Modern Statistical Modeling

Lectured by Wei Lin LaTeXed by Chengxin Gong

2023 年 3 月 10 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 $\mathcal{X} \subset [0, 1]^d$, 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 $\mathcal{L}(h^*) = \mathbb{E}_x\{\eta(x) \wedge (1 - \eta(x))\} \geq \mathbb{E}_x(\eta(x)(1 - \eta(x)))$. □

- **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 $\leq \sum_{i=1}^r P(C_i)e^{-nP(C_i)} \leq r \max P(C_i)e^{-nP(C_i)} \leq \frac{r}{ne}$.□

- **Theorem** 1.1 (Generalization upper bound for 1-NN) $\mathbb{E}_S\mathcal{L}(\hat{h}_S) \leq 2\mathcal{L}(h^*) + 2c\sqrt{d}n^{-\frac{1}{d+1}}$.

  Proof Take C_i of the form $\{x : x_j \in [(\alpha_j - 1)/T, \alpha_j/T], \forall j\}$, where $\alpha_1, \cdots, \alpha_d \in \{1, \cdots, T\}^d$.
  Case 1: If $x, x' \in C_i$ for some i, then $||x - x'|| \leq \sqrt{d}\epsilon$.
  Case 2: Otherwise, $||x - x'|| \leq \sqrt{d}$.
  Hence, $\mathbb{E}_{S,x}||x - x_{\pi_1(x)}|| \leq \mathbb{E}_S\{P(\cup_{i:C_i\cap S\neq\emptyset}C_i)\sqrt{d}\epsilon + P(\cup_{i:C_i\cap S=\emptyset})\sqrt{d}\} \leq \sqrt{d}(\epsilon + \frac{r}{ne})$. Since $r = (\frac{1}{\epsilon})^d$, $\cdots \leq \sqrt{d}(\epsilon + \frac{1}{\epsilon^d ne})$. Matching the two terms gives $\epsilon = (\frac{1}{ne})^{\frac{1}{d+1}}$ and the optimal bound $2\sqrt{d}(ne)^{-\frac{1}{d+1}} \leq 2\sqrt{d}n^{-\frac{1}{d+1}}$. □

- **Theorem** 1.2 (Generalization upper bound for k-NN) $\mathbb{E}_S\mathcal{L}(\hat{h}_S) \le (1+\sqrt{\frac{8}{k}})\mathcal{L}(h^*) + (6c\sqrt{d}+k)n^{-\frac{1}{d+1}}$.

  Remark 1.1 k is called regularization parameter/hyperparameter and the optimal $k \sim n^d$.

  Remark 1.2 Exponential dependence on d: "curse of dimensionality".

- **Theorem** 1.3 (Lower bound) $\forall c > 1$ and any learning rule h, $\exists$ a distribution over $[0,1]^d \times \{0,1\}$ s.t. $\eta(x)$ is c-Lipschitz, the Bayes error is 0, but for $n < (c+1)^d/2$, $\mathbb{E}\mathcal{L}(h) > \frac{1}{4}$ (i.e. minimax bound $\inf_h \sup_y \mathbb{E}\mathcal{L}(h) \ge Cn^{-\frac{1}{d+1}}$).

  Hint Let G_c^d be the regular grid on $[0,1]^d$ with distance $1/c$ between points. Then any $\eta : G_c^d \to \{0,1\}$ is c-Lipschitz. Then use the following theorem. □

- **Theorem** 1.4 (No free-lunch theorem) Let A be any learning rule for binary classification with 0-1 loss over $\mathcal{X}^d$ and $n < |\mathcal{X}|/2$. Then $\exists$ distribution D over $\mathcal{X} \times \{0,1\}$ s.t. $\mathbb{E}\mathcal{L}(A) \ge \frac{1}{4}$. Furthermore, with prob $\ge \frac{1}{7}$, $\mathcal{L}(A_S) \ge \frac{1}{8}$.

2 Linear Regression

- $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$, $\mathbb{E}(\epsilon|X) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and X fixed.

- Least squares estimator (LSE) solves the normal equation $X^TX\hat{\beta} = X^TY, \hat{\beta} = (X^TX)^-X^TY$.

- ANOVA: $y_{ij} = \mu + \alpha_j + \epsilon_{ij}, i = 1, \cdots, n_j, j = 1, \cdots, J$. $\sum_j n_j = n, \sum_j \alpha_j = 0$.

- **Definition** 2.1 θ is estimable if $\exists$ an unbiased estimator of θ. $c^T\beta$ is linearly estimable if $\exists l \in \mathbb{R}^n$ s.t. $\mathbb{E}(l^TY) = c^T\beta, \forall\beta \in \mathbb{R}^p \Leftrightarrow c = X^Tl \in \mathcal{C}(X^T)$.

- **Theorem** 2.1 (1) If $c^T\hat{\beta}$ is unique, then $c \in \mathcal{C}(X^TX) = \mathcal{C}(X^T)$.
  (2) If $c \in \mathcal{C}(X^T)$, then $c^T\hat{\beta}$ is unique and unbiased for $c^T\beta$.
  (3) If $c^T\beta$ is estimable and $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, then $c \in \mathcal{C}(X^T)$.

  Proof (1) Let $b \in \mathcal{C}(X^TX)^\perp$ be arbitrary, then $X^TY = X^TX\hat{\beta} = X^TX(\hat{\beta}+b) \Rightarrow c^T\hat{\beta} = c^T(\hat{\beta}+b) \Rightarrow c^Tb = 0$.
  (2) $c = X^Tl$ for some $l \in \mathbb{R}^n$, then $c^T\hat{\beta} = lX^T\hat{\beta} = lX^T(X^TX)^-X^TY = lP_XY$ is unique. $\mathbb{E}(c^T\hat{\beta}) = l^TP_x\mathbb{E}Y = l^TP_XX\beta = l^TX\beta = c^T\beta$.
  (3) If $\exists$ an estimator $T(X,Y)$ unbiased for $c^T\beta$, then $c^T\beta = \int T(X,y)\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\{-\frac{1}{2\sigma^2}||y-X\beta||^2\}dy$. Differentiate with β, $c = X^T\int \frac{y-X\beta}{(2\pi\sigma^2)^{\frac{n}{2}}\sigma^2}T(X,y)\exp\{-\frac{1}{2\sigma^2}||y-X\beta||^2\}dy$. □