How do we find a prob. dist for a r.v. x?

3 steps

(1) choose a parametric model (eg Gaussian)

θ = parameters

(2) collect samples from r.v. x:

$D = \{x_1, \ldots, x_N\}$

We assume x_i are i.i.d. samples

(3) MLE (maximum likelihood principle)

The optimal parameter θ^* is that which maximizes the probability (likelihood) of the training data.

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)$$

"likelihood function"

likelihood of data w.r.t parameter θ

Note ① D is known, so $p(D|\theta)$ is a function of θ

It is not a prob. w.r.t θ

② $\log = \ln$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{\log p(D|\theta)}$$

$l(\theta) = \log.$ likelihood. function.

$$= -\underset{\theta}{\operatorname{argmin}}\ \underbrace{-\log p(D|\theta)}$$

negative log likelihood function (loss)

data LL

$$l(\theta) = \log p(D|\theta)$$

$$= \sum_{i=1}^{N} \log p_i(D_i|\theta)$$

assume independence (i.i.d.)

To get optimal MLE solution:

- if θ is a scalar, at local optimal

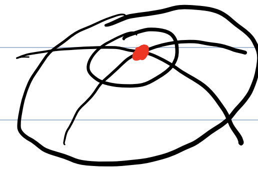
1) $\frac{\partial}{\partial\theta} \log p(D|\theta) = 0$ at θ^*

2) $\frac{\partial^2}{\partial\theta^2} \log p(D|\theta) < 0$ at θ^* (concave)

3> check the boundary condition of θ (if necessary)

- if θ is a vector

1> $\nabla_\theta \ell(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ell(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ell(\theta) \end{bmatrix} = 0$



2> Hessian Matrix

$$\nabla^2_\theta \ell(\theta) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \\ & \ddots & \\ & & \frac{\partial^2}{\partial \theta_p^2} \end{bmatrix} \ell(\theta) \prec 0 \text{ (negative definite)}$$

$H \prec 0$ negative definite: $\theta^T H \theta < 0, \forall \theta$

$\Rightarrow$ for all directions, "Concave" (mountain)

$H \succ 0$ positive definite: $\theta^T H \theta > 0, \forall \theta$

"convex" (bowl)

(semi-negative-definite) "ridge"

Ex. Bernoulli

$\theta = \pi \in \{0, 1\}$

$$\ell(\theta) = \sum_{i=1}^{N} \log p(x_i | \theta)$$

$$= \sum_{i=1}^{N} \log \left(\pi^{x_i} (1-\pi)^{(1-x_i)} \right)$$

$$= \sum_{i=1}^{N} x_i \log \pi + (1-x_i) \log (1-\pi)$$

$$= \underbrace{\left(\sum_{i=1}^{N} x_i \right)}_{\#\text{ of 1s}} \log \pi + \underbrace{\left(\sum_{i=1}^{N} (1-x_i) \right)}_{\#\text{ of 0s}} \log (1-\pi)$$

$m = \sum_{i=1} x_i \leftarrow$ "sufficient statistics" (a set of statistics) ... you need

$\Rightarrow$ the quantities you ... eg. $\sum x_i^2$ ...

$l(\theta)$ only depends on the N observations (dataset) through this value (these values)

$$= m \log \pi + (1-m) \log (1-\pi)$$

1) find max

$$\frac{\partial}{\partial \pi} l(\theta) = \frac{m}{\pi} + \frac{N-m}{1-\pi}(-1) = 0$$

$$\Rightarrow \quad \pi = \frac{m}{N} \qquad \hat{\pi} = \frac{m}{N} \text{ (sample mean)}$$

2) $$\frac{\partial^2}{\partial \pi^2} l(\theta) = -\frac{m}{\pi^2} - \frac{N-m}{(1-\pi)^2} < 0$$

3) boundary condition: $0 \le m \le N \quad 0 \le \frac{m}{N} \le 1$

Ex. Gaussian

① $\theta = \mu \quad (\alpha^2 \text{ known})$

$$l(\theta) = \sum_i \left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \alpha^2 - \frac{1}{2\alpha^2}(x_i-\mu)^2\right)$$

$$= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\alpha^2 - \frac{1}{2\alpha^2}\sum_i (x_i-\mu)^2$$

"sufficient statistics" $\Rightarrow$ $\{\sum_i x_i, \sum_i x_i^2\}$

1) $$\frac{\partial l(\theta)}{\partial \mu} = -\frac{1}{2\alpha^2}\sum_i 2(x_i-\mu)(-1) = 0$$

$$\Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1} x_i}{N}$$

② $\theta = \alpha^2$

$$\frac{\partial l(\theta)}{\partial \alpha^2} = -\frac{N}{2}\frac{1}{\alpha^2} - \frac{1}{\quad}(-1)\sum_i (x_i-\mu)^2 = 0$$

$$\hat{\alpha}^2 = \frac{1}{N}(X_i-\mu)^2 \text{ (Sample variance)}$$

Evaluation

Estimate (e.g. $\hat{\mu}$, $\hat{\alpha}_2$) is a number

Estimator is a r.v. (over possible datasets)

$$f(X_1, \dots, X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

(MLE is one of the ways to computing estimators)

r.v. for each sample

$X_i \sim p(X_i|\theta)$ true distribution

(assume X_i is drawn from some distribution)

The estimate is the value of the estimator for a given dataset D

$$\hat{\mu} = f(X_1, \dots, X_N)\big|_{X_1=x_1,\dots} : \frac{1}{N}\sum_i x_i$$

sample

Since the estimator is a r.v., we can derive the mean & variance to qualify "goodness"

Bias & Variance $\hat{\theta} = f(X_1, \dots, X_N)$

1) will it converge to the true value of θ?

$$\text{Bias}(\hat{\theta}) = E_{X_1\dots X_N}[\hat{\theta} - \underset{\text{true value}}{\theta}] = \underbrace{E_X[\hat{\theta}]}_{\text{mean of estimator}} - \theta$$

if the bias is non-zero, then we can never get the true value (even if infinite samples)

2) How long will it take to converge

(How many samples do we need)

$$Var(\hat{\theta}) = E_{x_1 \dots x_N}[(\hat{\theta} - E\hat{\theta})^2]$$

Ex. Gaussian.

Estimator $\hat{\mu} = \frac{1}{N}\sum_i x_i$

Mean of $\hat{\mu}$ $\quad E_{x_1 \dots x_N}[\frac{1}{N}\sum_i x_i] = \frac{1}{N} N \mu = \mu$

$Bias(\hat{\mu}) = 0$

Var of $\hat{\mu}$ $\quad E_{x_1 \dots x_N}[(\hat{\mu} - E\hat{\mu})^2]$

$$= E_{x_1 \dots x_N}[(\frac{1}{N}\sum x_i - E\hat{\mu})^2]$$

$(a+b)^2 = a^2 + ab + ba + b^2$

$$= \frac{1}{N^2} E((\sum_i (x_i - \mu))^2)$$

$(\sum_i x_i)^2 = \sum_i x_i \sum_j x_j$

$$= \frac{1}{N^2} E(\sum_i \sum_j (x_i - \mu)(x_j - \mu))$$

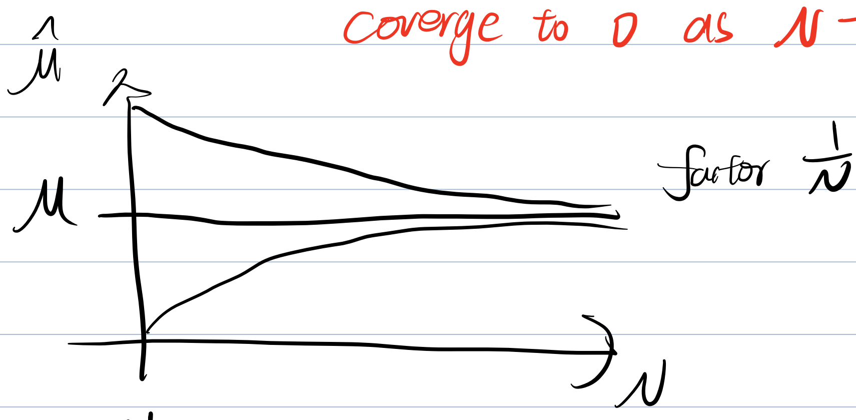
$i = j \Rightarrow E((x_i - \mu)^2) = \alpha^2$

$i \neq j \Rightarrow E[(x_i - \mu)(x_j - \mu)]$

$= E(x_i - \mu) E(x_j - \mu) = 0$

$$= \frac{1}{N^2}(N\alpha^2) = \frac{\alpha^2}{N} = Var(\hat{\mu})$$

coverge to 0 as $N \to \infty$



Gaussian Variance (PS 2-12)

$$E(\hat{\alpha}^2) = \frac{N-1}{N}\alpha^2 \Rightarrow Bias(\hat{\alpha}^2) = -\frac{1}{N}\alpha^2 \neq 0$$

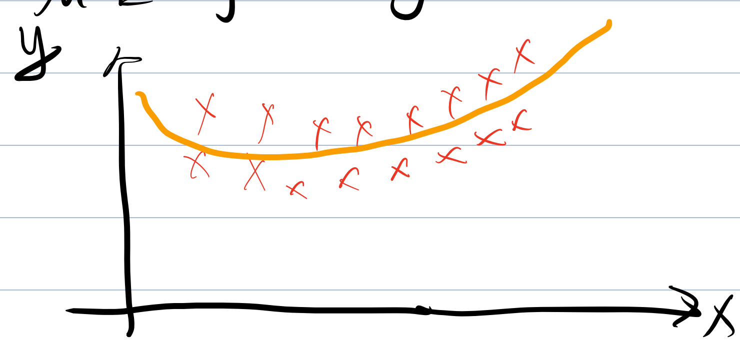
to make it unbiased:

$$\hat{\hat{\sigma}}^2 = \frac{N}{N-1}\hat{\sigma}^2 = \frac{1}{N-1}\sum_i (x_i-\mu)^2$$

Important Asymptotic Properties of MLE

1) consistent: As $N \to \infty$, the estimate converges to the true value
   Asymptotically unbiased.
2) efficient: achieves Cramer-Rao Lower Bound (CRLB) as $N \to \infty$
   - CRLB is a theoretical bound on the variance of any unbiased estimator for a given $p(x|\theta)$
   - i.e. no unbiased estimator can get lower variance.

MLE for regression



$x \in \mathbb{R}$ input
$y \in \mathbb{R}$ output
learn $f(x)$

Consider a k^{th} order polynomial

$$f(x,\theta) = \sum_{d=0}^{k} x^d \theta_d = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}^T \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_K \end{bmatrix} = \phi(x)^T \theta$$

Observe a noisy output:

$$\underbrace{y}_{\text{random variable}} = \underbrace{f(x,\theta)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{random variable}}$$

noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

equivalently,

$$p(y \mid x, \theta) = \mathcal{N}(y \mid f(x,\theta), \sigma^2)$$

Given dataset $\{(x_i, y_i)\}_{i=1}^{N}$, estimate θ using MLE

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1} \log p(y_i \mid x_i, \theta)$$
$$= \vdots$$
$$= \arg\min_{\theta} \sum_i (y_i - f(x_i,\theta))^2$$
$$= \arg\min_{\theta} \| y - \Phi^T \theta \|^2, \quad \Phi = [\phi(x_1), \dots, \phi(x_N)], \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

Notes:

1) MLE is more general than least square

2> assumptions are explicit

i) Gaussian noise

(ii) $\mu = 0$, σ^2 variance (fixed)

(iii) noise is i.i.d.

3> MLE can describe other least square formulations:

change the regression to fit the tasks
↓
change the noise distribution

i> weighted LS (ps 2.8)
(different var)

ii> regularised LS (lec. 3)

iii> Lp-norm (ps 2-9)

eg: non-negtive error ⇒ gamma

non-neg integer error ⇒ poisson