Lecture 2 — Parameter Estimation

How do we find the prob. dist. to a r.v. X?

Three steps:

1) choose a parametric model (e.g. Gaussian)
   call the parameters θ.

2) assemble a collection of samples (observations) from X.

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

↑ sample of X

We assume x_i's are independent. x_i are *iid* samples.

independently & identically distributed

3) Maximum likelihood (ML) principle.

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

ML estimate (MLE)

$$\theta^* = \arg\max_\theta p(D|\theta)$$

$p(D|\theta)$: likelihood of the D wrt. θ. "likelihood function"

$$= \arg\max_\theta \log p(D|\theta)$$ "log-likelihood" (LL)

$$= \arg\min_\theta -\log p(D|\theta)$$ "negative log-likelihood" (NLL) "loss"

Note: $\arg\max_\theta \log p(D|\theta)$

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

This is not a prob dist on θ!!!

Note: log = natural logarithm (log base e)

Data log-likelihood term

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

↓ independence

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

$\log(a \cdot b) = \log a + \log b$

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

CS5487 Lecture Notes (2020)
Dr. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

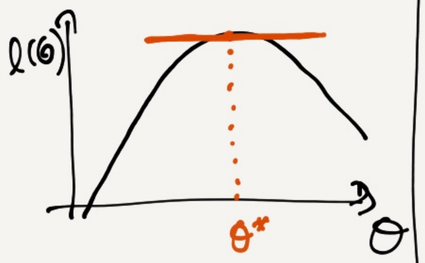
To get the ML Solution:

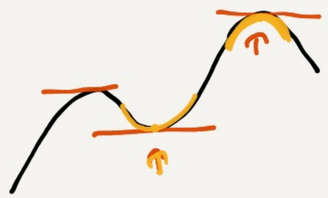
if θ is scalar, at local optimum of $\ell(\theta)$:

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

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

3) check the boundary conditions on θ (if necessary)

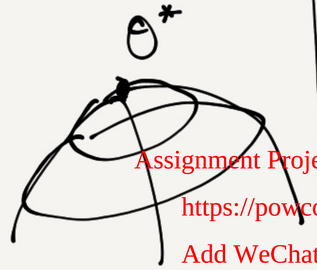




if θ is a vector: $\theta = \begin{bmatrix}\theta_1\\ \vdots \\ \theta_p\end{bmatrix}$

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

↑ gradient



2) $\nabla^2_\theta \ell(\theta) \prec 0$

↑ Hessian ↑ negative definite.

$H \prec 0 \Rightarrow \theta^T H \theta < 0 \ \forall\, \theta$
(all directions are curving downwards)

$$\nabla^2\ell(\theta) = \begin{bmatrix}\frac{\partial^2}{\partial\theta_1^2}\ell(\theta) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_p}\ell(\theta)\\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_1\partial\theta_p}\ell(\theta) & \cdots & \frac{\partial^2}{\partial\theta_p^2}\ell(\theta)\end{bmatrix}$$

"describes the curvature of $\ell(\theta)$"

Example: Bernoulli:

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

$\log a^b = b \log a$

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

$m = \sum_i x_i$ ← "sufficient statistic" - $\ell(\theta)$ only depends on N observations through this term.

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

Solve for π:

1) $\frac{\partial}{\partial\pi}\ell(\theta) = \boxed{\frac{m}{\pi} + \frac{N-m}{1-\pi}(-1)} = 0$ ↓ $\times(\pi)(1-\pi)$

$m(1-\pi) + (N-m)(-1)(\pi) = 0$

$m - N\pi = 0 \Rightarrow \boxed{\hat{\pi} = m/N = \frac{1}{N}\sum_{i=1}^N x_i}$

"fraction of 1's observed" (sample mean)

2) $\frac{\partial^2}{\partial\pi^2}\ell(\theta) = \frac{\partial}{\partial\pi}\left(\frac{\partial}{\partial\pi}\ell(\theta)\right) = \frac{\partial}{\partial\pi}\left(\frac{m}{\pi} - \frac{(N-m)}{(1-\pi)}\right)$

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

3) $0 \le m \le N \Rightarrow 0 \le \pi \le 1$ ✓

Example: Gaussian

$\theta = (\mu, \sigma^2)$, σ^2 is known.

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

$$= \sum_i \left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right]$$

$$= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

what are the suff. stats?

$(x_i - \mu)^2 = x_i^2 - 2x_i\mu + \mu^2$

$\sum_i x_i^2$, $\sum_i x_i$

Solve for μ:

$$\frac{\partial}{\partial \mu}\ell(\theta) = \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)\cdot 2\cdot(-1) = 0$$

$$\sum_i x_i - \sum_i \mu = 0$$

$$\sum_i x_i - N\mu = 0 \Rightarrow \boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

Sample mean of x.

- σ^2 is unknown, μ is known.

$$1)\ \frac{\partial}{\partial \sigma^2}\ell(\sigma) = -\frac{N}{2}\frac{1}{\sigma^2} - \frac{1}{2}\frac{1}{\sigma^4}(-1)\sum_i (x_i - \mu)^2 = 0 \quad \downarrow \times \sigma^4$$

$$= -\frac{N}{2}\sigma^2 + \frac{1}{2}\sum_i (x_i - \mu)^2 = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample variance of x.

Multivariate Gaussian

$$\begin{cases} \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \\ \hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T \end{cases}$$

Estimators

- the estimate (e.g. $\hat{\mu}$) is a number.
- the estimator is a r.v. over many datasets.
- the estimate is the value of the estimator for a given dataset D.

estimator: $f(X_1, \dots, X_N) = \frac{1}{N} \sum_{i=1}^{N} X_i$ ← r.v. for each sample $X_i \sim p(x_i \mid \theta)$ (true distribution)

ML estimate: $\hat{\mu} = f(X_1, .. X_N)\Big|_{X_1 = x_1, \dots, X_N = x_N}$

$= \frac{1}{N} \sum_i x_i$

- Since the estimator is a r.v, we can compute its mean & variance. Hence, we can quantify how good is the estimator.

Bias & Variance

- How do we measure "goodness"?
- $\hat{\theta} = f(X_1, \dots, X_N)$

1) Will it ever converge to the true value of θ?

- $\text{Bias}(\hat{\theta}) = E_{X_1 \dots X_N}[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$ (θ ← true value)
- if the bias is non-zero, we can never get the true value of θ, even with an infinite # of samples. "measures the expressiveness"

2) How long will it take to converge? How many samples do we need?

$\text{Var}(\hat{\theta}) = E_{X_1 \dots X_N}[(\hat{\theta} - E\hat{\theta})^2]$

"measures the uncertainty/variability"

Example: Gaussian (true value is μ)

Estimator: $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Bias: mean $E_{X_1 \cdots X_N}[\hat{\mu}] = E_{X_1 \cdots X_N}\left[\frac{1}{N}\sum_i X_i\right]$

$$= \frac{1}{N}\sum_i E_{X_i}[X_i] = \frac{1}{N}\sum_i \mu = \mu$$

$$\Rightarrow \boxed{\text{Bias}(\hat{\mu}) = \mu - \mu = 0} \quad = \text{"unbiased"}$$

Variance: $\text{Var}(\hat{\mu}) = E_{X_1 \cdots X_N}\left[(\hat{\mu} - \underbrace{E\hat{\mu}}_{\mu})^2\right]$

$$= E\left[\left(\frac{1}{N}\sum_i X_i - \underbrace{\mu}_{N\mu}\right)^2\right]$$

$$= \frac{1}{N^2} E\left[\left(\sum_i (X_i - \mu)\right)^2\right]$$

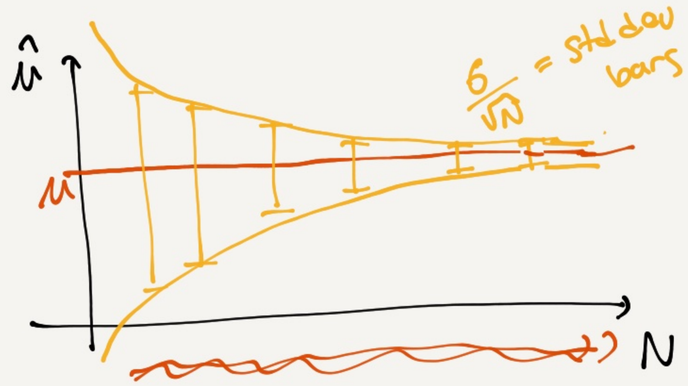
$$= \frac{1}{N^2} E\left[\sum_i \sum_j (X_i - \mu)(X_j - \mu)\right]$$

$i = j \Rightarrow E((X_i - \mu)^2) = \sigma^2$

$i \neq j \Rightarrow E((X_i - \mu)(X_j - \mu)) = \sigma_{ij}^2$

$$= \frac{1}{N^2}\left(\underbrace{\sum_{i=j} \sigma^2}_{N\sigma^2} + \underbrace{\sum_{i \neq j} \sigma_{ij}^2}_{=0}\right)$$

$$\boxed{\text{Var}(\hat{\mu}) = \frac{\sigma^2}{N}} \leftarrow \text{variance of the true Gaussian.}$$



Variance converges to 0 as $N \to \infty$

i.e. $p(\hat{\mu})$ concentrates around the true mean μ as N increases.

Another example: variance of Gaussian. (PS 2-12)

$$E[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$$

$$\boxed{\text{Bias}(\hat{\sigma}^2) = -\frac{1}{N}\sigma^2}$$

Important Asymptotic Properties of MLE

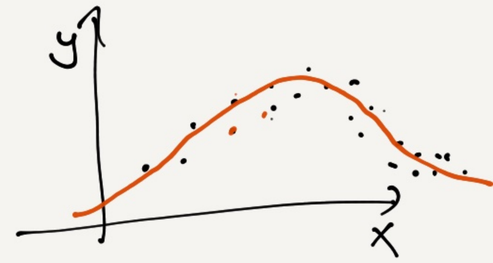
1) consistent - as $N \to \infty$, the estimated value converges to the true value. i.e. asymptotically unbiased.

2) efficient - achieves the Cramér-Rao Lower Bound (CRLB) as $N \to \infty$.

CRLB is a theoretical lower bound on the variance of an unbiased estimator, for some $p(x|\theta)$.
(no unbiased estimator can have lower variance than the CRLB, & MLE achieves it)

MLE & Regression

- supervised learning: $D = \{(x_i, y_i)\}$
- input $x \in \mathbb{R}$
- output $y \in \mathbb{R}$



E.g. estimate a polynomial function

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

$k=2$: $\theta_0 + \theta_1 x + \theta_2 x^2$; θ ; $\phi(x)$

(linear function of parameters θ)

Observe a noisy output y given an x:

$$y = f(x, \theta) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \text{ (iid)}$$

Equivalently:

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

Given a dataset $D = \{(x_i, y_i)\}$, we can use MLE to find θ:

$$\text{MLE: } \theta^* = \arg\max_\theta \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$$

$$\vdots \text{ (in tutorial)}$$

$$= \arg\min_\theta \sum_i (y_i - f(x_i, \theta))^2 \Leftarrow \text{least-squares formulation}$$

squared error

$$\Phi = \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_N) \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\theta^* = (\Phi \Phi^T)^{-1} \Phi y$$

Notes:

1) ML is more general than LS

2) all the assumptions are explicit.

   i) Gaussian Noise

   ii) iid samples (iid noise)

   iii) $\mu = 0$, σ^2 variance

3) We can change the assumptions to get other regression formulations (e.g. types of LS)

   i) weighted LS (PS 2-8)

   ii) regularized LS (lecture 3)

   iii) Lp norms (PS 2-9)

Problem 2-6

MLE for m.v. Gaussian

$D = \{x_1, \dots, x_N\}, \quad x_i \in \mathbb{R}^D$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\|x-\mu\|^2_\Sigma\right)$$

Data LL

$$\log p(D) = \sum_{i=1}^N \log p(x_i) =$$

$$= \sum_{i=1}^N -\frac{1}{2}\|x_i-\mu\|^2_\Sigma - \frac{1}{2}\log|\Sigma| - \frac{d}{2}\log 2\pi$$

(the last term: $\times$)

a) **mean μ**

$$\ell(\mu) = \sum_{i=1}^N -\frac{1}{2}\|x_i-\mu\|^2_\Sigma$$

$$= -\frac{1}{2}\sum_i (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$= -\frac{1}{2}\sum_i \underbrace{x_i^T\Sigma^{-1}x_i}_{\times} \underbrace{- x_i^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x_i}_{2x_i^T\Sigma^{-1}\mu} + \mu^T\Sigma^{-1}\mu$$

$a^Tb = b^Ta$ (scalar)

$$= -\frac{1}{2}\sum_i -2x_i^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu$$

($a^T\mu$)

$$\frac{\partial}{\partial\mu}\ell(\mu) = -\frac{1}{2}\sum_i -2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu = 0$$

$\Sigma *$ ($\sum_i -\Sigma^{-1}x_i + \Sigma^{-1}\mu$) $= 0$

$$\sum_i [-x_i + \mu] = 0$$

$$\mu = \frac{1}{N}\sum_{i=1}^N x_i$$

scalar version

$\frac{d}{dx}ax = a$ ✓

$\frac{d}{dx}ax^2 = 2ax$ ✓

⇓

vector version

$\frac{d}{dx}a^Tx = a$

$\frac{d}{dx}x^TAx = 2Ax$ (if A sym)

symmetric

$\frac{d}{dx}x^TAx = Ax + A^Tx$ (A not symmetric)

b) **covariance Σ**

$$\ell(\Sigma) = \sum_{i=1}^N -\frac{1}{2}\|x_i-\mu\|^2_\Sigma - \frac{1}{2}\log|\Sigma|$$

$$= \sum_{i=1}^N -\frac{1}{2}\underbrace{(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}_{a^Tb = \mathrm{tr}(a^Tb) = \mathrm{tr}(ba^T)} - \frac{1}{2}\log|\Sigma|$$

$$= \sum_i -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right) - \frac{1}{2}\log|\Sigma|$$

$\frac{\partial}{\partial X}\mathrm{tr}(X^{-1}A) = -(X^{-T}A^TX^{-T})$ $\quad \frac{d}{dx}\frac{a}{x} = \frac{-a}{x^2}$

$\frac{d}{dX}\log|X| = X^{-T}$ $\quad \Leftrightarrow \quad \frac{d}{dx}\log x = \frac{1}{x}$

$$\frac{\partial\ell(\Sigma)}{\partial\Sigma} = \sum_i -\frac{1}{2}\left(-\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right) - \frac{1}{2}\Sigma^{-1} = 0$$

$\times(\Sigma^* \Sigma)$

premultiply & postmultiply by Σ

$$\sum_i \left[+\frac{1}{2}(x_i-\mu)(x_i-\mu)^T - \frac{1}{2}\Sigma\right] = 0$$

$$\Rightarrow \quad \Sigma = \frac{1}{N}\sum_{i=1}^N (x_i-\mu)(x_i-\mu)^T$$

Problem 2.8

function

$$f(x) = \underbrace{\phi(x)}_{\text{features of } x}^T \overbrace{\theta}^{\text{parameters}}$$

polynomial: $\phi(x) = \begin{bmatrix} 1 \\ x \\ \vdots \\ x^p \end{bmatrix}$

$$\theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_p \end{bmatrix}$$

noisy observation

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

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

Dataset: $D = \{(x_i, y_i)\}_{i=1}^N$

MLE

$$l(D) = \sum_{i=1}^N \log p(y_i|x_i,\theta)$$

$$= \sum_i \underbrace{-\frac{1}{2\sigma^2}}_{\times} (y_i - f(x_i))^2 \underbrace{- \frac{1}{2}\log \sigma^2}_{\times} \underbrace{- \frac{1}{2}\log 2\pi}_{\times}$$

$$= \sum_i -(y_i - f(x_i))^2$$

$$\underset{\theta}{\operatorname{argmax}}\, l(D) = \underset{\theta}{\operatorname{argmin}} -l(D)$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_i (y_i - f(x_i))^2 \quad \text{← least squares objective.}$$

$$= \underset{\theta}{\operatorname{argmin}} \sum_i (y_i - \underbrace{\phi(x_i)^T\theta}_{\Phi_i})^2$$

$$= \underset{\theta}{\operatorname{argmin}} \|y - \Phi^T\theta\|^2$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\Phi = [\phi_1 \cdots \phi_N]$$

$$\underset{\theta}{\operatorname{argmin}} \|y - \Phi^T\theta\|^2 = \underset{\theta}{\operatorname{argmin}} (y - \Phi^T\theta)^T(y - \Phi^T\theta)$$

$$= \underset{\theta}{\operatorname{argmin}}\, y^Ty - \underbrace{2y^T\Phi^T\theta}_{a^Tx} + \underbrace{\theta^T\Phi\Phi^T\theta}_{x^TAx}$$

Solve for θ

$$\frac{\partial}{\partial\theta} = -2\Phi y + 2\Phi\Phi^T\theta = 0$$

$$(\Phi\Phi^T)^{-1} \times \left(\Phi\Phi^T\theta = \Phi y\right)$$

$$\boxed{\theta^* = (\Phi\Phi^T)^{-1}\Phi y}$$

(assume $\Phi\Phi^T$ is invertible)