FYS-STK4155 September 16

Resampling and statistical analysis

Central limit theorem

a set of data of i.i.d. events,

$$\mathbb{E}[x_i] = \bar{x}_i = \sum p(x_i) x_i$$

simplicity $\bar{x}_i \rightarrow x_i$

$$z = \frac{x_1 + x_2 + \dots x_m}{m}$$

x_i they follow $p(x)$

what form does $P(z)$ take?

$$P(z) = \int dx_1 p(x_1) \int dx_2 \, p(x_2) \dots \int dx_m \, p(x_m) \times \delta\left(z - \frac{x_1 + x_2 + \dots + x_m}{m}\right)$$

$$\delta\left(z - \frac{x_1 + x_2 + \dots + x_m}{m}\right)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} dq \, \exp\left(iq\left(z - \frac{x_1 + \dots x_m}{m}\right)\right)$$

$$E[x_i] = \mu$$

insert $e^{i\mu q - i\mu q}$

$$P(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dq \, e^{iq(z-\mu)}$$

$$\times \left[\int_{-\infty}^{\infty} dx \, p(x) \, e^{iq\frac{(\mu - x)}{m}} \right]^m$$

$$\int_{-\infty}^{\infty} dx \, p(x) \, e^{iq(\mu - x)/m}$$

$$= \int_{-\infty}^{\infty} dx \, p(x) \left[1 + iq\frac{(\mu - x)}{m} - \frac{q^2(\mu - x)^2}{2m^2} + \dots \right]$$

$$\int dx \, p(x) \, x = \mu \quad \wedge \quad \int dx \, p(x) = 1$$

$$= \left[1 - \frac{q^2 \sigma^2}{2m^2} + \dots \right]$$

$$p(z) = \frac{1}{\sqrt{2\pi \sigma^2/m}} e^{-\frac{(z-\mu)^2}{2\sigma^2/m}}$$

$$var[x] = \int p(x)(x-\mu)^2 dx$$
$$= \sigma^2$$

$$var[z] = \sigma^2/m \Rightarrow$$
$$stdv = \sigma/\sqrt{m}$$

Bootstrap Resampling

original sample

$$D = \{ z_0, z_1, \dots z_{n-1} \}$$

(i) Draw a bootstrap sample

$$D_1^* = \{ z_0^* z_1^*, \dots z_{n-1}^* \}$$

compute $\Theta_1^* = \frac{1}{n} \sum_{i=0}^{n-1} z_i^*$

(ii) Repeat B times yielding estimates

$\theta_1^* \quad \theta_2^* \quad \dots \quad \theta_B^*$

$$\bar{\theta} = \frac{1}{B} \sum_{j=1}^{B} \theta_j^*$$

(iii) can compute variance

$$s^2 = \frac{1}{B} \sum_{j=1}^{B} \left(\theta_j^* - \bar{\theta} \right)^2$$

(iv) output s^2 and $\bar{\theta}$

The bootstrap shows that it approaches the true mean and true variance and other expectation values.

Bias-variance tradeoff

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

$$= \text{Bias}\left[\left(y_i - \mathbb{E}[\tilde{y}]\right)^2\right]$$

$$+ \ \text{var}[\tilde{y}] + \sigma^2$$

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

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

$f(x)$ is deterministic.

$$f(x_i) \simeq \tilde{y}_i \qquad \mathbb{E}[f] = f$$

$$\mathbb{E}[\tilde{y}_i] = \tilde{y}_i \qquad = \mathbb{E}[y]$$

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

$$MSE = \mathbb{E}[(y - \tilde{y})^2]$$

if we know p

$$MSE = \int_D (y - \tilde{y})^2 p(x)\, dx$$

$$= \int_D (f + \varepsilon - \tilde{y})^2 p(x)\, dx$$

add and subtract

$\mathbb{E}[\tilde{y}]$

$$\int_D \left(f + \varepsilon - \tilde{y} + \mathbb{E}[\tilde{y}] - \mathbb{E}[\tilde{y}] \right)^2 p(x)\, dx$$

$$\mathbb{E}[\varepsilon] = 0 \qquad \mathbb{E}[\tilde{y}] = \mu_{\tilde{y}}$$

$$= \underbrace{\boxed{\int_D \left(f(x) - \mu_{\tilde{y}} \right)^2 p(x)\, dx}}_{\text{Bias}}$$

$$+ \underbrace{\boxed{\int_D \left(\tilde{y}(x) - \mu_{\tilde{y}} \right)^2 p(x)\, dx}}_{\text{var}[\tilde{y}]}$$

$$+ \ \text{var}[\varepsilon^2]$$

$$\sim \sigma^2$$

We have a discrete set and $p(x)$ is unknown $\Rightarrow$ sample expectation values (Reason for resampling)

$$MSE = \frac{1}{n} \sum_{i=0}^{n-1} \left(y_i - \tilde{y}_i \right)^2$$

$$= \frac{1}{n} \sum_{i=0}^{n-1} \underbrace{\left(y_i - \mathbb{E}[\tilde{y}] \right)^2}_{\text{Bias}}$$

$$+ \frac{1}{n} \sum_{i=0}^{n-1} \underbrace{\left(\tilde{y}_i - \mathbb{E}[\tilde{y}] \right)^2}_{\text{variance of model}}$$

$$+ \sigma^2$$

Done on test data.

<u>CV</u>

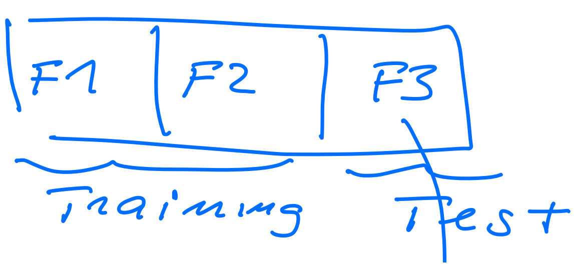
$X = [1, 2, 3, 4, 5, 6]$

$K = 3$ number of folds

Fold $\underline{1}$ $[5, 2]$

Fold 2 $[1, 3]$

Fold 3 $[4, 6]$



1) $Err_{Test} = Err_{F_3}$

2) $\boxed{\underbrace{F_1}_{Test} \mid \underbrace{F_2 \mid F_3}_{Training}}$

$Err_{Test} = Err_{F_1}$

3) $\boxed{\underbrace{F_1}_{Train} \mid \underbrace{F_2}_{Test} \mid \underbrace{F_3}_{Train}}$

$Err_{Test} = Err_{F_2}$

Total error $= \frac{1}{3}\left(Err_{F_1} + Err_{F_2} + Err_{F_3}\right)$

$K \sim 5-10$