Modern Multivariate Statistical Techniques

Jinhong Du,15338039

January 7, 2019

Content

1. 证明贝叶斯分类器是最优的

Proof.

(a) $(X_1,Y_1),\ (X_2,Y_2),\ \cdots,\ (X_n,Y_n) \sim F(X,Y)$;

(b) 简化成传统的假设检验问题:

对二分类问题（bianry/categorical）

$$H_0 : Y = 0 \qquad H_a : Y = 1$$

(c) 分类器

$$\delta(x) = \begin{cases} 0 & ,R \\ 1 & ,R^C \end{cases}$$

type I error $\alpha = \mathbb{P}\{\delta(\mathbf{X}) = 1|Y = 0\}$, type II error $\beta = \mathbb{P}\{\delta(\mathbf{X}) = 0|Y = 1\} = 1 - Power$, and the misclassification rate

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\{\delta(\mathbf{X})\neq\mathbf{Y}\}} = \mathbb{P}\{\delta(\mathbf{X}) \neq \mathbf{Y}\} &= \mathbb{P}\{\delta(\mathbf{X}) = 1|Y = 0\}\mathbb{P}\{\mathbf{Y} = 0\} + \mathbb{P}\{\delta(\mathbf{X}) = 0|Y = 1\}\mathbb{P}\{\mathbf{Y} = 1\} \\ &= \alpha\pi_0 + \beta\pi_1\end{aligned}$$

似然比准则

$$\delta_{L_c}(x) = \begin{cases} 0 & ,\dfrac{f_0}{f_1} > c \\ 1 & ,\dfrac{f_0}{f_1} \leq c \end{cases}$$

即 $R_c = \left\{x|\frac{f_0}{f_1} > c\right\}$, $R_c^C = \left\{x|\frac{f_0}{f_1} \leq c\right\}$

(d) Neyman-Pearson 引理: 在 $\alpha_c = \mathbb{P}\{\delta_{L_c}(x) = 1|Y = 0\}$ 的显著性水平下, $\delta(x)$ 是最大功效检验。即 $\forall\ \delta(x)$ 满足 $\mathbb{P}\{\delta(x) = 1|Y = 0\} \leq \alpha_c = \mathbb{P}\{\delta_{L_c}(x) = 1|Y = 0\}$, 其功效 $\mathbb{P}\{\delta(x) = 1|Y = 1\} \leq \mathbb{P}\{\delta_{L_c}(x) = 1|Y = 1\}$.
$\forall\ \delta(x)$, $\exists\ c \in \mathbb{R}$, s.t. $\mathbb{P}\{\delta(x) = 1|Y = 0\} = \alpha_c$,

$$\begin{aligned} 1 - \beta_c = \int_{R_c^C} f_1 \mathrm{d}x &\geq \int_{R^C} f_1 \mathrm{d}x \\ \beta_c = \int_{R_c} f_1 \mathrm{d}x &\leq \int_{R} f_1 \mathrm{d}x \\ \mathbb{E}\mathbb{1}_{\{\delta_{L_c}(\mathbf{X})\neq\mathbf{Y}\}} &\leq \mathbb{E}\mathbb{1}_{\{\delta(\mathbf{X})\neq\mathbf{Y}\}} \end{aligned}$$

Optimal decision rule is always a likelihood ratio test.

(e) 贝叶斯分类器

$$\delta_B(x) = \begin{cases} 0 & ,\dfrac{f_0}{f_1} > \dfrac{\pi_1}{\pi_0} \\ 1 & ,\dfrac{f_0}{f_1} \leq \dfrac{\pi_1}{\pi_0} \end{cases}$$

$$\mathbb{E}\mathbb{1}_{\{\delta_{L_c}(\mathbf{X})\neq\mathbf{Y}\}} = \pi_0 \int_{R_c^C} f_0 \mathrm{d}x + \pi_1 \int_{R_c} f_1 \mathrm{d}x$$

$$
\begin{aligned}
&= \int_{R_c^C} f_0\pi_0 \mathrm{d}x + \int_{R_c} f_1\pi_1 \mathrm{d}x \\
&= 1 - \int_{R_c} f_0\pi_0 \mathrm{d}x - \int_{R_c^C} f_1\pi_1 \mathrm{d}x \\
&= 1 - \int_{R_c \setminus R_{\frac{\pi_1}{\pi_0}}} f_0\pi_0 \mathrm{d}x - \int_{R_c^C \setminus R^C_{\frac{\pi_1}{\pi_0}}} f_1\pi_1 \mathrm{d}x - \int_{R_{\frac{\pi_1}{\pi_0}} \setminus R_c} f_0\pi_0 \mathrm{d}x - \int_{R^C_{\frac{\pi_1}{\pi_0}} \setminus R_c^C} f_1\pi_1 \mathrm{d}x \\
&\leq 1 - \int_{R_c \setminus R_{\frac{\pi_1}{\pi_0}}} f_1\pi_1 \mathrm{d}x - \int_{R_c^C \setminus R^C_{\frac{\pi_1}{\pi_0}}} f_0\pi_0 \mathrm{d}x - \int_{R_{\frac{\pi_1}{\pi_0}} \setminus R_c} f_0\pi_0 \mathrm{d}x - \int_{R^C_{\frac{\pi_1}{\pi_0}} \setminus R_c^C} f_1\pi_1 \mathrm{d}x \\
&= 1 - \int_{R_{\frac{\pi_1}{\pi_0}}} f_0\pi_0 \mathrm{d}x - \int_{R^C_{\frac{\pi_1}{\pi_0}}} f_1\pi_1 \mathrm{d}x \\
&= \mathbb{E}\mathbb{1}_{\{\delta_B(\mathbf{X}) \neq \mathbf{Y}\}}
\end{aligned}
$$

The inequality holds simply because when $x \in R_c \setminus R_{\frac{\pi_1}{\pi_0}} \subset R^C_{\frac{\pi_1}{\pi_0}}$, $f_0\pi_0 \leq f_1\pi_1$ and $x \in R_c^C \setminus R^C_{\frac{\pi_1}{\pi_0}} \subset R_{\frac{\pi_1}{\pi_0}}$, $f_0\pi_0 > f_1\pi_1$. And the following equality holds since $R_c \setminus R_{\frac{\pi_1}{\pi_0}} = R_c \cap R^C_{\frac{\pi_1}{\pi_0}}$, $R_c^C \setminus R^C_{\frac{\pi_1}{\pi_0}} = R_c^C \cap R_{\frac{\pi_1}{\pi_0}}$, $R_{\frac{\pi_1}{\pi_0}} \setminus R_c = R_{\frac{\pi_1}{\pi_0}} \cap R_c^C$, $R^C_{\frac{\pi_1}{\pi_0}} \setminus R_c^C = R^C_{\frac{\pi_1}{\pi_0}} \cap R_c$,

□

2. 用 PDEC 框架来分解 AdaBoost

   (a) 重编码，标签改为 ±1；
   (b) 寻找分类器集，C；
   (c) 选取初始权重，PDE；
   (d) 训练分类器，PDE；
   (e) 迭代，PDE；
   (f) 加权和，C。