Homework 2 ML Mar 5th 2023

1. $E_D(w) = \frac{1}{2}\sum r_n(t_n - w^T\phi(x_n))^2$

To minimize the square error, we need to find the largest w^T. So take the gradient w^T

$$\frac{\partial}{\partial w}\frac{1}{2}\sum r_n(t_n - w^T\phi(x_n))^2 = 0$$

$$\therefore \sum r_n(t_n - w^T\phi(x_n))\cdot\phi(x_n) = 0$$

$$\therefore \sum r_n t_n \phi_n = \sum w^T\phi(x_n)\phi(x_n)^T\cdot r_n$$

$$\therefore w^T = \frac{\sum r_n t_n \phi_n}{\sum \phi(x_n)\phi(x_n)^T r_n}$$

① Assume $\varepsilon_i \sim N(0, \sigma^2)$ so $y_i \sim N(x_i w^T, \sigma^2)$

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

Suppose $\sigma^2 = \frac{1}{2r_n}$ then $E_D(w) = \frac{1}{\frac{2}{2r_n}}\sum(t_n - \phi(x_n)w^T)^2$

$$= \sum r_n(t_n - \phi(x_n)w^T)^2 \text{ proved!}$$

② ~~If~~ r_n is the replicated data points

then $\sum(t_n - \phi(x_n)w^T)^2 = \sum(t_n^* r_n^2 - r_n^2\phi^*(x_n)w^T)^2$

$$= \sum r_n(t_n^* - \phi^*(x_n)w^T)^2$$

Question 1a

2. $y(x, w) = w_0 + \sum_{i=1}^{D} w_i x_i = \sum_{i=0}^{D} w_i x_i$

$$E_D(w) = \frac{1}{2} \sum_{j=1}^{N} \{ y(x_n, w) - t_n \}^2$$

$$= \cancel{\frac{1}{2} \sum_{n=1}^{N} \{ \sum_{i=0}^{D} w_i x_i - t_n \}^2}$$

~~To minimize $E_D(w)$, we need differentiate $E_D(w)$ over w~~

$$\cancel{\therefore \frac{\partial E_D(w)}{\partial w} = 0 = \sum_{n=1}^{N} (\sum_{i=0}^{D} w_i x_i - t_n) x_i}$$

$$\cancel{\therefore \sum_{n=1}^{N} t_n x_i = \sum_{n=1}^{N} \sum_{i=0}^{D} w_i x_i}$$

$$\hat{y} = w_0 + \sum_{i=1}^{D} w_i (x_i + \varepsilon) = w_0 + \sum^{D} w_i x_i + \sum w_i \varepsilon = y + \sum w_i \varepsilon$$

$$\therefore \hat{E}_D(w) = \frac{1}{2} \sum_{j=1}^{N} \{ \sum_{i=1}^{D} w_i \varepsilon_i + y - t_n \}^2$$

$$= \frac{1}{2} \sum^{N} \{ \sum^{D} w_i^2 \varepsilon_i^2 + 2 \sum^{D} w_i \varepsilon_i (y - t_n) \cancel{+ (y - t_n)^2} \}$$

$$= \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{D} w_i \varepsilon_i \varepsilon_j + \sum_{j=1}^{N} \sum_{i=1}^{D} w_i \varepsilon_i (y - t_n) + \frac{1}{2} \sum^{N} (y - t_n)^2$$

Given that $\mathbb{E}(\varepsilon_i \varepsilon_j) = \delta_{ji} \sigma^2$ $\mathbb{E}(\varepsilon_i) = 0$ $\therefore \sum w_i \varepsilon_i = 0$

$$\therefore \hat{E} = \frac{1}{2} \sum \sum w_i^2 \delta_{ij} + 0 + \frac{1}{2} \sum^{N} (y - t_n)^2$$

$$= \frac{1}{2} \sum^{N} \{ y - t_n \} + \frac{\sigma^2}{2} \sum \sum w_i^2 \delta_{ij}$$

$$= E_D(w) + \frac{\sigma^2}{2} \sum \sum w_i^2 \delta_{ij}$$ proved.

Question1b

HW2 Question 3

From the formula (3.59) in the textbook:

$$\sigma^2(x) = \underbrace{\beta^{-1}}_{\text{noise}} + \underbrace{\phi(x)^T S_N \phi(x)}_{\text{uncertainty}}$$

formula (3.51)

$$S_N^{-1} = S_0^{-1} + \beta \Phi \Phi^T \Rightarrow S_0^{-1} + \beta \sum_{i=1}^{N} \phi(x_i) \phi(x_i)^T$$

Therefore

$$S_{n+1}^{-1} = S_N^{-1} + \beta \phi(x_{N+1}) \phi(x_{N+1})^T$$

$$\therefore \quad \sigma_{N+1}^2(x) - \sigma_N^2(x) = \phi(x_{N+1}) \underline{(S_{n+1}^{-1} - S_n^{-1})} \phi(x_{N+1})^T \quad ①$$

Now, Given ~~these~~ formula from the question,

$$(M + vv^T)^{-1} = S_N^{-1} + \sqrt{\beta}\phi(x_{n+1}) \sqrt{\beta}\phi(x_{n+1})^T = S_{n+1}$$

If $M = S_N^{-1}$ $\quad v = \sqrt{\beta}\phi(x_{n+1})$

$$\therefore S_{n+1}^{-1} = S_N^{-1} - \frac{(S_N \sqrt{\beta}\phi(x_{n+1}))(\sqrt{\beta}\phi(x_{n+1})^T S_N)}{1 + \sqrt{\beta}\phi(x_{n+1})^T S_n \phi(x_{n+1})} \quad ②$$

So replace $S_{n+1}^{-1} - S_n^{-1}$ in ① by ②

$$\sigma_{n+1}^2(x) - \sigma_n^2 = - \frac{\phi(x_{n+1}) S_N \sqrt{\beta}\phi(x_{n+1}) (\sqrt{\beta}\phi(x_{n+1})^T S_N) \phi(x_{n+1})^T}{1 + \sqrt{\beta}\phi(x_{n+1})^T S_n \phi(x_{n+1})} \quad ③$$

Knowing that $\phi(x_{n+1})$, $\sqrt{\beta}$, S_N, 1 are all possitive

so ③ is negative

which means $\sigma_{n+1}^2(x) - \sigma_n^2(x) < 0$ so $\sigma_{n+1}^2(x) < \sigma_n^2(x)$

Question1c

4. $p(w,t) \propto \left(\prod_{n=1}^{N} \mathcal{N}(t_n; y(x_n, w), \sigma_1^2)\right) \times \left(\mathcal{N}(w; 0, \sigma_2^2 \Pi)\right)$

solution:

Given the Gaussian prior, the likelihood can be written as:

$$p(t|w) = \prod_{n=1}^{N} \mathcal{N}(t_n; w^T\phi(x_n), \sigma_1^2)$$

So the posterior can be written in form of Gaussian as the product of likelihood x prior

prior: $p(w|\sigma_2^2 \mathrm{I}) = \frac{1}{\sqrt{2\pi}\,\sigma_2 \mathrm{I}} \exp(-\frac{1}{2} w^T \sigma_2^{-1} w)$

take the ln $\underset{\alpha}{\underset{\parallel}{\sigma_2^2 \mathrm{I}}}$ suppose $\beta = \sigma_1^2$, $\alpha = \sigma_2^2 \mathrm{I}$

likelihood: $\therefore \ln p(w|\alpha) = -\frac{N}{2}\ln w^T w \alpha - \frac{N}{2}\ln 2\pi = -\alpha E_w(w) + \text{const}.$

and $p(t|w) = \frac{1}{\sqrt{2\beta^{-1}\pi}} \prod_{n=1}^{N} \exp(-\frac{(t_n - w^T\phi(x_n))^2}{2\beta^{-1}})$

$$\therefore \ln p(t|w) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2\beta + \ln(2\beta^{-1})^{-\frac{1}{2}}$$

$$= \underbrace{\frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2\beta}_{-\beta E_D(w)} + \text{const}.$$

Therefore to maximize $\ln p(w|\alpha) + \ln(w|\beta)$ which is derived from $\arg\max p(w|\alpha) \cdot p(t|w) \stackrel{\ln}{=} \ln(w|\alpha) + \ln(t|\beta)$ can be written as $-\alpha E_w(w) + \beta E_D(w)$ if we discard the constant term that don't depend on w.

Therefore max posterior is equivalent to max $-(\alpha E_w(w) + \beta E_D$ which is minimization of $\alpha E_w(w) + \beta E_D(w)$

To uniform this equation to $E_D(w) + \lambda E_w(w) = \beta E_D(w) + \alpha E_w$

$$\lambda = \frac{\alpha}{\beta} = \frac{\sigma_2^2 \mathrm{I}}{\sigma_1^2}$$

Question 1d

Question2

https://colab.research.google.com/drive/1tNr-Jz3DrmaUb360j9fDPYYs2ZjTweol?authuser=1#scrollTo=XmJ_Ei7bYV6I