

MTH9898_Chenyu_Zhao



北京大学
PEKING UNIVERSITY

orthogonal complement of ϕ

$$3.2 \quad A = \{V \mid \phi^T V = 0\} \quad B = \{V \mid \text{column space of } \phi, \exists w, \phi w = V\}$$

$$\text{for } V_1 \in A, \quad \phi(\phi^T \phi)^+ \phi^T V_1 = 0$$

$$\text{for } V_2 \in B, \quad \phi(\phi^T \phi)^+ \phi^T V_2 = \phi(\phi^T \phi)^+ \phi^T \phi w = \phi w = V_2$$

Any vector V can be decomposed to $V_1 + V_2$, where $V_1 \in A, V_2 \in B$
this is orthogonal decomposition which can be maximally done by ~~splitting~~ orthogonalize

$$\therefore \phi(\phi^T \phi)^+ \phi^T V = \phi(\phi^T \phi)^+ \phi^T (V_1 + V_2) = V_2$$

\therefore it projects V to ϕ 's column space

$$\therefore \text{the solution of OLS } \hat{\beta} = (X^T X)^+ X^T y \quad \hat{y} = X \hat{\beta} = X (X^T X)^+ X^T y \text{ is } y\text{'s}$$

④

orthogonal ~~proj~~ projection

$$3.4. \quad \begin{aligned} E_0(\vec{w}) &= \frac{1}{2} \sum_{n=1}^N \left(w_0 + \sum_{i=1}^D w_i (x_{ni} + \epsilon_{ni}) - t_n \right)^2 \\ &= \frac{1}{2} \sum_{n=1}^N \left(w_0 + \sum_{i=1}^D w_i x_{ni} - t_n \right)^2 + \sum_{n=1}^N \sum_{i=1}^D w_i \epsilon_{ni} \left(w_0 + \sum_{i=1}^D w_i x_{ni} - t_n \right) \\ &\quad + \frac{1}{2} \sum_{n=1}^N \sum_{i=1}^D \sum_{j=1}^D w_i w_j \epsilon_{ni} \epsilon_{nj} \end{aligned}$$

$$\begin{aligned} E[E_0(\vec{w})] &= \frac{1}{2} \sum_{n=1}^N \left(\sum_{i=1}^D w_i x_{ni} - t_n \right)^2 + \frac{1}{2} \sum_{n=1}^N \sum_{i=1}^D w_i^2 \sigma^2 \\ &= \tilde{E}_0 + \frac{1}{2} \sigma^2 \sum_{i=1}^D w_i^2 \end{aligned}$$

equivalent to noise-free error plus a regularization term

$$3.5 \quad \textcircled{1} \quad \frac{1}{2} \sum_{i=1}^N (t_i - w^T \phi(x_i))^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j|^2$$

$$\textcircled{2} \quad \begin{cases} \frac{1}{2} \sum_{i=1}^N (t_i - w^T \phi(x_i))^2 \\ \text{s.t.} \quad \sum_{j=1}^M |w_j|^2 \leq \eta \end{cases}$$

$$\textcircled{3} \quad J = \frac{1}{2} \sum_{i=1}^N (t_i - w^T \phi(x_i))^2 + \frac{\lambda'}{2} \left(\sum_{j=1}^M |w_j|^2 - \eta \right)$$

λ' is Lagrange multiplier
 $\lambda' \geq 0$

it has same expression with $\textcircled{1}$

when $\lambda \nearrow, \eta \downarrow$

$$\eta = \sum_{i=1}^M |w_i^*(\lambda)|^2$$

$$2.3 \quad \cancel{P(d \leq x) = \left(\frac{1}{1+x^p} \right)^N}$$

d = distance from origin to closest data point

one point's possibility that distance bigger than x is $1 - x^p$, $\therefore N$ points

$$\therefore P(d > x) = (1 - x^p)^N$$

$$P(d \leq x) = 1 - (1 - x^p)^N = \frac{1}{2}$$

$$x = \left(1 - \frac{1}{2^{\frac{1}{N}}} \right)^{\frac{1}{p}}$$

$$2.9 \quad E \left[\frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - \tilde{\beta}^T \tilde{x}_i) \right] = E(\tilde{y}_1 - \tilde{\beta}^T \tilde{x}_1) \quad \text{because } (\tilde{x}_i, \tilde{y}_i) \text{ are iid.}$$

$$= E \left[\frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - \tilde{\beta}^T \tilde{x}_i) \right] := R_{te}'(\tilde{\beta})$$

now we compare $E R_{tr}(\tilde{\beta})$ and $E R_{te}'(\tilde{\beta})$ where they have same size n

$$E R_{tr}(\tilde{\beta}) = E R_{te}'(\tilde{\beta}) \quad \text{where } \tilde{\beta} \text{ is trained on test set. because the two sets have same distribution}$$

And $E R_{te}'(\tilde{\beta}) \leq E R_{te}(\tilde{\beta}) = E R_{te}(\hat{\beta})$, because $\tilde{\beta}$ is least square fitted

$$\therefore E R_{tr}(\tilde{\beta}) \leq E R_{te}(\hat{\beta})$$