FYS4411/9411, lecture March 14

FYS4411/9411 March 19

- Resampling/statistical analysis
    - Bootstrap (Jackknife)
    - Blocking

standard Bootstrap assumes that the stochastic variables X_i are i.i.d.! ← distributed

independent, identically

The estimators are the

mean value (sample mean) and the sample variance

$$\mathbb{E}[x] = \langle x \rangle = \int_{x \in D} P_X(x)\, x\, dx$$

unknown

PDF

$$\text{or} \quad \sum_{i \in D} x_i P_X(x_i)$$

$$X = \{x_0\ x_1\ \dots\ x_{n-1}\}$$

not knowing $P_X(x_i)$, we evaluate the sample mean

$$\mu = \frac{1}{n}\sum_{i=0}^{n-1} x_i \neq \mathbb{E}[X]$$

$$\left(\sum_{i=0}^{n-1} P_X(x_i)\, x_i\right)$$

$$\approx \frac{1}{n}$$

$$\int_{x\in D} P_X(x)\,dx = 1 \quad \vee \quad \sum_i P_X(x_i) = 1$$

$$\mathrm{var}[X] = \int_{x \in D} P_X(x)\,(x - E[X])^2\,dx$$

$$\left(\sum_{i \in D} P_X(x_i)\,(x_i - E[X])^2 \right)$$

$$\rightarrow \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

in our code we need

$$\mathrm{var}[X] \geq 0$$

standard deviation (STD)

$$= \sqrt{\mathrm{var}[X]}$$

sample variance

$$\sigma_x^2 = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \mu_x)^2$$

sample mean

$$= \frac{1}{n}\sum_{i=0}^{n-1} x_i^2 - (\mu_x)^2$$

$$\neq \mathrm{var}[x]$$

what does it mean that x_i is i.i.d.?

x_1 and x_2 are stochastic variables

$P(x_1 x_2) = ?$

identical distribution:

$$P(x_1, x_2) = \tilde{P}(x_1)\,\tilde{P}(x_2)$$

Ex: normal distribution

$$\tilde{P}(x_1) \sim e^{-x_1^2/2\sigma^2}$$

Meet covariance

$$\mathrm{cov}(x_i\, x_j) =$$

$$\int dx_i \int dx_j\, P(x_i\, x_j)$$

$$\cancel{\times}\ (x_i - \bar{x}_i)(x_j - \bar{x}_j)$$

$\uparrow$ mean of x_i

if $x_i\, x_j$ are i.i.d.

$$P(x_i, x_j) = \bar{P}(x_i)\,\bar{P}(x_j)$$

$$\bar{x}_i = \bar{x}_j = \bar{x}$$

$$\int dx_i' \int dx_j' \; \tilde{p}(x_i') \, \tilde{p}(x_j') \, (x_i' - \bar{x})(x_j' - \bar{x})$$

$$= \int dx_i' (x_i' - \bar{x}) \, \tilde{p}(x_i') \times \int dx_j' (x_j' - \bar{x}) \, \tilde{p}(x_j') = 0$$

$$\Big(\int dx_i' (x_i' - \bar{x}) \, \tilde{p}(x_i') = \underbrace{\int dx_i' \, x_i' \, \tilde{p}(x_i')}_{\bar{x}} - \bar{x} \underbrace{\int \tilde{p}(x_i') \, dx_i'}_{=1} \Big)$$

Bootstrap

original sample

$$X = \{x_1, x_2, \ldots x_n\}$$

we assume we can produce new samples X^* and produce new estimates for μ and σ^2,

$$\Theta_n^{*(1)}, \Theta_n^{*(2)}, \ldots \Theta_n^{*(B)}$$

$$\Theta_n^{*(i)} = \frac{1}{n} \sum_{j=1}^{n} X_j^{*(i)}$$

$$\overline{\epsilon}^*_B = \frac{1}{B} \sum_{i=1}^{B} \epsilon_n^{*(i)}$$

$$var[\overline{\epsilon}^*_B] = \frac{1}{B-1} \sum_{i=1}^{B} (\epsilon_n^{*(i)} - \overline{\epsilon}^*_B)^2$$

Bootstrap algorithm

original data $X = \{1, 2, 3, 4, 5\}$

X^* by shuffling randomly with replacement

$X^* = \{2, 3, 3, 5, 1\}$

Central limit theorem

n- experiments, each of these have

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_j^{(i)}$$

i.i.d.

simply $\bar{X}_i \rightarrow X_i$ mean of expt i

same PDF

Total mean

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

what is the distribution

$\bar{p}(z)$?

we know $p(x_i)$

$$\bar{p}(z) = \int_{x_1 \in D} dx_1\, p(x_1) \int_{x_2 \in D} dx_2\, p(x_2)$$

$$\dots \int_{x_m \in D} dx_m\, p(x_m)\, \delta\left(z - \frac{x_1 + x_2 + \dots x_m}{m}\right)$$

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

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} dq \, \exp\left[iq\left(z - \frac{(x_1 + x_2 \dots x_m)}{m}\right)\right]$$

$$\int p(x_1)\, x_1\, dx_1 = x_1'$$

$$= \mu$$

insert in $\bar{p}(z)$ $\underline{\underline{1}}$

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

$$\mu = \int_{x \in D} p(x)\, x\, dx$$

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

$$\times \underbrace{\left[\int_{x\in D} dx\, p(x) \exp\left[\frac{iq(\mu-x)}{m}\right]\right]}^{m}$$

$$\int_{x\in D} p(x)\, dx = 1$$

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

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

$$= \left[1 + 0 - \frac{q^2 \sigma^2}{2m^2} + \dots \right]$$

in the limit $m \Rightarrow \infty$

$$\bar{P}(z) = \frac{1}{\sqrt{2\pi\sigma^2/m}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2/m} \right]$$

Gaussian with mean value μ and variance σ^2/m $\Rightarrow$

$STD = \sigma/\sqrt{m}$ and

$$\sigma^2 = \int_{x \in D} p(x)(x-\mu)^2 dx$$

single experiment

variance.

Central limit theorem.