07. Statistical Estimation

By Yang Lin[1] (2023 秋季, @NJU)

Contents

[1] Institute: Nanjing University. Email: linyang@nju.edu.cn.

1 Maximum Likelihood Estimation

Parametric distribution estimation

1. Distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed values

2. Parametric distribution estimation: choose from a family of densities $p_x(y)$, indexed by a parameter x

Maximum Likelihood Estimation

$$\texttt{maximize (over } x\texttt{)} \quad \log p_x(y)$$

1. y is observed value

2. $l(x) = \log p_x(y)$ is called log-likelihood function

3. can add constraints $x \in C$ explicitly, or define $p_x(y) = 0$ for $x \notin C$

4. a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

1 Maximum Likelihood Estimation

Linear Measurement Model

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m$$

$x \in \mathbf{R}^n$ is vector of unknown parameters

v_i is IID measurement noise, with density $p(z)$

y_i is measurement: y has density $p_x(y) = \Pi_{i=1}^m p(y_i - a_i^T x)$

maximum likelihood estimate: any solution x of

$$\texttt{maximize (over } x\texttt{)} \quad l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

y is observed value

1 Maximum Likelihood Estimation

1. **Gaussian noise** $\mathcal{N}(0, \sigma^2) : p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$,

$$l(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(a_i^T x - y_i)^2$$

ML estimate is LS solution

2. **Laplacian noise**: $p(z) = (1/(2a))e^{-|z|/a}$,

$$l(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^{m}|a_i^T x - y_i|$$

ML estimate is 1-norm solution

3. **uniform noise on** $[-a, a]$**:**

$$l(x) = \begin{cases} -m\log(2a) & |a_i^T x - y_i| \le a, \;\; i = 1, \ldots, m \\ -\infty & \texttt{otherwise} \end{cases}$$

ML estimate is any x with $|a_i^T x - y_i| \le a$

1 Maximum Likelihood Estimation

Logistic regression

Random variable $y \in \{0, 1\}$ with distribution

$$p = \mathbf{prob}(y = 1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

1. a, b are parameters; $u \in \mathbf{R}^n$ are (observable) explanatory variables

2. Estimation problem: estimate a, b from m observations (u_i, y_i)

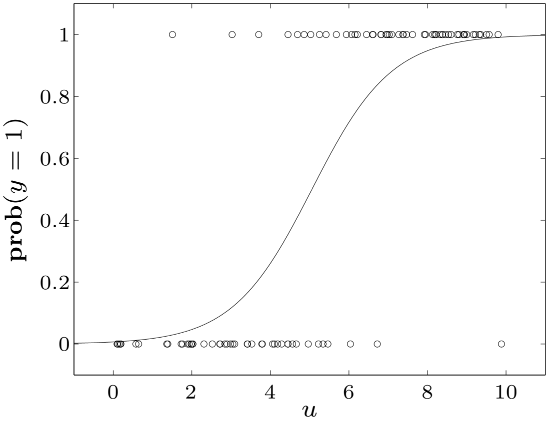
log-likelihood function (for $y_1 = \cdots = y_k = 1, y_{k+1} = \cdots = y_m = 0$):

$$\begin{aligned} l(a, b) &= \log\left(\Pi_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \Pi_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)}\right) \\ &= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b)) \end{aligned}$$

concave in a, b

1 maximum likelihood estimation

example ($n = 1$, $m = 50$ measurements)



circles show 50 points (u_i, y_i)

solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

2 Optimal Detector Design

(Binary) hypothesis testing

detection (hypothesis testing) problem

given observation of a random variable $X \in \{1, \dots, n\}$, choose between:

1. hypothesis 1: X was generated by distribution $p = (p_1, \dots, p_n)$
2. hypothesis 2: X was generated by distribution $q = (q_1, \dots, q_n)$

Randomized detector

a nonnegative matrix $T \in \mathbf{R}^{2\times n}$, with $\mathbf{1}^T T = \mathbf{1}^T$

if we observe $X = k$, we choose hypothesis 1 with probability t_{1k}, hypothesis 2 with probability t_{2k}

if all elements of T are 0 or 1, it is called a deterministic detector

2 Optimal Detector Design

detection probability matrix:

$$D = [Tp \;\; Tq] = \begin{bmatrix} 1 - P_{\mathrm{fp}} & P_{\mathrm{fn}} \\ P_{\mathrm{fp}} & 1 - P_{\mathrm{fn}} \end{bmatrix}$$

1. P_{fp} is probability of selecting hypothesis 2 if X is generated by distribution 1 (false positive)
2. P_{fn} is probability of selecting hypothesis 1 if X is generated by distribution 2 (false negative)

multicriterion formulation of detector design

$$\begin{array}{ll} \texttt{minimize (w.r.t. } \mathbf{R}^2_+\texttt{)} & (P_{\mathrm{fp}}, P_{\mathrm{fn}}) = ((Tp)_2, (Tq)_1) \\ \texttt{subject to} & t_{1k} + t_{2k} = 1, \;\; k = 1, \ldots, n \\ & t_{ik} \geq 0, \quad i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

variable $T \in \mathbf{R}^{2 \times n}$

2 Optimal Detector Design

scalarization (with weight $\lambda > 0$)

$$\begin{array}{ll} \texttt{minimize} & (Tp)_2 + \lambda (Tq)_1 \\ \texttt{subject to} & t_{1k} + t_{2k} = 1, \;\; t_{ik} \geq 0, i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

an LP with a simple analytical solution

$$(t_{1k}, t_{2k}) = \begin{cases} (1, 0) & p_k \geq \lambda q_k \\ (0, 1) & p_k < \lambda q_k \end{cases}$$

A deterministic detector, given by a likelihood ratio test

If $p_k = \lambda q_k$ for some k, any value $0 \leq t_{1k} \leq 1$, $t_{1k} = 1 - t_{2k}$ is optimal (i.e., Pareto-optimal detectors include non-deterministic detectors)

minimax detector

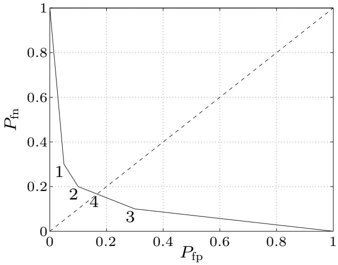
$$\begin{array}{ll} \texttt{minimize} & \max\{P_{\mathrm{fp}}, P_{\mathrm{fn}}\} = \max\{(Tp)_2, (Tq)_1\} \\ \texttt{subject to} & t_{1k} + t_{2k} = 1, \;\; t_{ik} \geq 0, i = 1, 2, \quad k = 1, \ldots, n \end{array}$$

an LP; solution is usually not deterministic

1 maximum likelihood estimation

example

$$P = \begin{bmatrix} 0.70 & 0.10 \\ 0.20 & 0.10 \\ 0.05 & 0.70 \\ 0.05 & 0.10 \end{bmatrix}$$



solutions 1, 2, 3 (and endpoints) are deterministic; 4 is minimax detector

3 Experiment Design

m linear measurements $y_i = a_i^T x + w_i, i = 1, \ldots, m$ of unknown $x \in \mathbf{R}^n$

1. measurement errors w_i are IID $\mathcal{N}(0, 1)$
2. ML(least-squares) estimate is

$$\hat{x} = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1} \sum_{i=1}^{m} y_i a_i$$

error $e = \hat{x} - x$ has zero mean and covariance

$$E = \mathbf{E} e e^T = \left(\sum_{i=1}^{m} a_i a_i^T \right)^{-1}$$

confidence ellipsoids are given by $\{x | (x - \hat{x})^T E^{-1} (x - \hat{x}) \leq \beta\}$

experiment design: choose $a_i \in \{v_1, \ldots, v_p\}$ (a set of possible test vectors) to make E 'small'

3 Experiment Design

vector optimization formulation

$$\begin{array}{ll} \texttt{minimize (w.r.t. } \mathbf{S}^n_+\texttt{)} & E = \left(\sum_{k=1}^p m_k v_k v_k^T\right)^{-1} \\ \texttt{subject to} & m_k \geq 0, \;\; m_1 + \cdots + m_p = m \\ & m_k \in \mathbf{Z} \end{array}$$

1. variables are m_k (# vectors a_i equal to v_k)
2. difficult in general, due to integer constraint

relaxed experiment design

assume $m \gg p$, use $\lambda_k = m_k/m$ as (continuous) real variable

$$\begin{array}{ll} \texttt{minimize (w.r.t. } \mathbf{S}^n_+\texttt{)} & E = (1/m)\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \texttt{subject to} & \lambda \succeq 0, \;\; \mathbf{1}^T\lambda = 1 \end{array}$$

common scalarizations: minimize $\log\det E$, $\mathbf{tr} E$, $\lambda_{\max}(E)$,...

can add other convex constraints, e.g., bound experiment cost $c^T\lambda \leq B$

3 Experiment Design

D-optimal design

$$\begin{array}{ll} \texttt{minimize} & \log\det\left(\sum_{k=1}^{p} \lambda_k v_k v_k^T\right)^{-1} \\ \texttt{subject to} & \lambda \succeq 0, \ \mathbf{1}^T\lambda = 1 \end{array}$$

interpretation: minimizes volume of confidence ellipsoids

dual problem

$$\begin{array}{ll} \texttt{maximize} & \log\det W + n\log n \\ \texttt{subject to} & v_k^T W v_k \le 1, \quad k = 1, \ldots, p \end{array}$$

interpretation: $\{x | x^T W x \le 1\}$ is minimum volume ellipsoid centered at origin, that includes all test vectors v_k

complementary slackness: for λ, W primal and dual optimal

$$\lambda_k (1 - v_k^T W v_k) = 0, \quad k = 1, \ldots, p$$

optimal experiment uses vectors v_k on boundary of ellipsoid defined by W