Foundations of Machine Learning

Assignment - 4

Q.1] Non-Uniform weights in Linear Regression.

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

i) Taking gradient of above function and equating it to zero.

$$\frac{\partial}{\partial w} E_D(w) = -\sum_{n=1}^{N} r_n \{t_n - w^T\phi(x_n)\} \phi(x_n) = 0$$

Solving for w

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

$$\therefore w = \left(\sum_{n=1}^{N} r_n \phi(x_n)\phi(x_n)^T\right)^{-1} \left(\sum_{n=1}^{N} r_n t_n \phi(x_n)\right)$$

ii) Matrix form: Rewriting the error function in terms of matrix products

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

$$= \frac{1}{2} (\phi w - t)^T R (\phi w - t)$$

$$= \frac{1}{2} (w^T\phi^T R \phi w - w^T\phi^T R t - t^T R \phi w + t^T R t)$$

$$= \frac{1}{2} (w^T\phi^T R \phi w - 2t^T R \phi w + t^T R t)$$

$R \Rightarrow \text{diag}(r_1, \ldots r_N)$

$\therefore$ Gradient of error function $\Rightarrow \phi^T R \phi w - t^T R \phi$

$$\therefore w = (\phi^T R \phi)^{-1} t^T R \phi$$

$$\therefore \boxed{w = (\phi^T R \phi)^{-1} \phi^T R t}$$

Scanned with CamScanner

b) Let Define ϕ as $N \times N$ matrix, $\phi(i,j) = \sqrt{g_i}\, \phi_j(x_i)$

The interpretation in 2 folds

i) In data dependence noise variance

$$\ln P(F/w, \beta) = \sum_{n=1}^{N} \ln N\left(t_n \left(w^T \phi(x_n), \beta^{-1}\right)\right)$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(w)$$

If we substitute β^{-1} by $x_n \beta^{-1}$, then expression would be,

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

ii) Replicated Datapoints,

g_n is effective number of observations of (x_n, t_n) as repeatedly occuring x_n times.

$$\sigma_\epsilon^2 = (1 - \rho_{xy}^2)\, \sigma_y^2$$

ρ_{xy} ⇒ correlation coefficient for population.

Scanned with CamScanner

Q.2) Bayes Optimal Classifier

Here,

$\Sigma P(F|h_i) \cdot P(h_i|D) = 0.4$

$\Sigma P(L|h_i) \cdot P(h_i|D) = 0.2 + 0.1 + 0.2 = 0.5$

$\Sigma P(R|h_i) \cdot P(h_i|D) = 0.1$

$\therefore$ Most probable hypothesis,

$(L) \cdot P(L|h_i) = 0.5$. i.e. Bayes Optimal Estimate

ii) Map Hypothesis:

$H_{MAP} \Rightarrow \text{argmax } P(D|h) \cdot P(h)$.

For Forward, $\Rightarrow \text{argmax } (0.4 \times 1) = 0.4$

For Left $\Rightarrow \text{argmax } (0.2 \times 1, \; 0.1 \times 1, \; 0.2 \times 1)$

$\Rightarrow \text{argmax } (0.2, \; 0.1, \; 0.2)$

$\Rightarrow 0.2$

For Right $\Rightarrow \text{argmax } (0.1 \times 1) = 0.1$

$\therefore$ For MAP Hypothesis, Forward has maximum value.

Hence, both are not same.

Scanned with CamScanner

Q.3 $H = \{p, q\}$. classified as I iff $p < x < q$.

The VC dimension of a set of hypothesis H is the size of the largest set $C \subseteq X$ such that C is shattered by H.

a) Let the training points are in sphere of Radius 'R'.

Let $G(x) = \text{sign}(f(x)) = \text{sign}(\beta_I x + \beta_0)$

b) So, class of functions.

$\{G(x), \|B\| \leq A\}$, it has VC dimension H satisfying $\leq R^2 A^2$

$\therefore$ VC dimension of 1-D data of hypothesis space H is $\boxed{H \leq R^2 A^2}$

Q.4 $y(x, w) = w_0 + \sum_{k=1}^{D} w_k x_k \rightarrow$ 1D data $\quad$ - I

Sum of squares error function for N data samples

$$E_D(w) = \frac{1}{2} \sum_{i=1}^{N} (y(x_i, w) - t_i)^2 \quad \text{-II}$$

$i = 1, 2, \ldots N$

Substituting I in II with noise ε

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} \left\{ \left[w_0 + \sum_{i=1}^{D} w_i (x_i + \varepsilon_i) \right] - t_i \right\}^2$$

$$= \frac{1}{2} \sum_{i=1}^{N} \left\{ (w_0 + \sum_{i=1}^{D} w_i x_i) - t_i + \sum_{i=1}^{D} w_i \varepsilon_i \right\}^2$$

Scanned with CamScanner

$$= \frac{1}{2}\sum_{i=1}^{N}\left\{\left[y(x_i,w)-t_i\right]+\sum_{i=1}^{D} w_i \varepsilon_i\right\}^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left\{(y(x_i,w)-t_i)^2+\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)^2+ 2\cdot\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)\left(y(x_i,w)-t_i\right)\right\}.$$

$y(x_i, w) \Rightarrow$ denotes the output of linear model when input variable is x_i without Gaussian noise added. The second term equation we have,

$$E_\varepsilon\left[\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)^2\right] = E_\varepsilon\left[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \varepsilon_i\varepsilon_j\right]$$

$$= \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j E_\varepsilon[\varepsilon_i\varepsilon_j]$$

$$\varepsilon \sim (0,\sigma^2) \Rightarrow \sigma^2 \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij}$$

$$\Rightarrow \quad E_\varepsilon\left[\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)^2\right] = \sigma^2\sum_{i=1}^{D} w_i^2$$

Last term,

$$E_\varepsilon\left[2\left(\sum_{i=1}^{D} w_i\varepsilon_i\right)\left(y(x_i,w)-t_i\right)\right]$$

$$\Rightarrow 2\left(y(x_i,w)-t_i\right) E_\varepsilon\left[\sum_{i=1}^{D} w_i\varepsilon_i\right]$$

$$\Rightarrow 2\left(y(x_i,w)-t_i\right)\sum_{i=1}^{D} E_\varepsilon[w_i\varepsilon_i] = 0$$

Therefore, if we calculate the expectation of $E_D(w)$, with respect to Gaussian noise ε, we can obtain.

$$E_\varepsilon\left[E_D(w)\right] = \frac{1}{2}\sum_{i=1}^{N}\left(y(x_i,w)-t_i\right)^2 + \frac{\sigma^2}{2}\sum_{i=1}^{D} w_i^2$$