Parameter Estimation: Maximum-Likelihood and Bayesian

Adway Mitra
MLFA AI42001 Center for Artificial Intelligence
Indian Institute of Technology Kharagpur

October 10, 2019

Common Discrete Distributions

Distribution	Support	PMF	Parameters
Bernoulli	$\{0,1\}$	$p^x(1-p)^{(1-x)}$	p
Binomial	$\mathcal{Z}$	$\binom{N}{x}p^x(1-p)^{N-x}$	N, p
Poisson	$\mathcal{Z}$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Geometric	$\mathcal{Z}^+$	$(1-p)^{x-1}p$	p
Categorical	$\{V_1,\ldots,V_K\}$	$\prod_{k=1}^K p_k^{I(x=k)}$	$(p_1,p_2,\ldots,p_k)$
Multinomial	$\mathcal{Z}^K$	$\frac{N!}{n_1!..n_K!}\prod_{k=1}^K p_k^{n_k}$	$(N_1,p_1,\ldots,N_K,p_K)$

Common Continuous Distributions

Distribution	Support	PDF	Parameters
Beta	$(0,1)$	$\frac{1}{B(a,b)} x^{(a-1)}(1-x)^{(b-1)}$	(a,b)
Gamma	$\mathcal{R}+$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} exp(-\beta x)$	(α,β)
Gaussian	$\mathcal{R}$	$\frac{1}{\sqrt{(2\pi)}\sigma} exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$	(μ,σ)
M.V. Gaussian	$\mathcal{R}^D$	$\frac{1}{2\pi^{\frac{D}{2}}\|\Sigma\|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$	(μ,Σ)

Parameter Estimation Problem

- Given: N observations $x_1, x_2, \ldots, x_N$
- Imagine these observations are observations of IID random variables
- Choose a suitable distribution for them
- Support, histogram important considerations to choose distribution
- Need parameters for the distribution!

Maximum Likelihood Estimation (MLE)

- Write down joint PMF/PDF of the observations
- $prob(X_1 = x1, X_2 = x_2, \ldots, X_N = x_N) = \prod_{n=1}^{N} prob(X_n = x_n)$
- This is also called **likelihood function** $\mathcal{L}(p)$ of parameters p of the distribution (prob)
- Choose parameters such that this likelihood is maximized!
- Differentiate w.r.t p, equate to 0, solve equations!

MLE of Bernoulli

- Input: results of N tosses, $X_i \in \{0, 1\}$
- Based on support, we choose Bernoulli Distribution
- Need to estimate parameter p

$$\mathcal{L}(p) = \prod_{n=1}^{N} prob(X_n = x_n) = \prod_{n=1}^{N} p^{x_n}(1-p)^{1-x_n} = p^{N_1}(1-p)^{N_0} \quad (1)$$

$$p_{MLE} = argmax_p \mathcal{L}(p) = \frac{N_1}{N_1 + N_0} \quad (2)$$

MLE for Poisson

- Input: N integer observations, $X_i \in \mathcal{Z}$
- Based on support and histogram, we may choose Poisson Distribution
- Need to estimate parameter λ

$$\mathcal{L}(\lambda) = \prod_{n=1}^{N} prob(X_n = x_n) \propto \prod_{n=1}^{N} e^{-\lambda}\lambda^{x_n}$$
$$= e^{-N\lambda}\lambda^{\sum_{n=1}^{N} x_n} \tag{3}$$

$$\lambda_{MLE} = argmax_{\lambda}\mathcal{L}(\lambda) = \frac{\sum_{n=1}^{N} x_n}{N} \tag{4}$$

MLE for Gaussian

- Input: N real observations, $X_i \in \mathcal{R}$
- Based on support and histogram, we may choose Gaussian/Normal Distribution
- Need to estimate parameter (μ, σ)

$$\begin{aligned} \mathcal{L}(\mu, \sigma) = \prod_{n=1}^{N} prob(X_n = x_n) &\propto \prod_{n=1}^{N} \frac{1}{\sigma} exp(-\frac{(x_n - \mu)^2}{\sigma^2}) \\ &= \frac{1}{\sigma^N} exp(-\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{\sigma^2}) \end{aligned} \quad (5)$$

$$\mu_{MLE}, \sigma_{MLE} = argmax_{\mu,\sigma} \mathcal{L}(\mu, \sigma)$$

$$\mu_{MLE} = \frac{\sum_{n=1}^{N} x_n}{N}, \sigma_{MLE} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{MLE})^2 \quad (6)$$

MLE for Multivariate Gaussian

- Input: N real vector observations, $X_i \in \mathcal{R}^D$
- Based on support and histogram, we may choose Gaussian/Normal Distribution
- Need to estimate parameter (μ, Σ)

$$\mathcal{L}(\mu,\Sigma) = \prod_{n=1}^{N} prob(X_n = x_n) \propto \prod_{n=1}^{N} \frac{1}{|\Sigma|^{\frac{1}{2}}} exp(-(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

$$= \frac{1}{|\Sigma|^{\frac{N}{2}}} exp(-\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu) \quad (7)$$

$$\mu_{MLE}, \Sigma_{MLE} = argmax_{\mu,\Sigma}\mathcal{L}(\mu,\Sigma)$$

$$\mu_{MLE} = \frac{\sum_{n=1}^{N} x_n}{N}, \Sigma_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{MLE})(x_n-\mu_{MLE})^T \quad (8)$$

Bayesian Parameter Estimation

- Maximum likelihood estimate - entirely based on data
- But if data is not reliable?
- Bayesian approach: we may have some prior beliefs
- Bayesian approach: combine our prior beliefs with evidence, i.e. data
- Bayesian approach: keep updating our beliefs as more and more data comes in!

Bayesian Parameter Estimation

- Consider the parameters as random variables
- Put a prior distribution on the parameters
- $posterior(param|data) \propto prob(data|param) * prior(param)$
- $prob(data|param) = \mathcal{L}(param)$ (likelihood function)
- Difference from MLE - we get a distribution of the parameter instead of single value
- Maximum A-Posteriori (MAP) estimate: $param_{Bayes} = argmax_{param} posterior(param|data)$

Bayesian Parameter Estimation

- How to choose prior distribution?
  - Reflect our belief on parameter
  - Mathematical tractability (posterior should be a valid distribution)
- Some likelihood functions have conjugate prior
- Prior and Posterior on parameters should be same distribution with different parameters!
  - Easy to interpret

Bayesian estimate of Bernoulli Distribution

- Data: $\{x_1, \ldots, x_N\} \in \{0,1\}$, Model: $X_i \sim Bernoulli(p)$
- $p \in (0,1)$, so $prior(p) = Beta(a,b)$
  - (a,b) hyperparameters - parameters of prior
  - Assume $a = b$ if no information

$$\begin{aligned} posterior(p|X) &\propto \prod_{i=1}^{N} prob(X_i = x_i|p) * prior(p) \\ &= p^{N_1}(1-p)^{N_0} * p^{a-1}(1-p)^{b-1} \\ &= p^{N_1+a-1}(1-p)^{N_0+b-1} \end{aligned} \tag{9}$$

- $posterior(p) : Beta(N_1 + a, N_0 + b)$
- $p_{Bayes} = \frac{N_1+a}{N+a+b}$

Bayesian estimate of Bernoulli parameters

- $prior(p) = Beta(5,7)$, $p_{MAP} = 5/12$
- $X_1 = TAIL$, $posterior(p) = Beta(5,8)$, $p_{MAP} = 5/13$
- $X_2 = HEAD$, $posterior(p) = Beta(6,8)$, $p_{MAP} = 6/14$
- $X_3 = HEAD$, $posterior(p) = Beta(7,8)$, $p_{MAP} = 7/15$
- $X_4 = TAIL$, $posterior(p) = Beta(7,9)$, $p_{MAP} = 7/16$
- $X_5 = HEAD$, $posterior(p) = Beta(8,9)$, $p_{MAP} = 8/17$
- $X_6 = HEAD$, $posterior(p) = Beta(9,9)$, $p_{MAP} = 9/18$

Bayesian estimate of Gaussian Distribution - variance known

- Data: $\{x_1, \ldots, x_N\} \in \mathcal{R}$, Model: $X_i \sim \mathcal{N}(\mu, \sigma)$
- $\mu \in \mathcal{R}$, so $prior(\mu) = \mathcal{N}(\mu_0, \sigma_0)$
- Assume σ is known for simplicity

$$
\begin{aligned}
posterior(\mu|X) &\propto \prod_{i=1}^{N} prob(X_i = x_i|\mu) * prior(\mu) \\
&= \frac{1}{\sigma^N} exp(-\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2}) * \frac{1}{\sigma_0} exp(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}) \\
&= \mathcal{N}(\frac{\frac{N}{\sigma^2}\hat{X} + \frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}})
\end{aligned}
\tag{10}
$$

where $\hat{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$

Bayesian estimate of Gaussian Distribution - Variance Unknown

- Data: $\{x_1, \ldots, x_N\} \in \mathcal{R}$, Model: $X_i \sim \mathcal{N}(\mu, \tau)$ where $\tau = \frac{1}{\sigma^2}$
- Define $prior(\mu, \tau) = prior_1(\mu|\tau), prior_2(\tau)$
- $prior_1(\mu|\tau) = \mathcal{N}(\mu_0, \eta\tau)$, $prior_2(\tau) = Gamma(a, b)$
- $posterior(\mu, \tau) = \mathcal{L}(\mu, \tau, X) * prior_1(\mu|\tau) * prior_2(\tau)$
- $posterior(\mu) = \int posterior(\mu, \tau) d\tau$: $GaussianDistribution$
- $posterior(\tau) = \int posterior(\mu, \tau) d\mu$: $GammaDistribution$