t_i - w^T\phi(x_i)\right)^2 = 0$$

$$\frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial w} r_i\left(t_i - w^T\phi(x_i)\right)^2 = 0$$

$$\cancel{\frac{1}{2}}\sum_{i=1}^{N}\cancel{2}r_i\left(t_i - w^T\phi(x_i)\right)\cdot\frac{\partial}{\partial w}\left(-w^T\phi(x_i)\right) = 0$$

$$\sum_{i=1}^{N} r_i\left(t_i - w^T\phi(x_i)\right)\frac{\partial}{\partial w}\left(-\phi(x_i)^T w\right) = 0$$

$$\sum_{i=1}^{N} r_i\left(t_i - w^T\phi(x_i)\right)\phi(x_i)^T = 0$$

$$\Rightarrow \sum_{i=1}^{N} w^T r_i \phi(x_i)\phi(x_i)^T = \sum_{i=1}^{N} r_i t_i \phi(x_i)^T$$

Taking transpose :

$$\left(\sum_{i=1}^{N} r_i \phi(x_i)\phi(x_i)^T\right) w = \sum_{i=1}^{N} r_i t_i \phi(x_i)$$

Let $\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & & & \vdots \\ \phi_0(x_N) & \cdots & \cdots & \phi_{M-1}(x_N) \end{bmatrix}_{N\times M}$

~~Let R =~~

$\sum_{i=1}^{N} r_i \phi(x_i)\phi(x_i)^T$ can be expressed as

$$\underbrace{\begin{bmatrix} r_1\phi_0(x_1) & r_2\phi_0(x_2) & \cdots & r_n\phi_0(x_N) \\ r_1\phi_1(x_1) & r_2\phi_1(x_2) & & \vdots \\ \vdots & \vdots & & \vdots \\ r_1\phi_{M-1}(x_1) & r_2\phi_{M-1}(x_2) & & r_n\phi_{M-1}(x_N) \end{bmatrix}_{M\times N}}_{\Phi^T R} \underbrace{\begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & & & \\ \phi_0(x_N) & \cdots & \cdots & \phi_{M-1}(x_N) \end{bmatrix}_{N\times M}}_{\Phi}$$

where $R = \begin{bmatrix} r_1 & 0 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & & 0 \\ 0 & 0 & \cdots & & r_n \end{bmatrix}_{N\times N}$ → diagonal matrix.

$$\Rightarrow (\Phi^T R \Phi) w = \Phi^T R t$$

$$\Rightarrow \boxed{w = (\Phi^T R \Phi)^{-1} \Phi^T R t}$$

where $t = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix}_{N\times 1}$