Lecture 2

- Parameter Estimation

How do we find a probability distribution for a r.v. X

Three steps:

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

2) collect samples from r.v. X:
   $\mathcal{D} = \{x_1, \ldots, x_N\}$
   we assume x_i's are independent; x_i are iid (independant identically distribution) samples

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

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

$p(\mathcal{D}|\theta)$ ← likelihood of data w.r.t θ
aka: "likelihood function"

$$= \arg\max_\theta \underbrace{\log p(\mathcal{D}|\theta)}_{L(\theta) = \log \text{ likelihood function}}$$

$$= \arg\min_\theta \underbrace{-\log p(\mathcal{D}|\theta)}_{\text{negative log likelihood function (loss)}}$$

Note: $\mathcal{D}$ is known, so $p(\mathcal{D}|\theta)$ is a function of

θ. It is not a probability w.r.t θ.

and log = nature log = ln

- data LL (Log likelihood)

$$L(\theta) = \log P(D|\theta)$$

independent assumption

$$= \log \prod_{i=1}^{n} P(x_i|\theta)$$

$$L(\theta) = \sum_{i=1}^{n} \log P(x_i|\theta)$$

To get the MLE solution

- if θ is a scaler, at local optimum

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

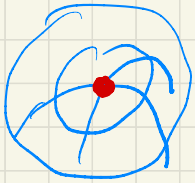
2) $\frac{\partial^2}{\partial \theta} \log P(D|\theta) < 0$ at θ^*

(local maximum; concave) : [2]

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

- If θ is a vector:

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



2) $\nabla_\theta^2 L(\theta) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \\ \vdots & \cdots & \frac{\partial^2}{\partial \theta_p^2} \end{bmatrix} L(\theta) < 0$

↑ Hessian

This Hessian need to be negative definate to make sure the local maximum

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

"mountain" concave in all directions

$H > 0$: positive definate : $\theta^T H \theta > 0 \quad \forall \theta$

"bowl" convex in all directions

凸函数：任意两点连成的线段皆位于图像的上方

Example: Bernoulli

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

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

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

$$= \log \pi \underbrace{\sum x_i}_{\#\text{ of 1's}} + \log (1-\pi) \underbrace{\sum (1-x_i)}_{\#\text{ of 0's}}$$

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

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

find the max:

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

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

$$\Rightarrow \hat{\pi} = \frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (=1)$$

"fraction of 1's observed"

(sample mean)

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

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

3) boundary condition: $0 \le m \le N$

$$\Rightarrow 0 \le \frac{m}{N} \le 1 \quad \text{(check)}$$

Example: Gaussian

$\theta = \mu$ (σ^2 known) (case 1)

$$l(\theta) = \sum \log p(x_i|\theta)$$

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

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

what are the sufficient statistics?

$\left\{\sum_i x_i,\ \sum_i x_i^2\right\}$

$$\frac{\partial L}{\partial \mu} = -\frac{1}{2\sigma^2} \sum 2(x_i - \mu)(-1) = 0$$

$$\Rightarrow \sum (x_i - \mu) = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{N} \sum x_i \quad \text{(sample mean)}$$

$\theta = \sigma^2$ (μ is known) (case 2)

($\neq \sigma^2$) $$\frac{\partial L}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - \mu)^2 = 0$$

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

"sample variance"

M.V. Gaussian:

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

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

- Estimators

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

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

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

(X_i is r.v for each sample, $X_i \sim P(x|\theta)$ true distr

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

$$\hat{\mu} = f(X_1, X_2, \cdots X_N)\Big|_{X_1 = x_1, \cdots} = \frac{1}{N}\sum_i x_i$$

(x_1: sample; x_i: sample)

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

- Bias and variance $\hat{\theta} = f(X_1, \cdots, X_N)$

  1) will it converge to the true value of θ

$$\text{Bias}(\hat{\theta}) = E_{X_1, \cdots X_N}[\hat{\theta} - \theta] = \underbrace{E_X[\hat{\theta}]}_{\text{mean of the estimator}} - \theta$$

  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?)

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

Example: Gaussian

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

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

$$= \frac{1}{N} E_{x_1 \cdots x_N}\left[\sum_i x_i\right]$$

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

Bias of $\hat{\mu} = 0$

var of $\hat{\mu} = E_{x_1 \cdots x_N}\left[(\hat{\mu} - E\hat{\mu})^2\right]$ ($\hat{\mu} \to \frac{1}{N}\sum x_i$, $E\hat{\mu} \to \mu$)

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

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

$$= \frac{1}{N^2} E\left[\sum_i \sum_j (x_i - \mu)(x_j - \mu)\right] \qquad = E\left[\frac{1}{N^2}\left(\sum (x_i - \mu)\right)^2\right]$$

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

$i \neq j \Rightarrow E(x_i - \mu)(x_j - \mu) = 0$ ← independent

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

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

Gaussian variance (ps 2-12)

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

to make it unbiased

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

Important Asymptotic properties of MLE

1) Consistent - As $N \to \infty$, the estimate converge to the true value. Asymptotic

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 $p(x|\theta)$)

No unbiased estimator can get lower variance than MLE.

- MLE for Regression.

$x \in \mathbb{R}$ input

$y \in \mathbb{R}$ output

learn $f(x)$

consider a polynomial function (Kth order)

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

$\phi(x)$ θ

observe a noisy output:

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

ε: noise $\varepsilon \sim N(0, \sigma^2)$ i.i.d

equivalently (y is a r.v.)

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

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

using MLE:

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

$$= \vdots$$

least square formulation

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

$$= \arg\min_\theta \| y - \Phi^T \theta \|^2, \quad \Phi = [\phi(x_1), \cdots \phi(x_N)]$$

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

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

Notes:

1) MLE is more general than LS

2) Assumption 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) regularised LS (Lecture 3)

iii) Lp-norm (ps 2.9)