CSC4020 Fundamental of Machine Learning
Frequentist and Bayesian Inference

Hongyuan Zha

Institute for Data and Decision Analytics
Chinese University of Hong Kong, Shenzhen

Parameter of Interest

- We focus on parametric models

$$\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$$

  with $\Theta \subset R^k$, and $\theta = [\theta_1, \ldots, \theta_k]$ on the population P
- Some times we are just interested some of the parameters: $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta = [\mu, \sigma^2]$. We just want to estimate $\mu = T(\theta)$, **parameter of interest** and σ^2 a **nuisance parameter**
- More generally, the parameter of interest can be complicated function of θ: suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we want to estimate

$$\tau \equiv P(X > 1) = 1 - P(X < 1) = 1 - P\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{1-\mu}{\sigma}\right)$$

Maximum Likelihood

Many of the main concepts and methods of MLE were developed by R. A. Fisher, 1921 and 1925

We first consider $\theta \in R$

- IID $X_1, \dots, X_n \sim f(x;\theta)$. The likelihood function

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

  The log-likelihood function is defined by $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$

- The likelihood function is just the joint density of the data, except that we treat it is a function of the parameter θ: $\mathcal{L}_n : \Theta \mapsto [0, \infty)$
- **MLE**. The θ that maximizes $\mathcal{L}_n(\theta)$, denoted by $\widehat{\theta_n}$

MLE: An Example

- Let $X_1, \ldots, X_n$ be IID with uniform distribution $\mathrm{Uniform}(0, \theta)$

$$f(x; \theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

- For a fixed θ, if for some i, $X_i > \theta$, then $f(X_i; \theta) = 0 \Rightarrow \mathcal{L}_n(\theta) = 0$
- $\mathcal{L}_n(\theta) = 0$ iff $X_{\max} > \theta$
- If $X_{\max} \le \theta$, then $f(X_i; \theta) = 1/\theta$

$$\mathcal{L}_n(\theta) = \begin{cases} 1/\theta^n & X_{\max} \le \theta \\ 0 & \text{otherwise} \end{cases}$$

- $\mathcal{L}_n(\theta)$ is strictly decreasing over $[X_{\max}, \infty) \Rightarrow \widehat{\theta_n} = X_{\max}$

Properties of MLE

- **Consistency**. The MLE is consistent: $\widehat{\theta_n} \xrightarrow{P} \theta^*$
- **Equivariance**. Let $\tau = g(\theta)$ and g is one-to-one. Let $\widehat{\theta_n}$ be the MLE of θ, then $\widehat{\tau_n} = g(\widehat{\theta_n})$ is the MLE of τ
- **Asymptotically normality**. The distribution of $\widehat{\theta_n}$ is approximately normal
- **Asymptotically optimality**. The MLE has the smallest (asymptotic) variance, i.e., the MLE is **efficient** or **asymptotically optimal**

The Bayesian Method for Parameter Estimation

- Three steps of Bayesian inference
  1. Start with $f(\theta)$, the **prior distribution**, expressing our beliefs about a parameter θ before we see any data
  2. Choose a statistical model $f(x|\theta)$, reflecting our beliefs about x given θ
  3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the **posterior distribution** $f(\theta|X_1, \ldots, X_n)$
- For Step 3, consider a simple case: Θ and X are discrete RVs

$$P(\Theta = \theta | X = x) = \frac{P(\Theta = \theta, X = x)}{P(X = x)}$$

$$= \frac{P(X = x|\Theta = \theta))P(\Theta = \theta)}{\sum_\theta P(X = x|\Theta = \theta))P(\Theta = \theta)}$$

A simple case of **Bayes' theorem**

Posterior Density

- When we have continuous RVs, we use density functions

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

- If we have n IID observations $X_1, \ldots, X_n$, $f(x|\theta)$ becomes

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \mathcal{L}_n(\theta)$$

  the likelihood function

- **Notation**. $X^n = [X_1, \ldots, X_n], x^n = [x_1, \ldots, x_n]$

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta)$$

Summarizing Posterior

- **Normalizing constant** (not a function of θ)

$$c_n = \int f(x^n|\theta) f(\theta) d\theta$$

- **Posterior is proportional to Likelihood times Prior**

$$f(\theta|x^n) \propto \mathcal{L}_n(\theta) f(\theta)$$

- With the posterior, we can summarize it using the posterior mean

$$\bar{\theta}_n = \int \theta f(\theta|x^n) d\theta = \frac{\int \theta \mathcal{L}_n(\theta) f(\theta) d\theta}{\int \mathcal{L}_n(\theta) f(\theta) d\theta}$$

- **MAP**

$$\widehat{\theta_n}^{\mathrm{MAP}} = \arg\max_{\theta \in \Theta} f(\theta|x^n) = \arg\max_{\theta \in \Theta} [\mathcal{L}_n(\theta) + \log f(\theta)]$$

The Bayesian Interval Estimate

- Bayesian interval estimate: pick a and b such that

$$\int_{-\infty}^{a} f(\theta|x^n)d\theta = \alpha/2, \quad \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$$

- Let $C = [a, b]$, then (when x^n is given a and b are fixed)

$$P(\theta \in C|x^n) = \int_{a}^{b} f(\theta|x^n)d\theta = 1 - \alpha$$

  C is $1 - \alpha$ posterior interval

Examples

- EXAMPLE. IID $X_1, \ldots, X_n$ Bernoulli(p)
- Take prior $f(p)$ uniform over $[0, 1]$ as the prior. Given the realization $x^n = [x_1, \ldots, x_n]$, the posterior

$$f(p|x^n) \propto \mathcal{L}_n(\theta)f(\theta) = p^s(1-p)^{n-s}, \quad s = \sum_i x_i$$

  Here, if $X_i = 1$ for head, and $X_i = 0$ for tail, then s is the number of heads in n trials

- Recall Beta distribution

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

IID Bernoulli Case

- The posterior is a Beta distribution

$$\theta | x^n \sim \mathrm{Beta}(s+1, n-s+1)$$

- The mean of a Beta distribution is $\alpha/(\alpha+\beta)$, so the posterior mean is

$$\bar{p} = \frac{s+1}{n+2} = \lambda_n \widehat{p} + (1-\lambda_n)\tilde{p}, \quad \tilde{p} = \frac{1}{2}, \lambda_n = \frac{n}{n+2} \approx 1$$

  here $\widehat{p} = s/n$ is the MLE

- The posterior mean is a convex combination of the MLE and the prior mean, which is $1/2$

IID Bernoulli Case: Conjugate Prior

- We can also use a Beta distribution prior: $p \sim \mathrm{Beta}(\alpha, \beta)$ with **known and fixed** α and β. The posterior becomes

$$\theta | x^n \sim \mathrm{Beta}(\alpha + s, \beta + n - s)$$

  which is also also a Beta distribution
- The flat prior $f(p) = 1 \Leftrightarrow \alpha = \beta = 1$. The posterior mean is

$$\bar{p} = \frac{\alpha + s}{\alpha + \beta + n} = \lambda_n \widehat{p} + (1 - \lambda_n)\tilde{p}, \quad \lambda_n = \frac{n}{\alpha + \beta + n}, \quad \tilde{p} = \frac{\alpha}{\alpha + \beta}$$

- The prior was a Beta distribution and the posterior was a Beta distribution
- **Conjugate prior**: When the prior and the posterior are in the same family, the prior is conjugate with respect to the model

IID Normal Case

- IID $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, assuming σ^2 is known
- The prior for $\theta \sim \mathcal{N}(a, b^2)$, a and b^2 known and fixed. Then the posterior for θ is (using the linear Gaussian systems theorem!)

$$\theta | x^n \sim \mathcal{N}(\bar{\theta}, \tau^2), \quad \bar{\theta} = \underbrace{w\bar{X} + (1-w)a}_{\text{posterior mean}},$$

$$w = \frac{\frac{1}{\text{se}^2}}{\frac{1}{\text{se}^2} + \frac{1}{b^2}}, \quad \frac{1}{\tau^2} = \frac{1}{\text{se}^2} + \frac{1}{b^2}, \quad \text{se}^2 = \frac{\sigma^2}{n}$$

  se is the standard error of the MLE $\widehat{\theta} = \bar{X}$
- When $n \to \infty$, $w \to 1$: for large n the posterior is approximately $\mathcal{N}(\widehat{\theta}, \text{se}^2)$
- Fix n, and let $b \to \infty$, a flat prior, we obtain the same result
- This is another example of a **conjugate prior**

IID Normal Case: Bayesian Interval

- We seek to find $C = [c, d]$ such that $P(\theta \in C|X^n) = .95$
  — We choose c such that $P(\theta < c|X^n) = .025$ and $P(\theta > d|X^n) = .025$

$$P(\theta < c|X^n) = P\left(\frac{\theta - \bar{\theta}}{\tau} < \frac{c - \bar{\theta}}{\tau}|X^n\right) = P\left(Z < \frac{c - \bar{\theta}}{\tau}\right) = 0.025$$

  assuming $(\theta - \bar{\theta})/\tau$ is approximately Gaussian
- We set

$$\frac{c - \bar{\theta}}{\tau} = 1.96 \quad \Rightarrow \quad c = \bar{\theta} - 1.96\tau$$

- Similarly, $d = \bar{\theta} + 1.96\tau$
- A 95% Bayesian interval for θ is $\bar{\theta} \pm 1.96\tau$

Simulation

- For general Bayesian inference, the posterior often needs to be approximated by simulation
- Draw IID sample $\theta_1, \ldots, \theta_B \sim f(\theta|x^n)$
- A histogram of $\theta_1, \ldots, \theta_B$ approximate $f(\theta|x^n)$, the posterior density
- Posterior mean $\mathcal{E}(\theta|x^n) \approx \frac{1}{B}\sum_i \theta_i$
- The $1-\alpha$ Bayesian interval is approximated by $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$, $\theta_{\alpha/2}$ is $\alpha/2$ sample quantile of $\theta_1, \ldots, \theta_B$
- For $\tau = g(\theta)$, $\tau_i = g(\theta_i)$. Then $\tau_1, \ldots, \tau_B \sim f(\tau|x^n)$. This avoids the need to do any analytical calculations
- However to effectively draw samples from $f(\theta|x^n)$ is the active research area of Monte Carlo methods

Flat Priors, Improper Priors, Noninformative Priors

- Flat prior $f(\theta) = \text{constant}$, encoding our ignorance about the prior knowledge of θ
- Improper Priors. $X \sim \mathcal{N}(\theta, \sigma^2)$, assuming σ^2 is known. If we impose a flat prior $f(\theta) = \text{constant}$, but $\int f(\theta) d\theta \neq 1$, not a probability density in the usual sense $\Rightarrow$ **improper prior**
- Another example (interpolating noise-free data)

$$p(x) = \mathcal{N}(x|0, \sigma^2(L^T L)^{-1}) \propto \exp\left(-\frac{1}{2\sigma^2}\|Lx\|_2^2\right)$$

- However, we can still form the posterior

$$f(\theta|x^n) \propto \mathcal{L}_n(\theta) f(\theta) = \mathcal{L}_n(\theta) \quad \Rightarrow \quad \theta|X^n \sim \mathcal{N}(\bar{X}, \sigma^2/n)$$

- The resulting point and interval estimators agree exactly with their frequentist counterparts

Flat Priors are not Invariant

- Let $X_n \sim \mathrm{Bernoulli}(p)$ and $f(p) = 1$, representing our lack of information about p before the experiment
- Now consider $\psi = \log p/(1-p)$, the density for ψ now is

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}$$

  not a flat prior
- But if we are ignorant about p then we are also ignorant about ψ so we should use a flat prior for ψ
- Flat priors are not transformation invariant

Jeffreys' Prior

- $f(\theta) \propto I(\theta)^{1/2}$, and it's transformation invariant
- EXAMPLE. Consider $\sim \mathrm{Bernoulli}(p)$. The Fisher information

$$I(\theta) = \frac{1}{p(1-p)}, \quad f(\theta) \propto \frac{1}{\sqrt{p(1-p)}}$$

- This is $\mathrm{Beta}(1/2, 1/2)$, not too different from flat prior
- Multiparameter case: $f(\theta) \propto (\det I(\theta))^{1/2}$

Revisit Linear Gaussian Systems

- Variable of interest $x \in R^{D_x}$ and observation $y \in R^{D_y}$
- Prior on x

$$p(x) = \mathcal{N}(x|\mu_x, \Sigma_x)$$

  and likelihood

$$p(y|x) = \mathcal{N}(y|Ax + b, \Sigma_y)$$

  Notice that the dependency of mean on x is a linear function of x
- The **posterior**

$$p(x|y) = \mathcal{N}(x|\mu_{x|y}, \Sigma_{x|y})$$

  In this case, the posterior is still Gaussian
- We can think of x as the latent variable, y the observed variable giving rise to the data

$$p(y) = \int p(y|x)p(x)dx$$

Auto-Encoding Variational Bayes (VAE)

- We consider a setting similar to that used in EM algorithm: iid data

$$Y = \{y_1, \ldots, y_N\} \sim p_\theta(y)$$

- Data generating process: Ancestral sampling
  1) sample $x_i \sim p_\theta(x)$; and 2) sample $y_i \sim p_\theta(y|x_i)$.
  θ is the **model parameter**, and x_i is the **latent variables**
- We are interested in
  1. MLE or MAP estimate of θ
  2. Approximate posterior inference of x given y and a value for θ (data encoding and representation)
  3. Marginal inference for x (image denoising and super-resolution)
- $p_\theta(y, x) = p_\theta(y|x)p_\theta(x)$ is the **generative model**, and $p_\theta(y|x)$ is the **decoder**

Auto-Encoding Variational Bayes (VAE)

- In particular, we use $q_\phi(x|y)$, a **recognition model**, also called an **encoder**, to approximate $p_\theta(x|y)$

$$p_\theta(x|y) = \frac{p_\theta(y|x)p_\theta(x)}{\int p_\theta(y|x)p_\theta(x)dx}, \quad q_\phi(x|y) \approx p_\theta(x|y)$$

- Consider the prior $p_\theta(x) = \mathcal{N}(x|0, I), p_\theta(y|x) = \mathcal{N}(y|\mu(x), \sigma(x)^2 I)$

$$\mu = W_1 h + b_1, \ \log \sigma^2 = W_2 h + b_2, \quad h = \tanh(W_3 x + b_3)$$

  $\mu(x)$ is a nonlinear function of x, and even $\sigma(x)^2$ is a constant, $p_\theta(x|y)$ in no longer Gaussian
- We assume that $q_\phi(z|x_i) = \mathcal{N}(z|\mu(x_i, \phi), \sigma(x_i, \phi)^2 I)$ with $\mu(x_i, \phi), \sigma(x_i, \phi)^2$ similarly built as above, and ϕ are the corresponding weights