

第七讲

王伟文 暨南大学

1 累积分布函数一致收敛性

给定服从某一分布的随机变量 X , 其累积分布函数 (*Cumulative Distribution Function, CDF*) 表示为

$$F(t) := \mathbb{P}[X \leq t] \quad \forall t \in \mathbb{R}.$$

设 $\{X_i\}_{i \in [n]}$ 为一组与 X 同分布且相互独立的样本, X 的一个经验 CDF 估计可以是

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, t]}[X_i].$$

因为

$$\begin{aligned} F(t) &= \mathbb{P}[X \leq t] = \int_{-\infty}^t p(x) dx \\ &= \int_{-\infty}^{+\infty} \mathbb{I}_{(-\infty, t]}[x] p(x) dx \\ &= \mathbb{E}[\mathbb{I}_{(-\infty, t]}[X]] \end{aligned}$$

故 $\mathbb{E}[\hat{F}_n(t)] = F(t)$.

给定定义域为 \mathbb{R} 的函数 f 和 g , 记两个函数间的距离为

$$\|f - g\|_{\infty} := \sup_{t \in \mathbb{R}} |f(t) - g(t)|.$$

定理 1.1: Glivenko-Cantelli 定理

对任意分布, $\hat{F}_n(t)$ 是 $F(t)$ 的强一致估计, 即

$$\|\hat{F}_n - F\|_{\infty} \xrightarrow{a.s.} 0,$$

换言之,

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \|\hat{F}_n - F\|_{\infty} = 0\right) = 1.$$

2 一般函数类的一致收敛性

考虑定义域为 \mathcal{X} 的可积实值函数类 \mathcal{F} , $\{X_i\}_{i \in [d]}$ 为相互独立的样本, 源自以 \mathcal{X} 为支撑集的分布 \mathbb{P} , 定义随机变量

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

累积概率密度函数. 取

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

此时

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\hat{F}_n - F\|_{\infty}$$

经验风险最小化. 考虑一族分布 $\{\mathbb{P}_{\theta} : \theta \in \Omega\}$. 给定一组独立同分布样本 $\{X_i\}_{i \in [n]}$, 产生于某一未知分布 $\mathbb{P}_{\theta^*} (\theta^* \in \Omega)$. 经验风险最小化从最小化如下经验风险估计未知参数 θ^* .

经验风险 (Empirical Risk)

$$\hat{R}_n(\theta, \theta^*) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\theta}(X_i)$$

经验风险最小化对未知参数 θ^* 的估计

$$\hat{\theta} \in \arg \min_{\theta \in \Omega} \hat{R}_n(\theta, \theta^*),$$

其中 \mathcal{L}_{θ} 表示损失函数.

对任意 θ , 其总体风险 (Population Risk) 定义为

$$R_n(\theta, \theta^*) = \mathbb{E} [\hat{R}_n(\theta, \theta^*)] = \mathbb{E}_{\theta^*} [\mathcal{L}_{\theta}(X)]$$

实际上, θ^* 是否属于参数空间 Ω 往往是未知的, 此时一个关键问题是如何控制 $\hat{\theta}$ 的超额风险.

超额风险 (Excess Risk)

$$\mathcal{E}_n(\hat{\theta}, \theta^*) := R_n(\hat{\theta}, \theta^*) - \inf_{\theta \in \Omega} R_n(\theta, \theta^*)$$

例 2.1. (极大似然估计) 记分布簇 $\{\mathbb{P}_{\theta} : \theta \in \Omega\}$ 每一个元素对应概率密度函数 p_{θ} , 定义损失函数

$$\mathcal{L}(x) := \log \frac{p_{\theta^*}(x)}{p_{\theta}(x)}$$

参数 θ^* 经验风险最小化估计为

$$\begin{aligned} \hat{\theta} \in \arg \min_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta^*}(X_i)}{p_{\theta}(X_i)} &= \arg \min_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{p_{\theta}(X_i)} \\ &= \arg \max_{\theta \in \Omega} \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \end{aligned}$$

为了控制超额风险 $\mathcal{E}_n(\hat{\theta}, \theta^*)$, 需要对其进行分解. 假设存在 θ_0 使得 $R_n(\theta_0, \theta^*) = \inf_{\theta \in \Omega} R_n(\theta, \theta^*)$.

$$\mathcal{E}_n(\hat{\theta}, \theta^*) = \underbrace{R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*)}_{T_1} + \underbrace{\hat{R}_n(\hat{\theta}, \theta^*) - \hat{R}_n(\theta_0, \theta^*)}_{T_2} + \underbrace{\hat{R}_n(\theta_0, \theta^*) - R_n(\theta_0, \theta^*)}_{T_3}$$

1. 由 $\hat{\theta}$ 的最优性, $\hat{R}_n(\hat{\theta}, \theta^*) \leq \hat{R}_n(\theta_0, \theta^*)$, 故 $T_2 \leq 0$,

2.

$$\begin{aligned} T_1 = R(\hat{\theta}, \theta^*) - \hat{R}_n(\hat{\theta}, \theta^*) &\leq \left| \mathbb{E}_{\theta^*}[\mathcal{L}_{\hat{\theta}}(X)] - \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\hat{\theta}}(X_i) \right| \\ &\leq \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\theta}(X_i) - \mathbb{E}[\mathcal{L}_{\theta}(X)] \right|. \end{aligned}$$

3. 类似 2

$$T_3 = \hat{R}_n(\theta_0, \theta^*) - R_n(\theta_0, \theta^*) \leq \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\theta}(X_i) - \mathbb{E}[\mathcal{L}_{\theta}(X)] \right|,$$

综上,

$$\mathcal{E}_n(\hat{\theta}, \theta^*) \leq 2 \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\theta}(X_i) - \mathbb{E}[\mathcal{L}_{\theta}(X)] \right| = 2 \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{L}(\Omega)}.$$

其中

$$\mathcal{L}(\Omega) \triangleq \{\mathcal{L}_{\theta} : \theta \in \Omega\}$$

3 基于 Rademacher 复杂度的统一律

定义 3.1: 函数类 \mathcal{F} 的 Rademacher 复杂度

给定分布 \mathcal{P} 独立产生的 n 个数据点 $(x_i)_{i \in [n]}$, 则函数类 \mathcal{F} 关于这 n 个数据点的经验 Rademacher 复杂度定义为

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|,$$

将数据点替换为 n 个独立同分布的随机变量 $X = (X_i)_{i \in [n]}$, 并关于 X 求期望, 得到函数类 \mathcal{F} 的 Rademacher 复杂度

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

定理 3.1

设存在 $c \geq 0$, 对于任意 $f \in \mathcal{F}$, 有 $\|f\|_{\infty} \leq c$. 对任意 $n \geq 1$ 及 $\delta \geq 0$, 不等式

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

至少以概率 $1 - \exp(-\frac{n\delta^2}{2c^2})$ 成立.

证明. 1. 有界差不等式控制 $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$.

定义函数

$$g(x) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)] \right|, \quad x = (x_1, x_2, \dots, x_n).$$

因为 $\|f\|_{\infty} \leq c$, 即 $\sup_x |f(x)| \leq c$, 故

$$\begin{aligned} g(x) - g(x^{\setminus k}) &= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i \neq k}^n f(x_i) + \frac{1}{n} f(x_k) - \mathbb{E}[f(X)] \right| - \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i \neq k}^n f(x_i) + \frac{1}{n} f(x'_k) - \mathbb{E}[f(X)] \right| \\ &\leq \sup_{f \in \mathcal{F}} \left\{ \left| \frac{1}{n} \sum_{i \neq k}^n f(x_i) + \frac{1}{n} f(x_k) - \mathbb{E}[f(X)] \right| - \left| \frac{1}{n} \sum_{i \neq k}^n f(x_i) + \frac{1}{n} f(x'_k) - \mathbb{E}[f(X)] \right| \right\} \\ &\leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} f(x_k) - \frac{1}{n} f(x'_k) \right| \\ &\leq \frac{2c}{n}. \end{aligned}$$

同理, $g(x^{\setminus k}) - g(x) \leq \frac{2c}{n}$. 因此 $g(x)$ 满足以 $\frac{2n}{c}$ 为参数的有界差不等式.

$$\mathbb{P} [\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \delta] \leq e^{-\frac{n\delta^2}{2c^2}} \quad \forall \delta \geq 0.$$

因此不等式

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} + \delta$$

至少以概率 $1 - e^{-\frac{n\delta^2}{2c^2}}$ 成立.

2. Rademacher 复杂度控制 $\mathbb{E}\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$

此过程应用对称化技巧 (symmetrization trick), 记 $X = (X_1, X_2, \dots, X_n)$, X' 为 X 的独立副

本.

$$\begin{aligned}
\mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} &= \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \\
&= \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{X'} \left[\frac{1}{n} \sum_{i=1}^n f(X'_i) \right] \right| \\
&= \mathbb{E}_X \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X'} \left[\frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right] \right| \\
&\leq \mathbb{E}_X \sup_{f \in \mathcal{F}} \mathbb{E}_{X'} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right| \quad (\text{Jensen 不等式}) \\
&\leq \mathbb{E}_{X, X'} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right| \quad (\mathbb{E}h(X) \leq \mathbb{E} \sup_{h \in \mathcal{H}} h(X), \forall h \in \mathcal{H}) \\
&= \mathbb{E}_{X, X'} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (f(X_i) - f(X'_i)) \right| \\
&\leq \mathbb{E}_{X, X'} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| + \mathbb{E}_{X, X'} \mathbb{E}_{\epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X'_i) \right| \\
&= 2\mathcal{R}_n(\mathcal{F})
\end{aligned}$$

3. 综合上述结论, 有

$$\begin{aligned}
\mathbb{P} [\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta] &\geq \mathbb{P} [\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} + \delta] \\
&\geq 1 - e^{-\frac{n\delta^2}{2c^2}}.
\end{aligned}$$

□