FOML ASSIGNMENT (Akshay kumar)

Group Number - 57 (Abhishree Gajanan)

Question 3: Part (a)

Given target variable 't_n' and input 'x_n', the likelihood function in heteroscedastic setting for a single data point is given by

$$P(t_n | x_n, \theta) = \mathcal{N}(t_n | x_n, \theta)$$

as it follows gaussian distribution with mean $\mu_n(x_n, \theta)$.

Now,

$$\mathcal{N}(t_n | x_n, \theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma_n^2}\right)}$$

Hence, the likelihood function in heteroscedastic setting for single data point is given by

$$\boxed{P(t_n | x_n, \theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma_n}\right)}}$$

Prior

The prior distribution for a heteroscedastic setting can be specified in different ways. It encode our belif or prior knowledge about the parameter θ.

$$P(\theta) = \mathcal{N}(0, \theta)$$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(0 - \theta^T 1)^2}{2\sigma_n^2}}$$

$$= \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\theta^2}{2\sigma_n^2}}$$

$$\boxed{P(\theta) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\theta\theta^T}{2\sigma_n^2}}}$$

Question 3: Part (b)

The maximum likelihood estimation of the parameter θ is obtained by the likelihood function that is product of the product of likelihood of each data point calculated in first part of this question.

$$L(\theta) = \prod_{n=1}^{N} P(t_n | x_n, \theta)$$

$$= \prod_{n=1}^{N} \mathcal{N}(t_n | x_n, \theta)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma^2}\right)}$$

To simplify computation, use logarithm of the likelihood (log-likelihood) because it converts the product into sum which is easier to work.

$$\log(L(\theta)) = \log\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma^2}\right)}\right)$$

$$\log(L(\theta)) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right) + \sum_{i=1}^{N} \log\left(\exp^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma^2}\right)}\right)$$

Ignore the constant term

$$\log(L(\theta)) = \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(t_n - \theta^T x_n)^2$$

$$\log(L(\theta)) = \underset{(\theta)}{\arg\max}\left(\sum_{i=1}^{N} -\frac{1}{2\sigma_n^2}(t_n - \theta^T x_n)^2\right)$$

Remone -ve sign from expression and convert to argmin

$$\boxed{\log L(\theta) = \underset{(\theta)}{\arg\min}\left(\sum_{n=1}^{N} \frac{1}{2\sigma_n^2}(t_n - \theta^T x_n)^2\right)}$$

Maximum A Posteriori (MAP estimate):

In MAP estimation, we include the prior information about the parameters θ. The objective function of MAP estimation for θ can be given as

$$P(\theta \mid t_n, x_n) \propto P(t_n \mid x_n, \theta) \cdot P(\theta)$$

$$P(\theta \mid t_n, x_n) = \prod_{n=1}^{N}\left\{\frac{1}{\sqrt{2\pi\sigma_n^2}} \exp^{-\left(\frac{(t_n - \theta^T x_n)^2}{2\sigma_n^2}\right)}\right\} \cdot \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\theta\theta^T}{2\sigma_n^2}}$$

Taking log both sides. to simplify

$$\log(P(\theta \mid t_n, x_n)) = \sum_{n=1}^{N} \log\frac{1}{\sqrt{2\pi\sigma_n^2}} + \sum_{i=1}^{N} -\frac{(t_n - \theta^T x_n)^2}{2\sigma_n^2} + \log\frac{1}{\sqrt{2\pi\sigma_n^2}} + \left(-\frac{\theta\theta^T}{2\sigma_n^2}\right)$$

Ignoring the constant terms.

$$\log(P(\theta \mid t_n, x_n)) = \sum_{n=1}^{N} -\frac{1}{2\sigma_n^2}(t_n - \theta^T x_n)^2 + -\frac{1}{2\sigma_n^2}(\theta\theta^T)$$

Maximize the above equation

$$\log(P(\theta|t_n,x_n)) = \underset{(\theta)}{\operatorname{argmax}}\left\{\sum_{n=1}^{N} -\frac{1}{2\sigma_n^2}(t_n-\theta^T x_n)^2 - \frac{1}{2\sigma_n^2}(\theta\theta^T)\right\}$$

Remove -ve sign and convert to argmin

$$\boxed{\log(P(\theta|t_n,X_n)) = \underset{(\theta)}{\operatorname{argmin}}\left\{\sum_{n=1}^{N} +\frac{1}{2\sigma_n^2}(t_n-\theta^T x_n)^2 + \frac{1}{2\sigma_n^2}\theta\theta^T\right\}}$$

Question 3: Part (c)

In part (b) of the question we have calculated the ML objective which is as given below

$$\text{Log}(L(\theta)) = \underset{(\theta)}{\operatorname{argmin}}\left\{\sum_{n=1}^{N}\frac{1}{2\sigma_n^2}(t_n-\theta^T t_x)^2\right\}$$

comparing it with the expression given in the question

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}^2$$

Ignoring the constant term $\frac{1}{2}$ we find that

$r_n = \frac{1}{\sigma_n^2}$; as σ_n^2 is always >0 hence

$\frac{1}{\sigma_n^2}$ is also >0

And, hence the term $\boxed{r_n = \frac{1}{\sigma_n^2} > 0}$

Now,

To find w that minimizes the given error function, we find $\frac{\partial}{\partial w}(E_D(w))$ and equate it to '0'.

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

$$\frac{\partial}{\partial w}\left\{\frac{1}{2}\sum_{n=1}^{N} \gamma_n (t_n - w^T \phi(x_n))^2\right\} = 0$$

$$\frac{1}{\cancel{2}}\sum_{n=1}^{N} \gamma_n \cdot \cancel{2}(t_n - w^T \phi(x_n))(\phi(x_n))^T = 0$$

$$\sum_{n=1}^{N} (\gamma_n t_n - \gamma_n w^T \phi(x_n))(\phi(x_n))^T = 0$$

$$\sum_{n=1}^{N} \gamma_n t_n (\phi(x_n))^T - \sum_{n=1}^{N} \gamma_n w^T \phi(x_n)(\phi(x_n))^T = 0$$

$$\sum_{n=1}^{N} \gamma_n t_n (\phi(x_n))^T = \sum_{n=1}^{N} \gamma_n w^T \phi(x_n)(\phi(x_n))^T$$

$$\boxed{w^T = \frac{\sum_{n=1}^{N} \gamma_n t_n (\phi(x_n))^T}{\sum_{n=1}^{N} \gamma_n w^T \phi(x_n)(\phi(x_n))^T}}$$