STAT 576 Bayesian Analysis

Lecture 1: Review on Prerequisites

Chencheng Cai

Washington State University

Measurable Space

- $(\Omega, \mathcal{E})$ is called a **measurable space** if Ω is a nonempty set and $\mathcal{E}$ is a σ-algebra on Ω.
- The σ-algebra $\mathcal{E}$ on Ω is a collection of subsets of Ω such that
  - $\Omega \in \mathcal{E}$;
  - if $E \in \mathcal{E}$, then $E^c \in \mathcal{E}$; (closed under complementation)
  - if $E_1, E_2, \cdots \in \mathcal{E}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{E}$. (closed under countable union)
- A **measure** m on $(\Omega, \mathcal{E})$ is a set function $m : \mathcal{E} \to \bar{\mathbb{R}}$ such that
  - $m(\varnothing) = 0$;
  - $m(E) \geq 0$ for all $E \in \mathcal{E}$;
  - if $\{E_i\}_{i=1}^{\infty}$ are pairwise **disjoint** sets in $\mathcal{E}$, then $m\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} m(E_i)$.
- Common measurable spaces:
  - for $\Omega = \mathbb{R}$, $\mathcal{E}$ contains all **Borel** sets, and we use the **Lebesgue measure**:

$$\mu((a, b)) = b - a$$

  - for $\Omega = \mathbb{Z}$, $\mathcal{E}$ contains all the subsets, and we use the **counting measure**:

$$\mu(E) = \begin{cases} |E| & \text{if } E \text{ finite} \\ +\infty & \text{if } E \text{ infinite} \end{cases}$$

Probability Space

- A measurable space $(\Omega, \mathcal{E}, \mathbb{P})$ is called a **probability space** if $\mathbb{P}(\Omega) = 1$.
- In this case,
  - Ω: sample space.
  - $E \in \mathcal{E}$: event.
  - $\mathbb{P}(E)$ for $E \in \mathcal{E}$: the probability of event E.
  - $\mathbb{P}$ is called the **probability measure**
- Example:
  - $\Omega = [0, 1]$, $\mathcal{E}$ is all Borel sets restricted to $[0, 1]$.
  - Then (1) the set of all rational numbers $\mathbb{Q} \cap [0, 1]$ is measurable.
  - and (2) $\mathbb{P}(\mathbb{Q} \cap [0, 1]) = 0$.

Random Variable

- A random variable $X(\omega)$ is a **measurable** function mapping from a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to a measurable space $(\Omega_X, \mathcal{X})$.
- Here **measurable** means
  - for any $E_X \in \mathcal{X}$, its preimage $X^{-1}(E_X)$ is measurable, i.e. $X^{-1}(E_X) \in \mathcal{E}$.
- Then we can define a probability measure $\mathbb{P}_X$ on $(\Omega_X, \mathcal{X})$ for any $E_X \in \mathcal{X}$ by

$$\mathbb{P}_X(E_X) = \mathbb{P}(X^{-1}(E_X)).$$

- Example:
  - Consider $\Omega = \mathbb{Z}$ with $\mathcal{E}$ all of its subsets, and $\mathbb{P}$ some probability measure on it.
  - Let $X(\omega) = |\omega| \in \mathbb{Z}^+$
  - Then for any $a \in \mathbb{Z}^+$,

$$\mathbb{P}_X(X = a) = \mathbb{P}(X^{-1}(\{a\})) = \mathbb{P}(\{a, -a\}) = \mathbb{P}(a) + \mathbb{P}(-a)$$

Distribution

- For a probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$, its **distribution function** $F_X : \mathbb{R} \to [0,1]$ is

$$F_X(t) = \mathbb{P}_X((-\infty, t])$$

- The distribution function is *cadlag*:
  - *continue à droite*: $\lim_{t\uparrow c} F(t)$ exists for all c.
  - *limite à gauche*: $\lim_{t\downarrow c} F(t) = F(c)$ for all c.
- The distribution function is non-decreasing.
- $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$
- The probability space is uniquely determined by its distribution function because

$$\mathbb{P}_X((a, b]) = F_X(b) - F_X(a)$$

Distribution

- The probability measure $\mathbb{P}_X$ is **absolutely continuous** with respect to the Lebesgue measure μ if

$$\mathbb{P}_X(E) = 0 \quad \text{whenever} \quad \mu(E) = 0.$$

- If $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure μ, we call $p : \mathbb{R} \to \mathbb{R}$ the **probability density function** of F_X if

$$\mathbb{P}_X(E) = \int_E p d\mu$$

  for any Borel set E.
- The Leibniz rule gives $F'(t) = p(t)$.
- Similarly, if we replace all previous arguments for Lebesgue measure to counting measure, the corresponding p is called the **probability mass function**.

Expectation

- Suppose the random variable X is in a probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ with distribution function F that is absolutely continuous to the Lebesgue measure.
- Let f be a measurable function of X. Then the expecation of f can be written as
  - classical Rieman integral: $\int_{-\infty}^{\infty} f(x)p(x)dx$.
  - Lebesgue integral: $\int_{\mathbb{R}} f(x)p(x)d\mu$
  - or simply
$$\int_{\mathbb{R}} f(x)dF(x)$$
- Since the formula is almost the same for continous and discrete random variables, except that the base measure μ is Lebesgue (for continuous r.v.) and counting (for discrete), we simply use the integral for all types of random variables.

Estimation

- We observe X from a distributon from a distribution family $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.
- The distribution family $\mathcal{F}$ is called **identifiable** if for any $\theta \neq \theta'$

$$\sup_t |F_\theta(t) - F_{\theta'}(t)| > 0$$

- The left-hand side is called Komogorov-Smirnov distance.
- If the distribution F_θ has a density function p_θ for all θ, the **likelihood** function is

$$L(\theta) = p_\theta(X)$$

- The **score** function is

$$\dot{\ell}(\theta) = \frac{\partial}{\partial\theta} \log L(\theta)$$

- The **Fisher's information** is

$$I(\theta) = \mathbb{E}_\theta[(\dot{\ell}(\theta))^2] = -\mathbb{E}_\theta[\ddot{\ell}(\theta)] = -\int \ddot{\ell} dF_\theta$$

Maximum Likelihood Estimator

- The **Maximum Likelihood Estimator (MLE)** is

$$\hat{\theta} = \arg\max_{\theta} \ L(\theta) = \arg\max_{\theta} \ \ell(\theta)$$

- If ℓ is differentiable and $\hat{\theta}$ is an interior point of Θ, then

$$\dot{\ell}(\hat{\theta}) = 0.$$

  The above is called the **estimating equation(s)**.

- Counter-example: $X \sim \mathrm{unif}[0, \theta]$.

Consistency of MLE

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples drawn from F_{θ_0} for some $\theta_0 \in \Theta$.
- The log-likelihood function is now

$$\ell_n(\theta) = \log \prod_{i=1}^{n} p_\theta(X_i) = \sum_{i=1}^{n} \log p_\theta(X_i)$$

- (1) $\hat{\theta}_n$ maximizes $\ell_n(\theta)$.
- (2) By the Law of Large Numbers, we have

$$n^{-1}\ell_n(\theta) \to \mathbb{E}_{\theta_0}[\log p_\theta(X)] =: \ell(\theta)$$

- (3) We can show that θ_0 is the maximum of the (point-wise) limit function $\ell(\theta)$:

$$\ell(\theta) < \ell(\theta_0) \quad \text{for all } \theta \neq \theta_0.$$

- Under certain regularity conditions (uniform convergence of ℓ_n), with (1)-(3), we have

$$\hat{\theta}_n \xrightarrow{P} \theta_0.$$

CLT for MLE

- We can have a Taylor expansion of $\dot{\ell}$ at θ_0:

$$0 = \dot{\ell}_n(\hat{\theta}_n) = \dot{\ell}_n(\theta_0) + (\hat{\theta}_n - \theta_0)\ddot{\ell}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 \dddot{\ell}_n(\theta'),$$

  for some θ' between θ_0 and $\hat{\theta}$.
- Then we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\dot{\ell}_n(\theta_0)}{\ddot{\ell}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_n)\dddot{\ell}_n(\theta')}$$

- The CLT gives $\sqrt{n}\dot{\ell}_n(\theta_0) \xrightarrow{D} \mathcal{N}(0, I(\theta))$.
- The LLN gives $\ddot{\ell}_n(\theta_0) \xrightarrow{P} I(\theta)$.
- If $\dddot{\ell}_n$ is bounded, by consistency, we have $(\hat{\theta}_n - \theta_n)\dddot{\ell}_n(\theta') \xrightarrow{P} 0$.
- By Slutsky's lemma, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}(0, I^{-1}(\theta))$$

Example

- Let $X_1, \ldots, X_n$ be i.i.d. Binomial distribution with size K (fixed) and probability $\theta \in (0,1)$.
- Likelihood function:

$$L(\theta) = \prod_{i=1}^{n} \binom{K}{X_i} \theta^{X_i} (1-\theta)^{K-X_i}$$

- The log-likelihood function:

$$\ell(\theta) = S_n \log \theta + (nK - S_n) \log(1-\theta) + C,$$

  where $S_n = \sum_{i=1}^{n} X_i$ and C is a constant of θ.
- The score function is

$$\dot{\ell}(\theta) = \frac{S_n}{\theta} + \frac{S_n - nK}{1-\theta}$$

- By setting the score function to 0, we have

$$\hat{\theta}_n = \frac{S_n}{nK}$$

Example

- The consistency is followed by LLN:

$$\frac{S_n}{n} \xrightarrow{P} \mathbb{E}[X_1] = K\theta$$

- The CLT is followed by the CLT of S_n:

$$n^{-1/2} S_n \xrightarrow{D} \mathcal{N}(0, I(\theta))$$

- where the Fisher's information is

$$I(\theta) = -\mathbb{E}[\ddot{\ell}(\theta)] = \mathbb{E}_\theta \left[\frac{X_1}{\theta^2} + \frac{K - X_1}{(1-\theta)^2} \right] = \frac{K}{\theta(1-\theta)}$$

- Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{\theta(1-\theta)}{K}\right)$$

Counter-example

- Let $X_1, \ldots, X_n$ be i.i.d. from unif$[0, \theta]$ with $\theta \in \mathbb{R}^+$.
- The likelihood function is

$$L_n(\theta) = \prod_{i=1}^{n} \frac{\mathbb{I}\{X_i \leq \theta\}}{\theta} = \frac{\mathbb{I}\{X_{(n)} \leq \theta\}}{\theta^n}$$

- The likelihood is **not** differentiable, but we can maximize it directly to have $\hat{\theta}_n = X_{(n)}$.
- The consistency is followed by that for any $0 < \epsilon < \theta$,

$$\mathbb{P}[|\hat{\theta}_n - \theta| > \epsilon] = \left(1 - \frac{\epsilon}{\theta}\right)^n \to 0.$$

Counter-example

- We have the distribution function for $n(\theta - \hat{\theta}_n)$ as

$$F(t) = 1 - \mathbb{P}[\hat{\theta}_n \leq \theta - t/n] = 1 - \left(1 - \frac{t}{n\theta}\right)^n \to 1 - e^{-t/\theta}$$

- Therefore, we have the limit distribution of $\hat{\theta}_n$ as

$$n(\theta - \hat{\theta}_n) \xrightarrow{D} \operatorname{Exp}\left(\theta^{-1}\right)$$

- The CLT does not hold for this example.