CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Lecture 2

Parameter Estimation

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

Three Steps:

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

2) collect samples from r.v. X:
   $D = \{x_1, \dots, x_N\}$
   we assume x_i's are independent; x_i's are iid samples
   independent & identically distributed

3) Maximum likelihood principle:
   the optimal parameter θ^* is that which maximizes the probability (likelihood) of the training data.

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

likelihood of data w.r.t. param θ.
"likelihood function"

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

$\ell(\theta) = \underbrace{\text{log-likelihood}}_{LL}$ function.

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

negative LL function (loss)

Note: D is known, so $p(D|\theta)$ is a function of θ.
It is not a probability w.r.t. θ.

log = natural log (log base e)

data LL

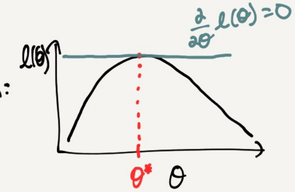
$$\ell(\theta) = \log p(D|\theta)$$

independence assumption

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

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

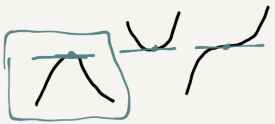
To get the MLE solution



if θ is a scalar, at local optimum:

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^*
   (local maximum; concave)

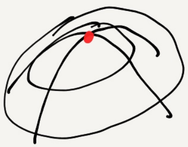


3) check the boundary conditions of θ (if necessary)

if θ is a vector:

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

   gradient



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

   Hessian

$H \prec 0$: negative definite: $\theta^T H \theta < 0,\ \forall\theta$ — "mountain" concave in all dir.

$H \succ 0$: positive defn: $\theta^T H \theta > 0,\ \forall\theta$ — "bowl" - convex in all directions.

Example: Bernoulli

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

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

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

$$= \sum_i \left[x_i \log \pi + (1-x_i) \log(1-\pi)\right]$$

$$= \underbrace{\left(\sum_i x_i\right)}_{\#\text{ of 1's}} \log \pi + \underbrace{\left[\sum_i (1-x_i)\right]}_{\#\text{ of 0's}} \log(1-\pi)$$

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

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

find the max:

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

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

$$m - \cancel{m\pi} - N\pi + \cancel{m\pi} = 0$$

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

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

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

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

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

Example: Gaussian

① $\theta = \mu$ (σ^2 known)

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

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

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

what are the sufficient statistics?
$\{\sum_i x_i, \sum_i x_i^2\}$

max wrt μ:

$$\frac{\partial \ell(\theta)}{\partial \mu} = \cancel{-}\frac{1}{\cancel{2}\sigma^2}\sum_i \cancel{2}(x_i-\mu)\cancel{(-1)} = 0$$

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

(sample mean)

② $\theta = \sigma^2$ (μ is known)

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

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

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

M.v. Gaussian

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

See next tutorial...

Estimators

the **Estimate** (e.g. $\hat{\mu}$) is a number.

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

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

r.v. for each sample

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

sample

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

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

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

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

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?)

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

Example: Gaussian

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

Mean of $\hat{\mu}$: $E_{X_1 \dots X_N}\left[\frac{1}{N} \sum_i X_i\right] = \frac{1}{N} \sum_i \overbrace{E_{X_i}[X_i]}^{\text{mean of true distribution } (\mu)} = \frac{1}{N} \cdot N\mu$

$= \mu$

$\boxed{\text{Bias of } \hat{\mu} = 0}$ ✓

Var of $\hat{\mu}$: $E_{X_1 \dots X_N}[(\hat{\mu} - E\hat{\mu})^2] = E\left[\underbrace{\left(\frac{1}{N}\sum_i X_i - \mu\right)^2}_{\left(\frac{1}{N} \sum_i (X_i - \mu)\right)^2}\right]$

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

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

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

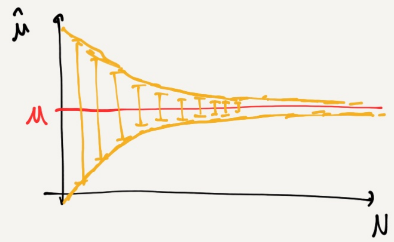
$\left(\sum_i a_i\right)^2 = \sum_{i,j} a_i a_j$

$i = j \Rightarrow E[(x_i - \mu)^2] = \sigma^2$

$i \neq j \Rightarrow E[(x_i - \mu)(x_j - \mu)] = 0$

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

variance converges to 0 as $N \to \infty$.



Gaussian variance (PS 2-12)

$$E(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 \quad \Rightarrow \quad \text{Bias}(\hat{\sigma}^2) = \underbrace{\frac{-1}{N}\sigma^2}_{} \neq 0$$

true variance

to make it unbiased:

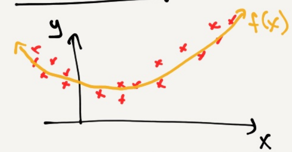
$$\hat{\hat{\sigma}}^2 = \frac{N}{N-1}\hat{\sigma}^2 = \frac{N}{N-1}\frac{1}{N}\sum_i (x_i-\mu)^2 = \boxed{\frac{1}{N-1}\sum_i (x_i-\mu)^2}$$

Important Asymptotic Properties of MLE

1) consistent - As $N \to \infty$, the estimate converges to the true value. Asymptotically unbiased.

2) efficient - achieves the Cramér-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 polynomial function (k^{th} order)

$$f(x,\theta) = \sum_{d=0}^{K} x^d \theta_d = \underbrace{\begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^K \end{bmatrix}^T}_{\phi(x)} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} = \underbrace{\phi(x)^T\theta}_{\text{linear function in } \theta.}$$

observe a noisy output:

$$y = f(x,\theta) + \epsilon$$

noise $\quad \epsilon \sim N(0,\sigma^2)$, iid.

equivalently, (y is a r.v.)

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

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

$$\hat{\theta} = \arg\max_\theta \sum_i \log p(y_i|x_i,\theta)$$

$$\vdots$$

$$\boxed{= \arg\min_\theta \sum_i (y_i - f(x_i,\theta))^2}$$

least-squares formulation

$$= \arg\min_\theta \|y - \Phi^T\theta\|^2, \quad \Phi = [\phi(x_1) \dots \phi(x_N)], \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\vdots$$

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

Notes:

1) MLE is more general than LS.

2) Assumptions are explicit

   i) Gaussian noise

   ii) $\mu = 0$, σ^2 variance (fixed)

   iii) noise is iid.

3) MLE can describe other LS formulations:

   i) weighted LS (PS 2.8)

   ii) regularized LS (lecture 3)

   iii) Lp-norm (PS 2.9)

generalized linear models (GLM)