Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 6

Yun Yang

- Uniform laws of large numbers

Uniform convergence of CDFs

First example of a uniform law of large numbers.

Suppose $X_1, \ldots, X_n$ are i.i.d. with CDF F. Define the empirical CDF as

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty,\, t]}(X_i) \quad \text{for all } t \in \mathbb{R}.$$

Theorem (Glivenko-Cantelli)

Empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF F,

$$\|\widehat{F}_n - F\|_\infty \overset{a.s.}{\to} 0,$$

where $\|F - G\|_\infty = \sup_{t \in \mathbb{R}} |F(t) - G(t)|$ is the supreme norm of $F - G$.

Uniform convergence of CDFs

Why it is a uniform law of large numbers?

$$\|\widehat{F}_n - F\|_\infty = \sup_t \left|\mathbb{P}_n(X \le t) - \mathbb{P}(X \le t)\right| \xrightarrow{\text{a.s.}} 0,$$

where $\mathbb{P}_n$ is the empirical measure

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

For any fixed t, the LLN says that $\left|\mathbb{P}_n(X \le t) - \mathbb{P}(X \le t)\right| \xrightarrow{\text{a.s.}} 0$. The Glivenko-Cantelli theorem says that this happens uniformly over all $t \in \mathbb{R}$.

Application of Glivenko-Cantelli theorem

In many estimation problems, the quantity of interest can be formulated as $\theta(F)$, where the functional θ maps any CDF F to a real number $\theta(F)$.

Plug-in principle

Estimating $\theta(F)$ by replacing the unknown F with $\widehat{F}_n$, yielding a plug-in estimator $\theta(\widehat{F}_n)$.

Examples

- Mean: $\theta(F) = \int x \, dF(x)$, and $\theta(\widehat{F}_n) = n^{-1} \sum_{i=1}^n X_i$.
- Quantile: $\theta(F) = \int \{x : F(x) \geq \alpha\}$, the α-quantile, and

$$\theta(\widehat{F}_n) = \inf \left\{ x : \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \geq \alpha \right\}.$$

If θ is continuous w.r.t. $\| \cdot \|_\infty$, then we get $\theta(\widehat{F}_n) \stackrel{\text{a.s.}}{\to} \theta(F)$.

Empirical process

- Let $\mathcal{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$.
- Let $X_1^n = (X_1, \ldots, X_n)$ be a collection of i.i.d. samples from $\mathbb{R}$ over $\mathcal{X}$.
- For any probability measure $\mathcal{Q}$ and function $f \in \mathcal{F}$, denote $\mathcal{Q}f = \mathbb{E}_{X\sim Q}[f(X)]$.
- The stochastic process $\mathbb{P}_n - \mathbb{P} = \{\mathbb{P}_n f - \mathbb{P}f : f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ is called an empirical process over $\mathcal{F}$.
- Define random variable (measurability issue?)

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\mathbb{P}_n f - \mathbb{P}f\right| = \sup_{f\in\mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f]\right|.$$

Glivenko-Cantelli class

Definition

$\mathcal{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$ if

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0 \text{ in probability as } n \to \infty.$$

Example: Empirical CDF

Consider the function class

$$\mathcal{F} = \{\mathbb{I}_{(-\infty, t]}(\cdot) : t \in \mathbb{R}\}.$$

For each fixed t, we have $\mathbb{P}_n \mathbb{I}_{(-\infty, t]} = F_n(t)$ and $\mathbb{P} \mathbb{I}_{(-\infty, t]} = F(t)$. Therefore, the classical Glivenko-Cantelli theorem implies $\mathcal{F}$ is a Glivenko-Cantelli class.

Note: not all classes of functions are Glivenko-Cantelli (counter-example?).

Empirical risk minimization

Variables of form $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ are ubiquitous in statistics.

- Given n i.i.d. samples $X_1^n = (X_1, \ldots, X_n)$ from an unknown distribution $\mathbb{P}$
- Θ is the space of all prediction rules, hypotheses, or parameters
- We have a loss function $\ell(\theta, x)$ that measures how bad it is to choose $\theta \in \Theta$ when the outcome is x.

Definition

For $X \sim \mathbb{P}$, the (population) risk is defined as $\mathcal{L}(\theta) = \mathbb{P}\,\ell(\theta, X)$.

We want to choose a $\theta \in \Theta$ that minimizes the population risk. Denote the minimizer by θ^*.

Empirical risk minimization

However, we cannot directly minimize the population risk $\mathcal{L}(\theta)$, since the underlying data generating distribution $\mathbb{P}$ is unknown. Instead, we consider the following surrogate.

Definition

For $X_1, \ldots, X_n$ i.i.d. from $\mathbb{P}$, the empirical risk is defined as

$$\mathcal{L}_n(\theta) = \mathbb{P}_n \, \ell(\theta, X) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, X_i).$$

Empirical risk minimization aims to minimize the empirical risk:

$$\widehat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathcal{L}_n(\theta).$$

We can quantify its performance via the excess risk

$$\mathcal{L}(\widehat{\theta}\,) - \inf_{\theta \in \Theta} \mathcal{L}(\theta).$$

Example: Maximum likelihood

- Suppose we have a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$, each $\mathbb{P}_\theta$ admits a density p_θ.
- The true underlying distribution $\mathbb{P} = \mathbb{P}_{\theta^*}$ for some unknown parameter θ^*
- Define loss function

$$\ell(\theta, x) = \log \frac{p_{\theta^*}(x)}{p_\theta(x)}.$$

- The population risk is the Kullback-Leibler divergence between p_{θ^*} and p_θ,

$$\mathbb{P}_{\theta^*} \log \frac{p_{\theta^*}}{p_\theta},$$

which attains minimum zero at $\theta = \theta^*$.
- Empirical risk minimization corresponds to the MLE.

Example: Binary classification

- Have n i.i.d. samples $(X_i, Y_i) \in \mathcal{X} \times \{0, 1\}$ from some unknown distribution $\mathbb{P}$.
- Want to find a best prediction rule $\theta : \mathcal{X} \to \{0, 1\}$ to predict the binary part Y from X.
- The loss function is the 0-1 loss

$$\ell(\theta,\ (x, y)) = \mathbb{I}(\theta(x) \neq y).$$

- The population risk is the mis-classification probability $\mathbb{P}(\theta(X) \neq Y)$, which is minimized at the Bayes classifier

$$\theta^*(x) = \begin{cases} 0, & \text{if } \mathbb{P}(Y = 1 \,|\, X = x) \leq \mathbb{P}(Y = 0 \,|\, X = x), \\ 1, & \text{if } \mathbb{P}(Y = 1 \,|\, X = x) > \mathbb{P}(Y = 0 \,|\, X = x). \end{cases}$$

- Empirical risk minimization chooses θ to minimize mis-classifications on the sample.

Control excess risk

Recall: θ^* is the population risk minimizer, and $\widehat{\theta}$ is the empirical risk minimizer.

Excess risk decomposition:

$$\mathcal{L}(\widehat{\theta}\,) - \inf_{\theta \in \Theta} \mathcal{L}(\theta) =$$
$$\Big[\mathcal{L}(\widehat{\theta}\,) - \mathcal{L}_n(\widehat{\theta}\,)\Big] + \Big[\mathcal{L}_n(\widehat{\theta}\,) - \mathcal{L}_n(\theta^*)\Big] + \Big[\mathcal{L}_n(\theta^*) - \mathcal{L}(\theta^*)\Big].$$

The middle term is non-positive because $\widehat{\theta}$ is chosen to minimize $\mathcal{L}_n$. Therefore, we have

$$\mathcal{L}(\widehat{\theta}\,) - \inf_{\theta \in \Theta} \mathcal{L}(\theta) \leq 2 \sup_{\theta \in \Theta} \Big|\mathcal{L}_n(\theta) - \mathcal{L}(\theta)\Big| = 2\,\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{L}},$$

where $\mathfrak{L} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$.

Rademacher complexity

For any fixed collection $x_1^n = (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathcal{F}(x_1^n) = \Big\{ \big(f(x_1), \ldots, f(x_n)\big) \,\Big|\, f \in \mathcal{F} \Big\}.$$

Recall that the Ramemacher complexity of this set (rescaled by n^{-1}) is defined by

$$\mathcal{R}\big(\mathcal{F}(x_1^n)/n\big) = \mathbb{E}_\varepsilon \Bigg[\sup_{f \in \mathcal{F}} \bigg| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \bigg| \Bigg],$$

which is called the empirical Rademacher complexity.

Definition

Given random samples $X_1^n = (X_1, \ldots, X_n)$, the Rademacher complexity of the function class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X \big[\mathcal{R}\big(\mathcal{F}(x_1^n)/n\big) \big] = \mathbb{E}_{X,\,\varepsilon} \Bigg[\sup_{f \in \mathcal{F}} \bigg| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \bigg| \Bigg].$$

A uniform law via Rademacher complexity

Rademacher complexity characterizes the typical largest correlation between a random noise vector and any function in the class $\mathcal{F}$, thereby the "complexity" of $\mathcal{F}$.

Theorem

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ that is uniformly bounded by $b > 0$. Then for all $n > 0$ and $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\,\mathcal{R}_n(\mathcal{F}) + \delta$$

with $\mathbb{P}$ probability at least $1 - 2\,\exp\big(-\frac{n\delta^2}{8b^2}\big)$. Consequently, $\mathcal{R}_n(\mathcal{F}) = o(1)$ implies $\mathcal{F}$ to be Glivenko-Cantelli.

Proof step one: Concentration around mean

Consider the function

$$G(x_1, \ldots, x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right|.$$

It satisfies the bounded difference property: for all $x_1, \ldots, x_n, x'_k \in \mathbb{R}$,

$$\left| G(x_1, \ldots, x_n) - G(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n) \right| \leq \frac{2\|f\|_\infty}{n} \leq \frac{2b}{n}.$$

Therefore, the bounded difference inequality implies the following holds with probability at least $1 - 2\exp\left(-\frac{n\,t^2}{8b^2}\right)$,

$$\Big| \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \Big| \leq t, \quad \text{for any } t > 0.$$

Proof step two: Upper bound on mean

Applying the symmetrization technique.

Let $(Y_1, \ldots, Y_n)$ be a second independent copy of $(X_1, \ldots, X_n)$. Then

$$\begin{aligned}
\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &= \mathbb{E}_X\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}\Big|\Big] \\
&= \mathbb{E}_X\Big[\sup_{f\in\mathcal{F}}\Big|\mathbb{E}_Y\big[\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\big]\Big|\Big] \\
&\leq \mathbb{E}_{X,Y}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\Big|\Big],
\end{aligned}$$

where the last step is due to Jensen's inequality.

Proof step two: Upper bound on mean

Let ε_i be i.i.d. Rademacher random variables.

For any $f \in \mathcal{F}$, random variable $\varepsilon_i(f(X_i) - f(Y_i))$ has the same distribution as $f(X_i) - f(Y_i)$. Consequently,

$$
\begin{aligned}
\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &\leq \mathbb{E}_{X,Y}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big\{f(X_i) - f(Y_i)\big\}\Big|\Big] \\
&\leq 2\mathbb{E}_{X,\varepsilon}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\Big] = 2\mathcal{R}_n(\mathcal{F}).
\end{aligned}
$$