Modern Statistical Modeling

Lectured by Wei Lin   LaTeXed by Chengxin Gong

2023 年 3 月 3 日

目录

1 Prediction and Nearest Neighbor

- Goal: (1) predict y from x ("black box"); (2) which variable(s) in x contributes to the prediction of y ("$x^T\beta$"), estimation, testing, variable selection.

- Why are prediction and estimation different: (1) model parameters; (2) identifiability ($f_{\theta_1} \neq f_{\theta_2} \Rightarrow \theta_1 \neq \theta_2$).

- Find prediction function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $\mathbb{E}_{X,Y}\mathcal{L}(f(X), Y) = \mathbb{E}\{\mathbb{E}(\mathcal{L}(f(X), Y)|X)\}$ where loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- Optimal predictor conditioned on x: $f^*(x) = \arg\min_{f(x)\in\mathcal{Y}} \mathbb{E}\{\mathcal{L}(f(X), Y)|X = x\}$.

- Regression: y numerical, squared error (L_2-loss) $\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$, $\mathbb{E}\{(Y - f(X))^2|X\} = \{\mathbb{E}(Y|X) - f(X)\}^2 + \mathbb{E}\{(Y - \mathbb{E}(Y|X))^2|X\} = \text{bias}^2 + \text{variance}$. Optimal $f^*(X) = \mathbb{E}(Y|X)$.

- To model f^*, $\begin{cases} \text{parametric: linear}, f*(x) = x^T\beta, \beta \in \mathbb{R}^2 \\ \text{nonparametric: infinite dimension}, f^*(x) = m(x), m \text{ satisfying certain smoothness} \end{cases}$.

- Classification: 0-1 loss $\mathcal{L}(\hat{y}, y) = I(\hat{y} = y)$, $\mathbb{E}\{\mathcal{L}(h(X), Y)|X = x\} = \sum_{j\neq h(x)} P(Y = j|X = x) = 1 - P(Y = h(X)|X = x)$. Optimal classification (Bayes classifier): $h^*(x) = \arg\max_{h(x)\in\mathcal{Y}} P(Y = h(X)|X = x)$.

- A fully nonparametric approach: k nearest neighbor (k-NN). Given training data $\{(x_i, y_i)\}_{i=1}^m$, use data "around" x to estimate $m(x) = \mathbb{E}(Y|X = x)$. Rationale: "Things that look alike must be alike". Classification: $h_{k\text{-NN}}(x) =$ majority label among $\{y_i, i \in N_k(x)\}$. Regression: $m_{k\text{-NN}}(x) = \frac{1}{k}\sum_{i\in N_k(x)} y_i$. k controls size of neighbor set. $k\uparrow$: effective sample size $\uparrow$, variance$\downarrow$, heterogeneity$\uparrow$, bias $\uparrow$.

- Theory for 1-NN: Consider binary classification: $\mathcal{Y} = \{0, 1\}$, $\mathcal{L}(h(x), y) = I(h(x) \neq y)$. Assume $\mathcal{X} \subset [-1, 1]^d$, ρ Euclidean distance, $S = \{(x_i, y_i)\}_{i=1}^n$. $\forall x \in \mathcal{X}$, let $\pi_1(x), \cdots, \pi_n(x)$ be an ordering of $\{1, \cdots, n\}$ with increasing distance to x. $\eta(x) = \mathbb{E}(Y = 1|X = x)$. Bayes classifier: $h^*(x) = I(\eta(x) > \frac{1}{2})$. Assumption on η: η is c-Lipschitz for some $c > 0$. Goal: Derive an upper bound on $\mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{L}(\hat{h}_S) = \mathbb{E}_{S\sim\mathcal{D}^n}\mathbb{E}_{(x,y)\sim\mathcal{D}}I(\hat{h}_S(x) \neq y)$.

- **Lemma** 1.1 The 1-NN rule $\hat{h}_S$ satisfies $\mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{L}(\hat{h}_S) \leq 2\mathcal{L}(h^*) + c\mathbb{E}_{S\sim\mathcal{D}^n, x\sim\mathcal{D}}||x - x_{\pi_1}(x)||$.

  Proof $\mathbb{E}_S\mathcal{L}(\hat{h}_S) = \mathbb{E}_{S_x\sim\mathcal{D}_x^n, x\sim\mathcal{D}_x, y\sim\eta(x), y'\sim\eta(\pi_1(x))}P(y \neq y')$. Note that $P(y \neq y') = \eta(x')(1 - \eta(x)) + (1 - \eta(x'))\eta(x) = (\eta - \eta + \eta')(1 - \eta) + (1 - \eta + \eta - \eta')\eta = 2\eta(1 - \eta) + (\eta - \eta')(2\eta - 1)$. Since η is c-Lipschitz and $|2\eta - 1| \leq 1$, $P(y \neq y') \leq 2\eta(1-\eta) + c||x - x'||$. Substituting back, $\mathbb{E}_S\mathcal{L}(\hat{h}_S) \leq 2\mathbb{E}_x\eta(x)(1-\eta(x)) + c\mathbb{E}_{S,x}||x - x_{\pi_1(x)}||$. The Bayes error $\mathcal{L}(h^*) = \mathbb{E}_x\{\eta(x) \wedge (1 - \eta(x))\} \geq \mathbb{E}_x(\eta(x)(1 - \eta(x)))$. $\square$

- **Lemma** 1.2 Let $C_1, \cdots, C_r$ be a collection of subsets of $\mathcal{X}$. Then $\mathbb{E}_{S\sim\mathcal{D}^n}\{\sum_{i:C_i\cap S=\emptyset}\}P(C_i) \leq \frac{r}{ne}$ ("probability of subsets that not hit by S").

  Proof By linearity, $\mathbb{E}_S\{\sum_{i:C_i\cap S=\emptyset} P(C_i)\} = \sum_{i=1}^r P(C_i)\mathbb{E}_S I(C_i \cap S = \emptyset) = \sum_{i=1}^r P(C_i)P(C_i \cap S = \emptyset)$. Note that $P(C_i \cap S = \emptyset) = (1 - P(C_i))^n \leq e^{-nP(C_i)}$. Thus, LHS $\leq \sum_{i=1}^r P(C_i)e^{-nP(C_i)} \leq r \max P(C_i)e^{-nP(C_i)} \leq \frac{r}{ne}$.$\square$