Theoretical Machine Learning

Lectured by Zhihua Zhang LaTeXed by Chengxin Gong

February 21, 2024

Contents

1 简介

- 机器学习的主要任务: 生成、预测、决策. 生成: $X_1, \cdots, X_n \sim F$, 推断分析 F, 无监督学习, GAN, GPT, $\cdots$. 预测: 数据对 $(X^{(1)}, Y^{(1)}), \cdots, (X^{(n)}, Y^{(n)})$, $X^{(i)} \in \mathbb{R}^d$ 输入变量, $f: \mathcal{X} \to \mathcal{Y}, x \in \mathcal{X}, y \in \mathcal{Y}$, 归因, 有监督学习. 决策: 强化学习, Agent$\leftarrow$action, state, reward$\rightarrow$ 环境.

- 求解问题的途径: 参数/非参数, 频率 (MLE)/贝叶斯.

- 误差模型: 有监督: $X = (X_1, \cdots, X_d)^T \in \mathbb{R}^d$, 回归: $Y \in \mathbb{R}$; 分类: $Y \in \{0,1\}(\{-1,1\}, \{1, \cdots, M\}, \{0,1\}^M)$; X 随机, Random design(生成模型), $Y = g(X) + \epsilon \overset{\text{or}}{=} g(X, Z), Y^{(i)} = g(X^{(i)}, Z^{(i)})$; X 固定 $X = x$, Fixed design(判别模型), $Y^{(i)} = g(x^{(i)}, Z^{(i)})$. 无监督: $X = g(Z)$(因子模型: $X = AZ + \epsilon, Z \in \mathcal{N}(0,1), \epsilon \sim \mathcal{N}(0, \Sigma)$).

2 统计决策理论

- Consider a state space Ω, data space $\mathcal{D}$, model $\mathcal{P} = \{p(\theta, x)\}$, action space $\mathscr{A}$. Loss function: $\mathcal{L}: \Omega \times \mathscr{A} \to [-\infty, +\infty]$, measurable, nonnegative. A measurable function $\delta: \mathcal{D} \to \mathscr{A}$ is called a nonrandomized decision rule. Risk function is defined as $\mathcal{R}(\theta, \delta) = \int \mathcal{L}(\theta, \delta(x)) \mathrm{d}P_\theta(x) = \mathbb{E}_\theta \mathcal{L}(\theta, \delta(X))$. Randomized decision: for each $X = x$, $\delta(x)$ is a probability distribution: $[A|X = x] \sim \delta_x$. Risk function for δ: $\mathcal{R}(\theta, \delta) = \mathbb{E}_\theta \mathcal{L}(\theta, A) = \mathbb{E}_\theta \mathbb{E}_a \mathcal{L}(\theta, A|X) = \iint \mathcal{L}(\theta, a) \mathrm{d}\delta_x(a) \mathrm{d}P_\theta(x)$.

- Example [参数估计]: $\theta \in \Omega, \mathscr{A} = \Omega, \mathcal{L}(\theta, a) = \|\theta - a\|_2^2 \overset{\text{or}}{=} \|\theta - a\|_p^p (p \geq 1) \overset{\text{or}}{=} \int \log \frac{P_\theta(x)}{P_a(x)} P_\theta(x) \mathrm{d}m(x)$(KL). $\mathcal{R} = \mathrm{Var}(a) + \mathrm{bias}^2(a)$. Bregmass loss: $\phi: \mathbb{R}^d \to \mathbb{R}$ describe any strictly convex differentiable function. Then $\mathcal{L}_\phi(\theta, a) = \phi(a) - \phi(\theta) - (\phi - a)^T \nabla \phi(a)$.

- Example [Testing]: $\mathscr{A} = \{0, 1\}$ with action "0" associated with accepting $H_0: \theta \in \Omega_0$ and "1": $H_1: \theta \in \Omega_1$. δ_x is a Bernolli distribution. $\mathcal{L}(\theta, a) = I\{a = 1, \theta \in \Omega_0\} + I\{a = 0, \theta \in \Omega_1\}$. Risk $\mathcal{R}(\theta, \delta) = \mathbb{P}_\theta(A = 1) 1_{\theta \in \Omega_0} + \mathbb{P}_\theta(A = 0) 1_{\theta \in \Omega_1}$.

- A decision rule δ is called inadmissible if a competing rule δ^* such that $\mathcal{R}(\theta, \delta^*) \leq \mathcal{R}(\theta, \delta)$ for all $\theta \in \Omega$ and $\mathcal{R}(\theta, \delta^*) < \mathcal{R}(\theta, \delta)$ for at least one $\theta \in \Omega$. Otherwise, δ is admissible.

- The maximum risk $\bar{\mathcal{R}}(\delta) = \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$ and the Bayes risk $r(\Lambda, \delta) = \int \mathcal{R}(\theta, \delta) \mathrm{d}\Lambda(\theta)$ ($\Lambda(\theta)$ is a prior for θ). A decision rule that minimizes the Bayes risk is called a Bayes rule, that is, $\hat{\delta}: r(\Lambda, \hat{\delta}) = \inf_\delta r(\Lambda, \delta)$. Minimax rule $\delta^*: \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta^*) = \inf_\delta \sup_{\theta \in \Omega} \mathcal{R}(\theta, \delta)$.