Exercise 7.4

$$\mathrm{Err}_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y_0}\left[(Y_i^0 - \hat{f}(x_i))^2\right]$$

$$\overline{\mathrm{err}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 \qquad y = f(x) + \varepsilon$$

$$E_Y[op] = \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y)$$

$$E_Y[op] = E_Y\left[\frac{1}{N}\sum_{i=1}^{N}\left\{E_{Y_0}\left[(Y_i^0 - \hat{f}(x_i))^2\right] - (y_i - \hat{f}(x_i))^2\right\}\right]$$

$$= E_Y\left[\frac{1}{N}\sum_{i=1}^{N}\left\{E_{Y_0}[Y_i^{0\,2}] - 2E_{Y_0}[Y_i^0]\hat{f}(x_i) + \cancel{\hat{f}(x_i)^2} + {\color{red}+E[\hat{f}(x)] - E[\hat{f}(x)]} - y_i^2 + 2E_Y[y_i\hat{f}(x_i)] - \cancel{\hat{f}(x_i)^2}\right\}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E_Y\left[\cancel{E_{Y_0}[Y_i^{0\,2}]}\right] - 2E_{Y_0}[Y_i^0]E_Y[\hat{f}(x_i)] - \cancel{E_Y[y_i^2]} + 2E_Y[y_i\hat{f}(x_i)]$$

- $E_Y\left[E_{Y_0}[Y_i^{0\,2}] - f(x)^2 + f(x)^2\right] = \sigma_\varepsilon^2 + f(x)^2$
- $E_Y\left[y^2 - f(x)^2 + f(x)^2\right] = \sigma_\varepsilon^2 + f(x)^2$
- $E_{Y_0}[Y_i^0] = f(x) = E_Y[y]$ $\qquad \hat{f}(x) = \hat{y}$

$$E_Y[op] = \frac{1}{N}\sum_{i=1}^{N}\left\{-2E_Y[y]E_Y[\hat{y}] + 2E_Y[y_i\hat{y}_i]\right\}$$

$$= \frac{2}{N}\sum_{i=1}^{N}\mathrm{Cov}(y_i, \hat{y}_i)$$

Ex 7.5

For $\hat{y} = Sy$ show that $\sum_{i=1}^{N} Cov(y_i, \hat{y}_i) = trace(S)\sigma^2$

$$\sum_{i=1}^{N} Cov(y_i, \hat{y}_i) = trace(Cov(y, \hat{y})) \stackrel{!}{=} trace(Cov(y, Sy)) \stackrel{!}{=} trace(S\, Cov(y, y)) \stackrel{!}{=} trace(S\, Var(y)) \stackrel{!}{=} trace(S\sigma^2) \stackrel{!}{=} trace(S)\sigma^2$$

Cov matrix
$$\begin{pmatrix} Cov(y_1,y_1) & Cov(y_1,y_2) & \cdots & Cov(y_1,y_n) \\ Cov(y_2,y_1) & Cov(y_2,y_2) & & \vdots \\ \vdots & & \ddots & \\ Cov(y_n,y_1) & \cdots & & Cov(y_n,y_n) \end{pmatrix}$$

$\sum_{i=1}^{rank(M)} m_{ii} = trace\ M$

Exercise 7.7

Use the approximation: $\frac{1}{(1-x)^2} \approx 1 + 2x$

to show similarities between C_p/AIC and GCV

$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}_\varepsilon^2$ AIC $= -\frac{2}{N}\, loglik + 2\frac{d}{N}$

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

$$C_p = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + 2\frac{d}{N}\hat{\sigma}_\varepsilon^2$$

$trace(S) = \frac{\sum Cov(y_i, \hat{y}_i)}{\sigma^2}$

$\hat{\sigma}^2 = \frac{1}{N-1}\sum(y_i - \hat{f}(x_i))^2$

$$GCV = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 \cdot \left(\frac{1}{1 - \frac{trace(S)}{N}}\right)^2$$
$$\stackrel{!}{=} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 \left(1 + 2\cdot\frac{trace(S)}{N}\right)$$
$$\stackrel{!}{=} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + \frac{2}{N}\, trace(S)\, \frac{\sum(y_i - \hat{f}(x_i))^2}{N} = \overline{err} + \frac{2}{N}\, trace(S)\, \hat{\sigma}_\varepsilon^2$$

Gaussian regression

$$AIC \propto -\frac{2}{N}\log\left(\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2\right\}\right) + \frac{2d}{N}$$
$$\stackrel{!}{=} + \frac{1}{N\sigma_\varepsilon^2}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + \frac{2d}{N}$$
$$\propto \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 + \frac{2d}{N}\sigma_\varepsilon^2$$

$d \leftarrow trace(S)$

Bootstrap methods

→ what is bootstrap
→ how to use bootstrap for error estimation $\widehat{E[Err_\tau]}$

IDEA: generate bootstrap sample from the empirical distribution computed on original sample
→ by resampling with replacement from the original sample

- suppose $\tau = \{ \overset{z_1}{(x_1, y_1)}, \dots, \overset{z_n}{(x_n, y_n)} \}$
- by resampling, $\tau_1^* = \{ (x_1^*, \overset{z_1^*}{y_1^*}), \dots, \overset{z_n^*}{(x_n^*, y_n^*)} \}$
- repeat for B large, $\tau_1^*, \tau_2^*, \dots, \tau_B$

Based on the generated bootstrap sample (which mimic new experiments) we can estimate any aspect of the distribution of a map

Example

original sample $\tau = \{ z_1, z_2, z_3, z_4 \} = \{ 1, 3, 4, 6 \}$

generate B bootstrap sample by resampling with replacement from τ

$\tau_1^* = \{ z_1^*, z_2^*, z_3^*, z_4^* \}$
$= \{ 1, 4, 1, 6 \}$

$\tau_2^* = \{ 4, 4, 3, 3 \}$

$\vdots$

$\tau_B^* = \{ 1, 1, 1, 1 \}$

(x, y) $\quad Cov(x, y)$

τ $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$
$(4; 3), (1; 3), \dots, (4; 2)$

$\tau_1^* = \{ (1;3), (4;2), \dots, (4;2) \}$

Bootstrap approach for prediction error estimation

WRONG APPROACH
- estimate our $\hat{f}(x)$ from each bootstrap sample
- evaluate how well $\hat{f}^*_b(x)$ estimate y

$$\widehat{Err}_b^{WRONG} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{N}\sum_{i=1}^{N} L(y, \hat{f}^*_b(x_i))\right)$$

⚠ : training and test set are $\underline{\underline{\text{not}}}$ independent

$E[\widehat{Err}_b^{WRONG}] = 0.184$, 1NN , $Y \perp X$

- $y_i \in \mathcal{T}^*_b \longrightarrow$ error $= 0$
- $y_i \notin \mathcal{T}^*_b \longrightarrow$ error $= 0.5$

$$= 0.5 \times \underbrace{Pr[y_i \notin \mathcal{T}^*_b]}_{0.368} + 0 \times Pr[y_i \in \mathcal{T}^*_b]$$

$Pr[$observation $i \notin$ bootstrap sample $b]$

$Pr[\mathcal{T}^*_{b[1]} \neq y_i] = \dfrac{N-1}{N}$ $\Rightarrow Pr[y_i \notin \mathcal{T}^*_b] = \left(\dfrac{N-1}{N}\right)^N = e^{-1} \approx 0.368$

same for all positions

$$E[\widehat{Err}_b^{WRONG}]\Big|_{\substack{X \perp Y \\ 1NN}} \approx 0.5 \times e^{-1} = 0.184$$

An important fact

$Pr[$observation i belongs to a bootstrap sample $b] = 1 - e^{-1} \approx \boxed{0.632}$

approximation of $1 - \left(\frac{N-1}{N}\right)^N$

CORRECT APPROACH

NB.: bootstrap sample has the same size of the original sample

$\mathcal{T} = \{z_1, \dots, z_N\}$

$\mathcal{T}^*_b = \{z^*_1, \dots, z^*_N\}$ resampling with replacement

→ there are original observations which are included more than once

⇒ there are also original observations which are not included at all

↳ these can be used as a test set as they are not used in the training process.

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^*_b(x_i))$$

where $|C^{-i}|$ is the number of bootstrap samples that do NOT contain i

Issues

→ the average number of unique observations in the training set is $0.632N$ → not so far from $0.5N$, that is the value related to 2-fold CV

→ similar issues of training-set-size bias than 2-fold CV

→ result in a small overestimation of the error

To solve the issue, the .632 estimator has been developed

$$\widehat{Err}^{(.632)} = 0.368\,\overline{err} + 0.632\,\widehat{Err}^{(1)}$$

In general, it works well, but in some case it fails, like in our 1NN $X \perp Y$

$$\overline{err} = 0 \rightarrow \widehat{Err}^{(.632)} = 0.632\,\widehat{Err}^{(1)}$$

Further improvements ".632+ estimator"

- based on the quantity γ, the no-information-error rate

error that we obtain when inputs and class label are independent

$\hat{\gamma}$ is computed by permuting x and y separately, we compute the prediction error for each combination of y_i and x_j

$$\hat{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N} L(y_i, \hat{f}(x_j))$$

- $\hat{\gamma}$ is used to compute the overfitting rate

$$\hat{R} = \frac{\widehat{Err}^{(1)} - \overline{err}}{\hat{\gamma} - \overline{err}} \qquad 0 \leq R \leq 1$$

no overfitting

- finally

$$\widehat{Err}^{(.632+)} = (1-\hat{\omega})\,\overline{err} + \hat{\omega}\,\widehat{Err}^{(1)}$$

where $\omega = \dfrac{0.632}{1-0.368\hat{R}}$

Generalized Additive Models

- extensions of the (generalized) linear model

Linear model
- powerfull tool
- can be used in several cases (regression, classification, ...)

Main limitation
- it suppose linear effects, often not true in reality
  ($\hat{\beta}$ is the increments in y when the corresponding x increase of 1)

In the context of regression, the (generalized) additive model has the form

$$E[Y|x_1,\dots,x_p] = \alpha + f_1(x_1) + \dots + f_p(x_p)$$

where
- Y is the outcome
- x_j are the predictors
- f_j is a function which describe the effect of x_j

- we already saw that we can use $f_j(x_j) = x_j^2$, $f_j(x_j) = \log x_j$
- we can be more general, and use a nonparametric function (splines, kernel, ...)

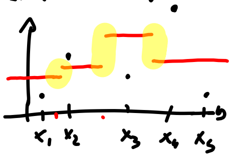
splines $\rightarrow$ § 5.2

kernel $\rightarrow$ § 6.1, § 6.2

- cubic splines $\rightarrow$ bottom right of Fig 5.2
- natural splines $\rightarrow$ since the estimation outside the observations' range can be dangerous, the line is forced to be linear outside the range

Kernel method

- extension of k-NN

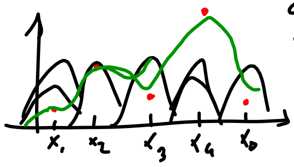


2-NN
- use the y_i of the 2 nearest neighbours
- costant
- ugly and unecessary discontinuities

kernel method
- weights through a kernel $D\left(\frac{|x - x_0|}{h_\lambda(x_0)}\right)$

different kinds of kernels

h, λ are parameters
h controls the shape
λ the width of the kernel



GAM

for functional effect, we can

- $E[Y|X] = \alpha + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \dots + \beta_p f_p(x_p)$

  $\log(x_i)$

  → the least square estimator approach is usable

- $E[Y|X] = \alpha + f_1(x_1) + \dots + f_p(x_p)$

  → backfitting algorithm

Generalized Additive Model

↳ extending the GLM (STK 3100)

GLM $g(\mu(x)) = \alpha + \beta^T X$ — extending the linear model to all exponential families sampling models

g ↳ link function

e.g. logistic model

$g() = \text{logit}$ $\qquad \mu(x) = \Pr[Y = 1 | X = x]$

$\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p$

GAM $g(\mu(x)) = \alpha + \sum_{j=1}^{p} f_j(x_j)$

additive logistic regression : $\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = \alpha + \sum_{j=1}^{p} f_j(x_j)$

Advantages of GAM:
- flexibility, due to f (we can capture non-linear effects)
- interpretability, due to the additivity (not so different from the usual interpretation of GLM)

Note: not all effect need to be non-linear / linear

- semiparametric model $g(\mu(x)) = \underbrace{X^T\beta}_{\text{parametric}} + \underbrace{f(z)}_{\text{non-parametric}}$

  e.g. semiparametric model : Cox model

  $\lambda(t) = \underbrace{\lambda_0(t)}_{\text{non-parametric}} \underbrace{\exp(X^T\beta)}_{\text{parametric part}}$