Parameter Estimation

Sunita Sarawagi

CS 215. Fall 2024

So far..

- Computing probabilities of outcomes given a fixed distribution.
- Distributions were given to us as a function..
- Functions had parameters with fixed values

What are Parameters?

Consider some probability distributions:

- Ber(p) — $\theta = p$
- Poi(λ) — $\theta = \lambda$
- Uni(α, β) — $\theta = (\alpha, \beta)$
- Normal(μ, σ^2) — $\theta = (\mu, \sigma^2)$
- Y = **m**X + **b** — $\theta = (m, b)$
- etc…

$X \sim N(0,1)$

Call these "parametric models"

Non parametric model

example – histogram

Given model, **parameters** yield actual distribution

- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

Today's class

How to determine the values of the parameters.

Parameters differ based on the task and application. These are not fixed like the speed of light.

The setup for parameter estimation in real-life

- Step 1: A real-life problem:
  1. Estimating the probability that at least two out of four servers will be alive next day
  2. The probability that stock price will rise by 10% in the next week
  3. The expected number of clicks on an advertisement in the next 3 hours
- Step 2: Model the problem: Choose a functional form of the uncertainty.
  1. Binomial?

     Assume that servers fail independently $X \sim \mathrm{Bin}($

     X = # of failures in a day
  2. Gaussian?

     X = change from one day to the next
  3. Poisson?

     X = # of clicks on the ad per hour

The setup for parameter estimation in real-life

Step 3: Collect a training sample by observing over several days.

1. Sample server failure data observed over 3 days

day1 ser1   day1 ser2 x_3 x_4 x_5 x_6 x_7 x_8 day 3 ser 3

x_1 x_2 x_{12}

0   0   1   0   0   0   0   1   ----   0

2. Stock price change over a 10 days

change $1 \to 2$ $2 \to 3$ - - - - -   9–10

1%   −2%   - -   3%

x_1 x_2 x_9

3. Number of clicks on the ad over the last 20 hour

Hour   1   2   3   -   20

10   15   5   7 x_{20}

x_1 x_2 x_2

- Step 4: Estimate the unknown parameters using the training sample

The overall setup in parameter estimation

density of pmf

- Given: a density or distribution function with parameters $f(x, \theta)$
- Given: sample: $D = \{x_1, x_2, \dots, x_N\}$
  - The i-th sample is a random variable X_i assumed to be independently identically distributed as per the unknown $f(x, \theta)$
- Find θ.

- Since D is a finite sample, we cannot really know the actual θ. Best we can do is obtain an estimate of θ.
- We will denote the estimate as $\boxed{\hat{\theta}}$
- Goodness of estimate will be discussed later.

Types of estimators

- Maximum likelihood: sample D is all you got. $\hat{\theta}$ point estimator
- Bayesian estimation: in addition to sample, we got prior beliefs.

Maximum Likelihood Estimation

- If θ were known we could have calculated the probability of getting the N outcomes in $D = \{x, x_2, \ldots, x_N\}$ from the distribution as

$x_1 \quad x_N$ are independent

- $P(D|\theta) = P(x_1, \ldots, x_N|\theta) = \prod_i P(x_i|\theta) = \boxed{\prod_i f(x_i; \theta)}$ ← for both continuous & discrete

- Likelihood refers to the above function. Often denoted as $L(\theta)$

- Maximum likelihood estimator:
  - Choose the parameter θ for which the above likelihood is maximized

$$\hat{\theta} \equiv \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} f(x_i, \theta) \rightarrow$$

Finding θ that maximizes likelihood

- Use log-likelihood instead of likelihood to convert products into sums
- $LL(\theta) =$ $\underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \log f(x_i, \theta)$ $\max_{\theta} LL(\theta)$

sum over observations

- Maximum likelihood estimator

$$\boxed{\hat{\theta} = \text{argmax}_{\theta} \sum_{i=1}^{N} \log f(x_i|\theta)}$$

Solved using numerical optimization methods applying calculus.

MLE for Bernoulli

$$X \sim Bern(p) \qquad X \in \{0, 1\}$$

$$f(x;p) = p^x(1-p)^{1-x}$$

Data sample: D $\quad N = 10$

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0	1	1	0	0	0	0	0	1	1

$$\max_p LL_D(\theta = [p]) = LL_D(p) = \max_p \sum_{i=1}^{N} \log p^{x_i}(1-p)^{1-x_i}$$

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

Let $\sum_{i=1}^{N} x_i = N_1$

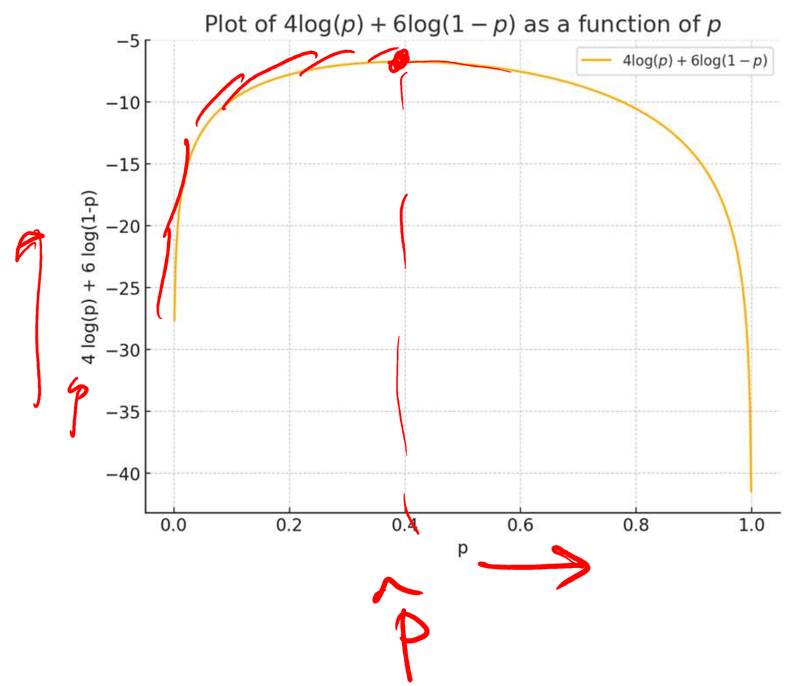
$$\max_{p} \underbrace{N_1 \log p + (N-N_1)\log(1-p)}_{LL(p)}$$

$\leftarrow$ concave in p

$\Rightarrow$ unique maxima at the p where $\frac{\partial LL}{\partial p} = 0$

$$\frac{\partial LL}{\partial p} = \frac{N_1}{\hat{p}} - \frac{N-N_1}{1-\hat{p}} = 0$$

$$\Rightarrow \hat{p} = \frac{N_1}{N}$$



Examples: MLE for Poisson

$$f(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$x \sim \exp(\lambda)$$

$$f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$D = \{x_1, x_2 - - x_N\}$$

$$LL(\lambda) = \sum_{i=1}^{N} \log\left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) = \left(\sum_{i=1}^{N} x_i\right)\log\lambda - \lambda N - \sum_{i=1}^{N}\log x_i!$$

$$\hat{\lambda} = \underset{\lambda}{\arg\max} \underbrace{\left(\sum_{i=1}^{N} x_i\right)\log\lambda - \lambda N}$$

$$\frac{\partial LL}{\partial\lambda} = \frac{\sum x_i}{\lambda} - N \; : \; \hat{\lambda} = \frac{\sum x_i}{N} \rightarrow \text{sample mean}$$

MLE for Gaussian

Homework

$$X \sim N(\underbrace{\mu, \sigma^2})$$

$$f(x, \theta = [\mu, \sigma^2]) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\frac{\partial LL}{\partial \mu} = 0$$

at μ calculated above

$$\frac{\partial LL}{\partial \sigma} = 0$$

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-\sum_1^n (x_i - \mu)^2}{2\sigma^2}\right]$$

The logarithm of the likelihood is thus given by

$$\log\, f(x_1, \ldots, x_n|\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_1^n (x_i - \mu)^2}{2\sigma^2}$$

In order to find the value of μ and σ maximizing the foregoing, we compute

$$\frac{\partial}{\partial\mu}\log\, f(x_1, \ldots, x_n|\mu, \sigma) = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial\sigma}\log\, f(x_1, \ldots, x_n|\mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_1^n (x_i - \mu)^2}{\sigma^3}$$

Equating these equations to zero yields that

$$\hat{\mu} = \sum_{i=1}^n x_i/n$$

and

$$\hat{\sigma} = \left[\sum_{i=1}^n (x_i - \hat{\mu})^2/n\right]^{1/2}$$

Example 7.2.d. The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:

4, 0, 6, 5, 2, 1, 2, 0, 4, 3

Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

Homework

- Most difficult question: what distribution to use to model accidents in a city?
  - Binomial? Will need to know total number of drivers
  - Gaussian?
  - Poisson?

Solution in textbook

MLE for a new distribution: Gamma distribution

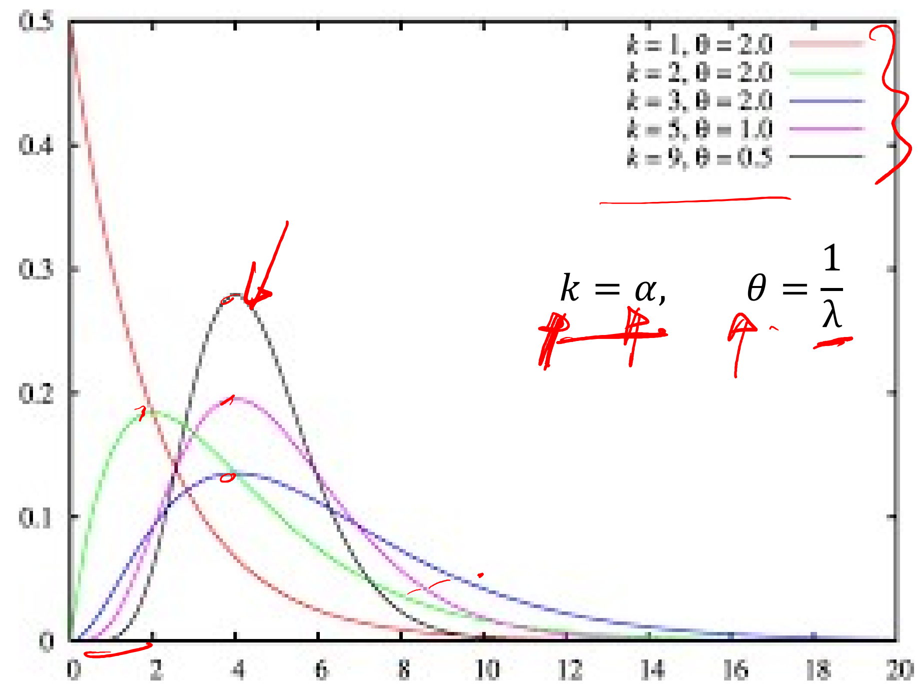
A random variable is said to have a gamma distribution with parameters (α, λ), $\lambda > 0$, $\alpha > 0$, if its density function is given by

$$f(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for what α is $f(x) \sim \exp(\lambda)$
$= \lambda e^{-\lambda x}$

$\alpha = 1$

- Can look like Gaussian for positive random variables.
- Reduces to exponential when $\alpha = 1$
- More flexible than exponential since mode is not at 0.
- Useful to model one-sided long tails e.g. blue curve here.



What is $\Gamma(\alpha)$? Gamma function

$$\Gamma(\alpha) = \int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{\alpha-1}\,dx$$

$$= \int_0^\infty e^{-y} y^{\alpha-1}\,dy \quad \text{(by letting } y = \lambda x\text{)}$$

The integration by parts formula $\int u\,dv = uv - \int v\,du$ yields, with $u = y^{\alpha-1}$, $dv = e^{-y}dy$, $v = -e^{-y}$, that for $\alpha > 1$,

$$\int_0^\infty e^{-y} y^{\alpha-1}\,dy = -e^{-y}y^{\alpha-1}\Big|_{y=0}^{y=\infty} + \int_0^\infty e^{-y}(\alpha-1)y^{\alpha-2}\,dy$$

$$\boxed{= (\alpha-1)\int_0^\infty e^{-y}y^{\alpha-2}\,dy}$$

or

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \tag{5.7.1}$$

$\Gamma(1) = 1$

If α is + integer then

$\Gamma(\alpha) = (\alpha-1)!$

Estimate MLE of parameter λ of Gamma distribution.

$$D = \{ x_1, x_2 \ldots x_N \}$$

$$\hat{\theta} = \hat{\lambda} = \underset{\lambda}{\arg\max} \sum_{i=1}^{N} \log \frac{e^{-\lambda x_i} (\lambda x_i)^{\alpha - 1} \lambda}{\Gamma(\alpha)}$$

$$= \underset{\lambda}{\arg\max} \underbrace{\sum_{i=1}^{N} \left[-\lambda x_i + (\alpha - 1) \log \lambda + \log \lambda \right]}_{F(\lambda)}$$

$$\frac{\partial F}{\partial \lambda} = -\sum_{i=1}^{N} x_i + \frac{(\alpha - 1) N}{\lambda} + \frac{N}{\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{\alpha N}{\sum_{i=1}^{N} x_i}$$

MLE for α

$$\underset{\alpha}{\operatorname{argmax}} \ \underbrace{(\alpha-1)\left[N\log\lambda + \sum_{i=1}^{N}\log x_i\right] - N\log\Gamma(\alpha)}_{G(\alpha)}$$

$$\frac{\partial G(\alpha)}{\partial \alpha} = \sum_{i=1}^{N}\log x_i - \underline{\frac{N}{\Gamma(\hat{\alpha})}\frac{\partial}{\partial \alpha}\Gamma(\alpha)\Big|_{\hat{\alpha}}} + N\log\lambda = 0$$

Not easy to solve in closed form.
But can be estimated numerically.