基础数理统计

第二十二章 分类

1. 22.1 引言
2. 22.2 错误率与贝叶斯分类器
3. 22.3 高斯分类器与线性分类器

22.1 引言

分类

定义 1 (分类)

从一个随机变量 X 来预测另一个离散的随机变量 Y 的问题被称作是分类，或有指导的学习，或判别，或者称为模式识别。

具体来说，考虑 i.i.d 数据 $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$，其中 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^\top \in \mathcal{X} \subset \mathbb{R}^d$ 为一个 d 维向量且 Y_i 在某个有限集 $\mathcal{Y}$ 中取值。一个分类规则就是一个函数 $h : \mathcal{X} \to \mathcal{Y}$。当观测到一个新的 $\boldsymbol{X}$，预测 Y 为 $h(\boldsymbol{X})$。

理论分类器

我们首先从理论上研究分类问题：

分类问题

已知两个总体分布 $F(\boldsymbol{x})$ 和 $G(\boldsymbol{x})$, 对一个新的观察值 $\boldsymbol{X} = \boldsymbol{x}$ 如何分类?

- 例如 $N(0,1)$ 与 $Exp(1)$;
- 例如 $N(0,1)$ 与 $U[-1,1]$;
- $\cdots$

思考 1

对于两个正态总体 $N(0,1)$ 和 $N(1,1)$，直观上应该如何分类?

思考 2

对于两个正态总体 $N(0,1)$ 和 $N(1,2)$，应该如何分类?

22.2 错误率与贝叶斯分类器

分类器

每一个分类准则或者分类方法是把整个样本空间 $\boldsymbol{x} \in \mathbb{R}^p$ 分成两部分 $\mathcal{A}$ 和 $\mathcal{A}^c$.

- 当 $\boldsymbol{x} \in \mathcal{A}$ 时候，认为新的观察值来自于 $F(\boldsymbol{x})$.
- 当 $\boldsymbol{x} \in \mathcal{A}^c$ 时候，认为新的观察值来自于 $G(\boldsymbol{x})$

好的分类器即找出好的区域 $\mathcal{A}$.

错分率-Error Rates

给定一个分类方法，即给定区域 $\mathcal{A}$，

$$
\begin{aligned}
& P(\boldsymbol{X}\text{ 被错分}) \\
=& P(\boldsymbol{X} \in \mathcal{A}^c | \boldsymbol{X} \sim F) P(\boldsymbol{X} \sim F) + P(\boldsymbol{X} \in \mathcal{A} | \boldsymbol{X} \sim G) P(\boldsymbol{X} \sim G) \\
=& \pi_1 \int_{\mathcal{A}^c} dF(\boldsymbol{x}) + \pi_2 \int_{\mathcal{A}} dG(\boldsymbol{x})
\end{aligned}
$$

其中 π_1, π_2 为先验概率。

错分损失

更一般的可以延伸为带有错分成本的损失函数：

$$\begin{aligned} L(\mathcal{A}) = & P(X \in \mathcal{A}^c | X \sim F) P(X \sim F) * \text{loss}_1 \\ & + P(X \in \mathcal{A} | X \sim G) P(X \sim G) * \text{loss}_2 . \end{aligned}$$

这里 loss_1 是把来自 F 的样本误判为 G 的损失，loss_2 是把来自 G 的样本误判为 F 的损失。形式上，

$$L(\mathcal{A}) = \pi_1 \int_{\mathcal{A}^c} dF(x) + \pi_2 \int_{\mathcal{A}} dG(x),$$

其中 π_1, π_2 为先验概率乘以对应损失。

最优分类器-Bayes Classifier

最优的分类方法 (Bayes 分类器) 为:

$$\begin{aligned}\mathcal{A}^* &= \arg\min_{\mathcal{A}} \pi_1 \int_{\mathcal{A}^c} dF(x) + \pi_2 \int_{\mathcal{A}} dG(x) \\ &= \arg\min_{\mathcal{A}} \left\{\pi_1 P(\mathcal{A}^c) + \pi_2 Q(\mathcal{A})\right\},\end{aligned}$$

其中 P, Q 为分别为分布函数 F, G 对应的概率函数, 即

$$P(\mathcal{A}) = \int_{\mathcal{A}} dF(x), Q(\mathcal{A}) = \int_{\mathcal{A}} dG(x).$$

(1) $F(x)$ 和 $G(x)$ 有密度函数 $f(x)$ 和 $g(x)$,

$$\mathcal{A}^* = \{x : \pi_1 f(x) > \pi_2 g(x)\}.$$

(2) $F(x)$ 和 $G(x)$ 是离散分布:

$$\mathcal{A}^* = \{x_i : \pi_1 P(X = x_i) > \pi_2 Q(X = x_i)\}.$$

证明.

以 (1) 为例，记 $h(\mathcal{A}) = \pi_1 P(\mathcal{A}^c) + \pi_2 Q(\mathcal{A})$，则易证明

$$\begin{aligned} h(\mathcal{A}) - h(\mathcal{A}^*) &= \int_{\mathcal{A}^c \cap \mathcal{A}^*} [\pi_1 f(x) - \pi_2 g(x)] dx \\ &\quad - \int_{\mathcal{A}^{*c} \cap \mathcal{A}} [\pi_1 f(x) - \pi_2 g(x)] dx. \end{aligned}$$

$\square$

在 $\pi_1 = \pi_2 = 1/2$ 的条件下，计算

例 1

假定 F, G 分别是正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ 的分布函数，计算理论分类器及错分率。

$$\mathcal{A}^\star = \left\{x : \frac{1}{\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} > \frac{1}{\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}\right\}.$$

例 2

假定 F, G 分别是抛硬币和扔色子的分布函数，即

$$P(X=1) = P(X=2) = 1/2;$$
$$Q(X=1) = Q(X=2) = \cdots = Q(X=6) = 1/6,$$

计算理论分类器及错分率。

$$\mathcal{A}^\star = \{1, 2\}.$$

错分率是 $1/6$.

例 3

假定 F, G 分别是正态分布 $N(0,1)$ 和指数分布 $Exp(1)$ 的分布函数，计算理论分类器及错分率。

$$\mathcal{A}^{\star} = \{x : x < 0\}.$$

错分率是 $1/4$。

例 4

假定 F, G 分别是正态分布 $N(0,1)$ 和均匀分布 $U[-1,1]$ 的分布函数，计算理论分类器及错分率。

$$\mathcal{A}^{\star} = \{x : |x| > 1\}$$

错分率是 $1/2(\Phi(1) - \Phi(-1))$.

定义 2 (误差率)

一个分类器 h 的真实误差率为

$$L(h) = P\left(h(\boldsymbol{X}) \neq Y\right),$$

而经验误差率或训练误差率为

$$\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(\boldsymbol{X}_i) \neq Y_i).$$

考虑特殊情况，$\mathcal{Y}=\{0,1\}$，令

$$r(x)=\mathrm{E}(Y|X=x)=\frac{\pi f_1(x)}{\pi f_1(x)+(1-\pi)f_0(x)},$$

其中
$f_0(x)=f(x|Y=0)$, $f_1(x)=f(x|Y=1)$, $\pi=P(Y=1)$.

定义 3 (贝叶斯分类规则)

贝叶斯分类规则 $h^\star$ 为

$$h^\star(x)=\begin{cases}1 & r(x)>1/2\\ 0 & \text{其他}\end{cases}$$

集合 $\mathcal{D}(h)=\{x: P(Y=1|X=x)=P(Y=0|X=x)\}$ 称为决策边界。

贝叶斯分类规则可以写成一些等价形式

$$
\begin{aligned}
h^{\star}(x) &= \begin{cases} 1 & P(Y=1|X=x) > P(Y=0|X=x) \\ 0 & \text{其他} \end{cases} \\
&= \begin{cases} 1 & \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{其他} \end{cases}
\end{aligned}
$$

定理 1

贝叶斯规则是最优的，即若 h 是任何其他分类准则，则 $L(h^\star) \leq L(h)$.

实际中贝叶斯规则是基于数据估计得到的，

- 经验风险极小化：选择一组分类器集 $\mathcal{H}$ 并且找到 $\hat{h} \in \mathcal{H}$ 使得能够极小化 $L(h)$ 的某个估计；
- 回归：找到回归函数 r 的一个估计 $\hat{r}$ 并定义

$$\hat{h}(x) = \begin{cases} 1 & \hat{r}(x) > \dfrac{1}{2} \\ 0 & \text{其他} \end{cases}$$

- 密度估计：估计 f_0, f_1 并且令 $\hat{\pi} = n^{-1}\sum_{i=1}^{n} Y_i$。定义

$$\hat{r}(x) = \hat{\mathbb{P}}(Y=1|X=x) = \frac{\hat{\pi}\hat{f}_1(x)}{\hat{\pi}\hat{f}_1(x) + (1-\hat{\pi})\hat{f}_0(x)},$$

$$\hat{h}(x) = \begin{cases} 1 & \hat{r}(x) > \dfrac{1}{2} \\ 0 & \text{其他} \end{cases}$$

定理 2

假设 $Y \in \mathcal{Y} = \{1, 2, \ldots, K\}$，最优规则为

$$h(x) = \operatorname*{argmax}_k P(Y = k|X = x) = \operatorname*{argmax}_k (\pi_k f_k(x)),$$

其中

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_r \pi_r f_r(x)},$$

$\pi_r = P(Y = r)$, $f_r(x) = f(x|Y = r)$。

22.3 高斯分类器与线性分类器

多元正态分布

假设 $\boldsymbol{x}|y=k$ 服从多元正态分布 $k=0,1$：

$$f_k(\boldsymbol{x})=(2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}.$$

其中 $\boldsymbol{\mu}_k$ 为 p 维均值，Σ_k 为 $p\times p$ 维的正定对称矩阵。

定理 3

若 $\boldsymbol{x}|y=0 \sim N_p(\boldsymbol{\mu}_0, \Sigma_0)$, $\boldsymbol{x}|y=1 \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$，则贝叶斯规则为

$$h^\star(x) = \begin{cases} 1 & r_1^2 < r_0^2 + 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & \text{其他} \end{cases}$$

其中 $r_i^2(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)$, $i=0,1$ 为 *Mahalanobis* 距离。等价于

$$h^\star(x) = \operatorname*{argmax}_k \delta_k(\boldsymbol{x}),$$

其中

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \log \pi_k.$$

QDA

整理得到 QDA(Quadratic Discriminant Analysis):

$$\boldsymbol{x}^\top(\Sigma_0^{-1} - \Sigma_1^{-1})\boldsymbol{x} - 2\boldsymbol{x}^\top(\Sigma_0^{-1}\boldsymbol{\mu}_0 - \Sigma_1^{-1}\boldsymbol{\mu}_1)$$
$$+ \boldsymbol{\mu}_0^\top\Sigma_0^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top\Sigma_1^{-1}\boldsymbol{\mu}_1 + 2\log\frac{\pi_1}{\pi_0} + \log\frac{|\Sigma_0|}{|\Sigma_1|} \le 0.$$

分类器具有二次型的形式:

$$\boldsymbol{x}^\top\Omega\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{\alpha} \le 0.$$

QDA 的参数估计

$$
\begin{aligned}
n_0 &= \sum_{i=1}^{n}(1-Y_i),\ n_1 = \sum_{i=1}^{n} Y_i, \\
\widehat{\pi}_0 &= \frac{1}{n}\sum_{i=1}^{n}(1-Y_i),\ \widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i, \\
\widehat{\boldsymbol{\mu}}_0 &= \frac{1}{n_0}\sum_{i:Y_i=0} \boldsymbol{x}_i,\ S_0 = \frac{1}{n_0}\sum_{i:Y_i=0}(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}}_0)(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}}_0)^\top, \\
\widehat{\boldsymbol{\mu}}_1 &= \frac{1}{n_1}\sum_{i:Y_i=1} \boldsymbol{x}_i,\ S_1 = \frac{1}{n_1}\sum_{i:Y_i=1}(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}}_1)^\top
\end{aligned}
$$

LDA

特别的，如果协方差结构相同 $\Sigma_0 = \Sigma_1 = \Sigma$, 我们得到 LDA(Linear Discriminant Analysis):

$$2\boldsymbol{x}^\top \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - \boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 - 2\log\frac{\pi_1}{\pi_0} \geq 0,$$

即

$$(\boldsymbol{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2})^\top \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) + \log\frac{\pi_0}{\pi_1} \geq 0,$$

LDA 的参数估计

$$S = \frac{n_0 S_0 + n_1 S_1}{n_0 + n_1},$$

分类规则等价为

$$h^\star(x) = \begin{cases} 1 & \delta_1(\boldsymbol{x}) > \delta_0(\boldsymbol{x}) \\ 0 & \text{其他} \end{cases}$$

其中

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^\top S^{-1} \widehat{\boldsymbol{\mu}}_k - \frac{1}{2} \widehat{\boldsymbol{\mu}}_j^\top S^{-1} \widehat{\boldsymbol{\mu}}_j + \log \widehat{\pi}_j.$$

作为判别函数。

Fisher 的线性判别分析

(1). 将数据投影到一条直线上，即用 $U=\omega^\top \boldsymbol{x}$ 代替 $\boldsymbol{x}$ 来进行分类；

(2). 不同类尽可能分开，Σ 为 $\boldsymbol{x}$ 的协方差阵，定义分离为

$$\begin{aligned} J(\omega) &= \frac{(\mathsf{E}(U|Y=0)-\mathsf{E}(U|Y=1))^2}{\omega^\top\Sigma\omega} \\ &= \frac{\omega^\top(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top\omega}{\omega^\top\Sigma\omega} \end{aligned}$$

(3) 实际中 $J(\omega)$ 用 $\widehat{J}(\omega)$ 进行估计

$$\widehat{J}(\omega)=\frac{\omega^\top S_B\omega}{\omega^\top S_W\omega},$$

其中

$$S_B=(\bar{\boldsymbol{x}}_0-\bar{\boldsymbol{x}}_1)(\bar{\boldsymbol{x}}_0-\bar{\boldsymbol{x}}_1)^\top,\ S_W=\frac{(n_0-1)S_0+(n_1-1)S_1}{(n_0-1)+(n_1-1)}.$$

定理 4

向量 $\omega = S_W^{-1}(\bar{\boldsymbol{x}}_0 - \bar{\boldsymbol{x}}_1)$ 为 $\widehat{J}(\omega)$ 的极小值点，称

$$U = \boldsymbol{\omega}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_0 - \bar{\boldsymbol{x}}_1)^\top S_W^{-1} \boldsymbol{x}$$

为 *Fisher* 线性判别函数。定义

$$m = \frac{1}{2}(\bar{\boldsymbol{x}}_0 - \bar{\boldsymbol{x}}_1)^\top S_W^{-1}(\bar{\boldsymbol{x}}_0 + \bar{\boldsymbol{x}}_1).$$

Fisher 分类规则为

$$h(x) = \begin{cases} 0 & \boldsymbol{\omega}^\top \boldsymbol{x} \geq m \\ 1 & \boldsymbol{\omega}^\top \boldsymbol{x} < m. \end{cases}$$

其他分类方法

- 逻辑回归;
- 支持向量机;
- 随机森林;
- 神经网络;
- 决策树;
- 最小近邻法;
- $\cdots$