High-dimensional data analysis, Final exam: **[+38 points]** +36

Name 宋柏元

- Not only answer but also calculation
- Derivations must be clear

+12 1. [+13] Let $\Im = \{(x_1, y_1), \ldots (x_N, y_N)\}$ be training data with $y_i = f(x_i) + \varepsilon_i$, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$.

Let $\{(x_1, Y_1^0), \ldots (x_N, Y_N^0)\}$ be test data with $Y_i^0 = f(x_i) + \varepsilon_i^0$, $E(\varepsilon_i^0) = 0$, $Var(\varepsilon_i^0) = \sigma^2$.

Let $\hat{f}(\cdot)$ be an estimate based on $\Im$.

+1 (i) [+1] Define training error ($\overline{err}$) under the squared error loss.

$$\overline{err} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2$$

+1 (ii) [+2] Define in-sample error (Err_{in}) under the squared error loss. (explain the meaning of "expectation" in your formula).

$$Err_{in} = \frac{1}{N} \sum_{i=1}^{N} Err(x_i), \quad \text{where } Err(x_0) = E_{\Im}\left\{ (Y_0^0 - \hat{f}(x_0))^2 \mid X = x_0 \right\}$$

(-1)

→ ~~expectation~~ conditional ~~over~~ given $\Im$

+3 (iii) [+3] Define the average optimism bias (ω). (explain the meaning of "expectation" in your formula).

$$\omega = E_{\Im}(op) = E_{\Im}(Err_{in} - \overline{err}) = E_{\Im}(Err_{in}) - E_{\Im}(\overline{err})$$

→ expectation over the distribution of $\Im$

+5 (iv) [+5] Assume that the test data and training data are independent.

Derive the relationship between ω and $Cov(y_i, \hat{f}(x_i))$, $1, \ldots, N$.

$$\omega = E_{\Im}(op) = E_{\Im}(Err_{in} - \overline{err}) = \frac{1}{N} \sum_{i=1}^{N} \left(E((Y_i^0 - \hat{f}(x_i))^2) - E((y_i - \hat{f}(x_i))^2) \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left((E(Y_i^{0\,2}) - 2E(Y_i^0 \hat{f}(x_i)) + \cancel{E(\hat{f}(x_i)^2)}) - (E(y_i^2) - 2E(y_i \hat{f}(x_i)) + \cancel{E(\hat{f}(x_i)^2)}) \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\cancel{Var(Y_i^0)} + \cancel{E(Y_i^0)^2} - 2E(Y_i^0) E(\hat{f}(x_i)) - \cancel{Var(y_i)} - \cancel{E(y_i)^2} + 2E(y_i \hat{f}(x_i)) \right)$$

$$= \frac{1}{N} \sum_{i=1}^{N} 2\left(E(y_i \hat{f}(x_i)) - E(y_i) E(\hat{f}(x_i))\right)$$

$$= \frac{2}{N} \sum_{i=1}^{N} cov(y_i, \hat{f}(x_i))$$

+2 (v) [+2] Is $Cov(y_i, \hat{f}(x_i))$ negative or positive? Why?

(LS) $cov(y, \hat{y}) = cov(y, X(X^TX)^{-1}X^Ty)$

$= cov(y, y) X(X^TX)^{-1}X^T$

$= \sigma^2 X(X^TX)^{-1}X^T$

$= \sigma^2 H$

$\therefore cov(y_i, \hat{y}_i) = \sigma^2 h_{ii} > 0$

($\hat{f}(x_i)$ is based on $\Im$

$\therefore$ It has positive relationship between $\hat{f}(x_i)$ and y_i)

+10 2. [+11] Consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}^T = (y_1, \dots, y_N)$, $\boldsymbol{\varepsilon} \sim N_N(\mathbf{0}, \sigma^2 I_N)$,

$$\mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$$ is an ($N \times p$)-design matrix, and $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$.

$$\hat{\beta} = (X^TX + \lambda I)^{-1} X^T y$$
$$\hat{y}_i = x_i^T (X^TX + \lambda I)^{-1} X^T y$$

+2 (i) [+2] Define an estimate $\hat{y}_i$ of y_i based on the ridge estimator.

✓ $\hat{y}_i = x_i^T (X^TX + \lambda I)^{-1} X^T y$

+3 (ii) [+4] Express $\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$ in terms of the degree of freedom.

$$\sum_{i=1}^{N} cov(\hat{y}_i, y_i) = tr\begin{pmatrix} cov(\hat{y}_1, y_1) & \cdots & cov(\hat{y}_1, y_N) \\ \vdots & & \\ cov(\hat{y}_N, y_1) & & cov(\hat{y}_N, y_N) \end{pmatrix} = tr(cov(\hat{y}, y))$$

$$= tr(cov(X(X^TX+\lambda I)^{-1}X^Ty, y))$$

$$= tr(X(X^TX+\lambda I)^{-1}X^T \overset{\sigma^2 I}{cov(y, y)})$$

✓ $$= \sigma^2 tr(X(X^TX+\lambda I)^{-1}X^T)$$

$$= \sigma^2 df(\lambda)$$

+2 (iii) [+2] Write the average optimism bias by using the degree of freedom.

$$w = \frac{2}{N}\sum_{i=1}^{N} (cov(y_i, \hat{y}_i)) = \frac{2\,df(\lambda)}{N}\sigma^2$$ ✓

+3 (iv) [+3] Define a cross-validation $CV(\hat{f}_\lambda)$ and explain how to estimate the shrinkage parameter.

✓ $$CV(\hat{f}_\lambda) = \sum_{i=1}^{N} \left(y_i - \hat{f}_\lambda^{(-i)}(x_i)\right)^2$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\ CV(\hat{f}_\lambda)$$

Find the parameter λ s.t. $CV(\hat{f}_\lambda)$ is minimized

$\delta_2(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}$

$e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}$

+14 **3. [+14]** Let G be a class ($G = 1$ or 2) and $X = (X_1, \ldots, X_p)$ be inputs. Assume $X \mid G = k \sim N(\mu_k, \Sigma)$ and $\pi_k = \Pr(G = k)$ for $k = 1, 2$.

+3 (i) [+3] Write down $\log \frac{\Pr(G=2 \mid X=x)}{\Pr(G=1 \mid X=x)}$. (simplify the formula by using $\mu_2 - \mu_1$)

$$\log \frac{\Pr(G=2 \mid X=x)}{\Pr(G=1 \mid X=x)} = \log \frac{\delta_2(x) \cdot \pi_2}{\delta_1(x) \cdot \pi_2} = \log \frac{\pi_2}{\pi_1} + \boxed{\log \frac{\delta_2(x)}{\delta_1(x)}}$$

$$= \log \frac{\pi_2}{\pi_1} - \frac{1}{2}\left((x-\mu_2)^T \Sigma^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)$$

$$= \log \frac{\pi_2}{\pi_1} - \frac{1}{2}\left(\cancel{x^T \Sigma^{-1} x} - 2x^T \Sigma^{-1}\mu_2 + \mu_2^T \Sigma^{-1}\mu_2 - \cancel{x^T \Sigma^{-1} x} + 2x^T \Sigma^{-1}\mu_1 - \underline{\mu_1^T \Sigma^{-1}\mu_1}\right) \quad \underline{\underline{-\mu_2^T \Sigma^{-1}\mu_1 + \mu_2^T \Sigma^{-1}\mu_1}}$$

$$= \log \frac{\pi_2}{\pi_1} + x^T \Sigma^{-1}(\mu_2 - \mu_1) - \frac{1}{2}(\mu_2 + \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1)$$

+1 (ii)[+1] Define a linear discriminant function $\delta_k(x)$ s.t. x belongs class 2 if $\delta_2(x) > \delta_1(x)$. (assuming all parameters are known)

$$\delta_k(x) = x^T \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + \log \pi_k$$

We have 6 gene expressions from 3 patients as follows:

	Prognosis	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
Patient 1	Poor (class 2)	1	0	2	1	1	0
Patient 2	Good (class 1)	0	1	-1	-2	1	-1
Patient 3	Good (class 1)	-1	-1	-1	1	-2	1

+3 (iii) [+3] Calculate

$\frac{1}{3\times 6} \sum_{i=1}^{N} \sum_{j=1}^{p}$

$\hat{\pi}_1 = \frac{2}{3}$ $\quad \hat{\pi}_2 = \frac{1}{3}$ $\quad \hat{\mu}_1 = \begin{pmatrix} -\frac{1}{2} \\ 0 \\ -1 \\ -\frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{pmatrix}$ $\quad \hat{\mu}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 1 \\ 1 \\ 0 \end{pmatrix}$ $\quad \hat{\Sigma} = \begin{pmatrix} 4/3 & & & & & 0 \\ & 4/3 & & & & \\ & & 4/3 & & & \\ & & & 4/3 & & \\ & & & & 4/3 & \\ 0 & & & & & 4/3 \end{pmatrix}_{6\times 6}$

where $\hat{\Sigma} = \hat{\sigma}^2 I$ and $\hat{\sigma}^2 = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2$ and $\bar{x} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij}$.

+3 (iv) [+3] A new patient $x = (x_1, x_2, x_3, x_4, x_5, x_6)$ belongs to class 2 if

$\underline{\frac{9}{8}(x_1 + 2x_3 + x_4 + x_5)}$ $>$ $\underline{\frac{63}{32} + \log 2}$

↑ a linear function of $(x_1, x_2, x_3, x_4, x_5, x_6)$ ↑ constants

+2 (v) [+2] Which genes are useless for prognosis? Why?

Gene 2 and 6, the coefficients of x_2 and x_6 are zeros

∴ No matter how x_2, x_6 change, they are useless for prognosis.

+2 (vi) [+2] Which gene is the most useful for prognosis? Why?

Gene 3. The coefficient size is the largest of all

It is more sensitive than the others