CompSci, Nov 23, 2022

- statistical interpretation
- Resampling techniques

$$y(x) = f(x) + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$f(x)$ is a deterministic function

$$\mathcal{D} = \{(x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$$

ideally we have PDF $p(x)$, $p(y)$ etc.

$$\mathbb{E}[y] = \int_{D} p(y)\, y\, dy = \mu_y$$

$$\left(\sum_{i \in D} p(y_i) y_i \right)$$

$$\mathrm{var}[y] = \int_D p(y)(y - \mu_y)^2$$

$$\mathrm{cov}[y, y'] = \int_D p(y, y')(y - \mu_y) \times (y' - \mu_{y'})\, dy\, dy'$$

Sample mean:

$$\bar{\mu}_y = \frac{1}{n} \sum_{i=0}^{n-1} y_i \neq \mu_y$$

sample variance

$$var[\bar{y}] = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \bar{\mu}_y)^2$$

$$\neq var[\sigma_y^2]$$

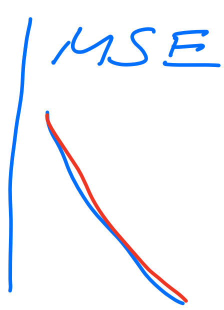
$$\mathbb{E}[y] = \frac{1}{n} \sum_{i=0}^{n-1} y_i$$

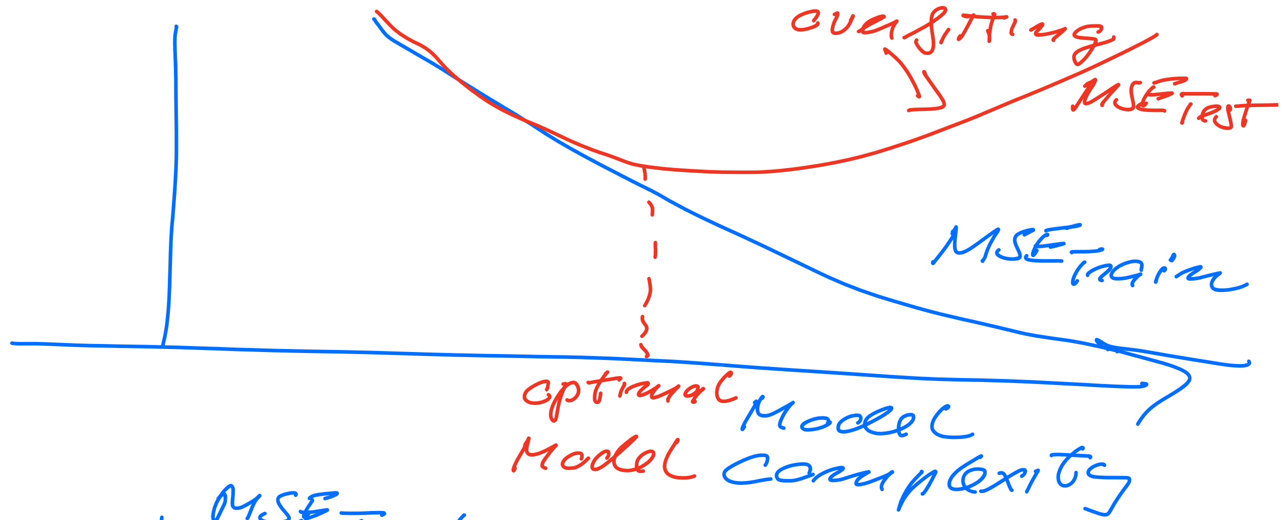
$$MSE = \mathbb{E}[(y - \tilde{y})^2] = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

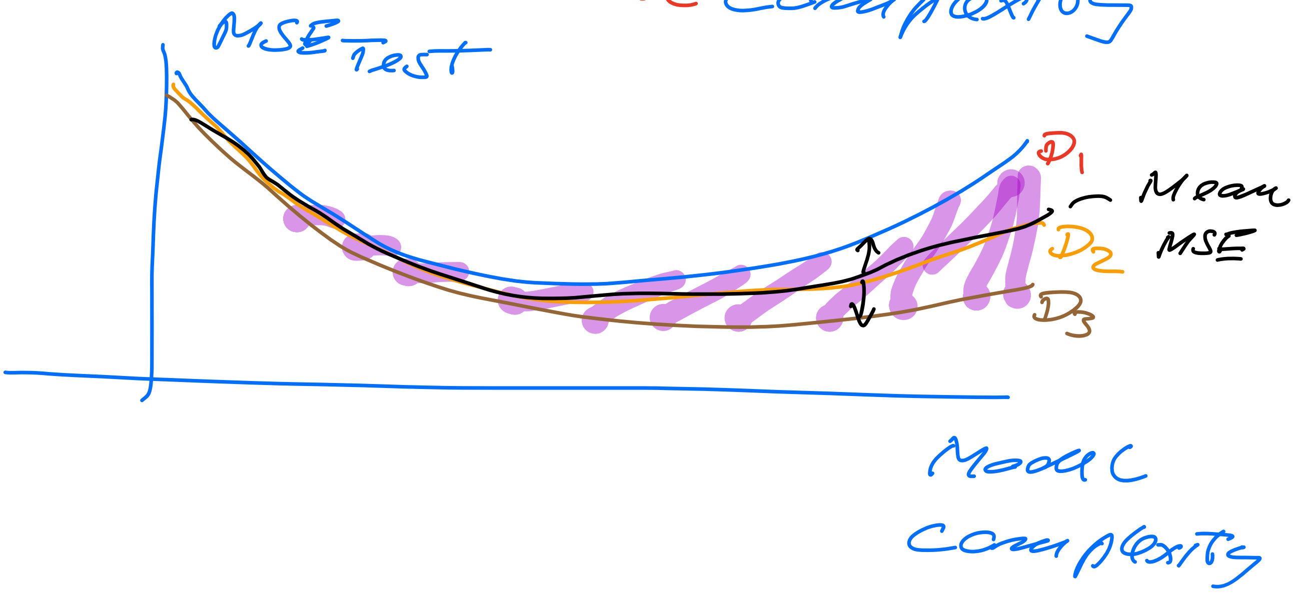
Resampling techniques

- Bootstrap
- Cross-validation

Look at MSE as example







Bootstrap strategy

Require $D = \{(x_0, y_0) \dots (x_{n-1}, y_{n-1})\}$

Split in train and test data.

Require $M = \#$ bootstraps samples

Have defined D_{train} and

D_{Test}

For $i = 1, M$

- Make $D_{Train}(i)$ by randomly selecting with replacement.

$$\begin{cases} D_{Train} = [1, 2, 3, 4, 5] \\ D^*_{Train} = [2, 3, 2, 4, 1] \end{cases}$$

- Train Model (i)
- compute $MSE(i)$ (on train) and $MSE_{Test}(i)$

end loop

- compute

$$MSE_{Test} = \frac{1}{M} \sum_{i=0}^{M-1} MSE_{Test}(i)$$

Cross-validation

Define folds = K

Example K = 5

T = Train

T = Test

D_1:	T	T	T	T	**T**	$MSE_{Test}(1)$
D_2:	T	T	**T**	T	T	$MSE_{Test}(2)$
D_3:	T	T	T	**T**	T	⋮
D_4:	T	**T**	T	T	T	⋮
D_5:	**T**	T	T	T	T	$MSE_{Test}(5)$

$$MSE_{Test} = \frac{1}{K} \sum_{i=0}^{K-1} MSE_{Test}^{(i)}$$

---

statistics

$$y(x_i) = y_i = f(x_i) + \varepsilon_i$$

$$\mathbb{E}[y_i] = \mathbb{E}[f(x_i)] + \mathbb{E}[\varepsilon_i]$$

$$\mathbb{E}[f(x_i)] \overset{?}{=} f(x_i), \quad \mathbb{E}[\varepsilon_i] = 0$$

$$f(x_i) \approx \sum_{j=0}^{p-1} x_{ij}\beta_j = X_{i*}\beta$$

$$\mathbb{E}[y_i] = X_{i*}\beta \Rightarrow \mathbb{E}[y] = X\beta$$

$$\mathrm{var}[y] = \sigma^2 = \mathrm{var}[\varepsilon]$$

(exercise)

$$\mathbb{E}[\beta] = \beta$$

y has mean value $X\beta$ and variance σ^2

$$y \sim N(X\beta, \sigma^2) \Rightarrow$$

$$y_i \sim N(X_{i*}\beta, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - X_{i*}\beta)^2}{2\sigma^2}} = p(y_i|\beta)$$

assume y_i are i.i.d.

$$P(D \mid \beta) = \prod_{i=0}^{n-1} P(y_i \mid \beta)$$

$$P(y_i \mid \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i \beta)^2}{2\sigma^2}}$$

$\sim -\frac{1}{2} \log$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\arg\max}\ P(D \mid \beta)$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\arg\max}\ \log P(D \mid \beta)$$

Or

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^P}{\arg\min}\left(-\log P(D \mid \beta)\right)$$

$$= \sum_{i=0}^{n-1} \left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{(y_i - x_i \beta)^2}{2\sigma^2}\right]$$

$$= \frac{n}{2}\log(2\pi\sigma^2) + \left\| \frac{(y - X\beta)}{2\sigma^2} \right\|_2^2$$

Taking derivatives wrt β

$$\Rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T \beta$$

$$\left(MSE = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - x_{i*}\beta)^2 \right.$$

$$= \frac{1}{n} \| (y - X\beta) \|_2^2$$

Ridge & Lasso ?

we are going to make an ansatz (prior) about the distribution of β,

Ridge $\quad p(\beta) = \frac{1}{\sqrt{2\pi \gamma^2}} e^{-\beta^2/2\gamma^2}$

Lasso $\quad p(\beta) \simeq e^{-|\beta|/2\gamma}$

(Laplace distribution)

Bayes' theorem,

- product rule of probabilities

$$p(A, B) = p(A|B)\, p(B)$$

$$= P(B|A)P(A)$$

- conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \text{if } P(B) > 0$$

$$P(B|A) = \frac{P(A,B)}{P(A)} \quad P(A) > 0$$

if A and B are i.i.d

$$P(A,B) = P(A)P(B)$$

- Marginal distribution

$$P(A) = \sum_b P(A|B=b)P(B=b)$$

$$P(B) = \sum_a P(B|A=a)P(A=a)$$

Combining we have Bayes' theorem

$$P(A|B) = \frac{P(A,B)}{P(B)} =$$

$$\frac{P(B|A)\,P(A)}{\sum_a P(B|A=a)\,P(A=a)}$$

likelihood prior

posterior

what is optimal β given D

$$P(\beta|D) \propto P(D|\beta)\,P(\beta)$$

$$N(X\beta, \sigma^2)$$