

第一章 绪论.

$D = \{x_1, x_2, \dots, x_m\}$ 表示 m 个示列数据集. 每个属性由 d 个属性描述.

$x_i = (x_{i1}, x_{i2}, \dots, x_{id})$ 是 d 维样本空间 X 中向量. $x_i \in X$, 其中 x_{ij} 是 x_i 在属性上的取值. d 称为样本 x_i 的维数.

(x_i, y_i) 表示第 i 个样例. 其中 $y_i \in Y$ 是 x_i 的标记, Y 是所有标记的集合, 亦称“标记空间.”

训练: $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$. 建立一个从输入空间 X 到输出空间 Y 的映射 $f: X \rightarrow Y$. 对二分类任务, 通常令 $Y = \{-1, +1\}$ 或 $\{0, 1\}$; 对多分类任务, $|Y| > 2$; 对于回归任务, $Y = \mathbb{R}$. \mathbb{R} 为实数集.

预测标记: $y = f(x)$.

分布: D . 学习算法 E

假设样本空间 X 和假设空间 H 是离散的. 令 $P(h|X, \varepsilon_a)$ 代表算法 E_a 基于训练数据 X 产生假设 h 的概率. 再令 f 代表我们希望学习的真实目标函数.

E_a 在训练集之外的所有样本的误率为:

$$E_{te}(E_a|X, f) = \sum_h \sum_{x \in X - X} P(x) \mathbb{I}(h(x) \neq f(x)) P(h|X, \varepsilon_a).$$

$\mathbb{I}(\cdot)$ 是指示函数. 为真是 1, 否则为 0.

考虑二分类. $X \mapsto \{0, 1\}$, 函数空间为 $\{0, 1\}^{|X|}$, 对所有可能的 f 均均匀分布对误差求和.

$$\sum_f E_{te}(E_a|X, f) = \sum_f \sum_{h \in H} \sum_{x \in X - X} P(x) \mathbb{I}(h(x) \neq f(x)) P(h|X, \varepsilon_a).$$

$$= \sum_{x \in X - X} P(x) \sum_h P(h|X, \varepsilon_a) \sum_f \mathbb{I}(h(x) \neq f(x)).$$

我们对 f 的假设是能将任意样本映射到 $\{0, 1\}$ 的函数且服从均匀分布. 即不止一个且每个 f 概率相同. 当样本空间只有两个样本, $X = \{x_1, x_2\}$, $|X| = 2$. 则所有真实目标函数为: $f_1: f_1(x_1) = 0, f_1(x_2) = 0$;

$$f_2: f_2(x_1) = 0, f_2(x_2) = 1;$$

$$f_3: f_3(x_1) = 1, f_3(x_2) = 0;$$

$$f_4: f_4(x_1) = 1, f_4(x_4) = 1;$$

- 共 $2^{|X|} = 4$ 个真实函数. 无论 ε_a 预测正确的模型 $h(x)$ 无论预测值为 0 还是 1 只有一半与之预测值相等.

$$\therefore \sum_f I(h(x) \neq f(x)) = \frac{1}{2} 2^{|X|}$$

$$* = \sum_{x \in X - x} P(x) \sum_h P(h|x, \varepsilon_a) \stackrel{\downarrow}{=} 2^{|X|}$$

$$= \frac{1}{2} 2^{|X|} \cdot \sum_{x \in X - x} P(x) \cdot \sum_h P(h|x, \varepsilon_a). \quad \text{所有可能的 } h \text{ 的概率和为 1.}$$

$$= 2^{|X|-1} \sum_{x \in X - x} P(x).$$

因此, 在假设下, 任何算法 $\varepsilon_a, \varepsilon_b$ 期望相同. \Leftarrow 在共同前提下.

$$\text{习题 1.4. 证明 } E_{\text{ote}}(\varepsilon_a|x, f) = \sum_h \sum_{x \in X - x} P(x) \ell(h(x), f(x)) P(h|x, \varepsilon_a) \text{ 与 } \varepsilon_a \text{ 无关.}$$

$$\text{证: } \bar{\text{原式}} = \sum_{x \in X - x} P(x) \cdot \sum_h P(h|x, \varepsilon_a) \cdot \ell(h(x) \neq f(x))$$

$$\text{二分类. } \sum_f E_{\text{ote}}(\varepsilon_a|x, f) = \sum_{x \in X - x} P(x) \cdot \sum_h P(h|x, \varepsilon_a) \cdot \sum_f \ell(h(x) \neq f(x)).$$

由于 f 的分布, $\ell(h(x) \neq f(x)) + \ell(h(x) = f(x)) = A$.

证明: 由于二分类, 正例分类得分为 1, 错分类得分为 0.

$$\ell(0, 0) = \ell(1, 1) = a. \quad \ell(0, 1) = \ell(1, 0) = b. \quad \therefore \ell(0, 0) + \ell(0, 1) = A$$

$$\therefore \sum_f \ell(h(x), f(x)) = \frac{1}{2} \cdot 2^{|X|} \ell(h(x) \neq f(x)) + \frac{1}{2} 2^{|X|} \cdot \ell(h(x) = f(x))$$

$$\therefore \bar{\text{原式}} = 2^{|X|-1} A \sum_{x \in X - x} P(x).$$

第2章 模型评估与选择

错误率 $\epsilon = a/m$.

精度 accuracy = $1 - \epsilon/m$.

从训练集 S 中产生测试集 T :

1. 留出法. D 中分离出训练集 S 和测试集 T . $D = S \cup T$, $S \cap T = \emptyset$.

2. 交叉验证. D 划分为 k 互斥子集. $D = D_1 \cup D_2 \dots \cup D_k$, $D_i \cap D_j = \emptyset$.

k -fold. 每次 $k-1$ 个训练集, 剩下一个测试.

返回平均结果.

3. 自助法. 每次选一个样本放入 D' . m 次后得到同尺寸的 D' . 取样可能重复.
(数据不放). D 为训练集, D' 为测试集 (集合减法).

性能度量. 回归: 预测误差. $E(f; D) = \int_{\mathcal{X} \times \mathcal{Y}} (f(x) - y)^2 p(x) dx$.

分类: 错误率. $E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) \neq y_i)$.

精度 $\text{acc}(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) = y_i) = 1 - E(f; D)$.

查准率 P : $P = \frac{TP}{TP + FP}$. 预测为正. precision

查全率 R : $R = \frac{TP}{TP + FN}$. 真实为正. recall

$P-R$ 曲线. $P=R$. 平衡点 (BEP)

$F_1 = \frac{2PR}{P+R}$. 混合平均. 更重视较小值.

$F_\beta = \frac{(1+\beta^2)PR}{(\beta^2 P) + R}$ $\beta > 1$ 查全率更影响. $\beta < 1$ 查准率更影响.

$$\text{macro-}P = \frac{1}{n} \sum_{i=1}^n P_i$$

$$\text{macro-}R = \frac{1}{n} \sum_{i=1}^n R_i$$

$$\text{macro-}F_1 = \frac{2 \text{macro-}P \text{macro-}R}{\text{macro-}P + \text{macro-}R}$$

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$

$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}}$$

$$\text{micro-}F_1 = \frac{2 \text{micro-}P \text{micro-}R}{\text{micro-}P + \text{micro-}R}$$

ROC curve: 

AUC: $\frac{1}{2} \sum_{i=1}^{m^-} (x_{if} - x_i) (y_i + y_{i+1})$. 分类器的通过率降低.

i.e.: 沿长 x 轴 $\frac{1}{m^-}$, y 轴 $\frac{1}{m^+}$ \uparrow 即梯形面积公式.

$$\begin{aligned} \text{rank} &= \frac{1}{m^+ m^-} \sum_{x \in D^+} \sum_{x' \in D^-} (\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-))) \\ &= \sum_{x \in D^+} \frac{1}{2} \cdot \frac{1}{m^+} \left[\frac{2}{m^-} \sum_{x' \in D^-} \mathbb{I}(f(x^+) < f(x')) + \frac{1}{m^-} \sum_{x' \in D^-} \mathbb{I}(f(x^+) = f(x')) \right] \end{aligned}$$

证明: 设 Y 中预测值从大到小 $p_1 \sim p_n$. 则 $\forall p_i$ 有 $L_i \rightarrow$ TPR 为 s_i .

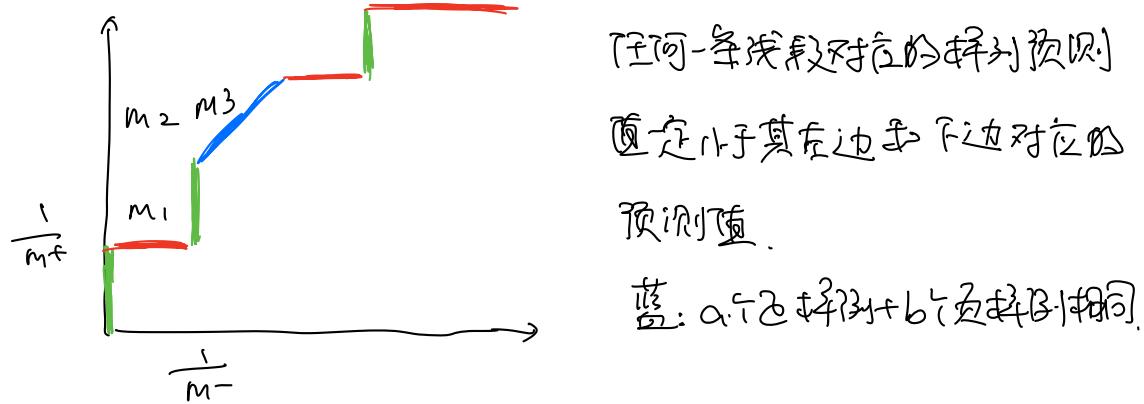
$$\sum_{i=1}^n s_i = 1 - \text{AUC}.$$

$f(x^+)$: 每一个正样本被认为是正确的.

$f(x^-)$: 每一个负样本被认为是正确的.

$f(x^+) < f(x^-)$ 代表正样本预测数值小于负样本.

因为是按预测数值从高到底操作的.



$\frac{1}{m^+} \sum_{x \in D^+} \mathbb{I}(f(x^+) < f(x^-))$ 为上式.

$\frac{1}{m^-} \sum_{x \in D^-} (\mathbb{I}(f(x^+) < f(x^-)) + \sum_{x' \in D^-} \mathbb{I}(f(x') = f(x^-)))$ 为下式.

代价敏感.

| 预 期 | 正 | 负 |
|--------|-----------|-----------|
| 0 | 0 | $cost_01$ |
| 1 | $cost_10$ | 0 |

$$E(f; D; cost) = \frac{1}{m} \left(\sum_{x \in D^+} \mathbb{I}(f(x) \neq y_i) \times cost_{01} + \sum_{x \in D^-} \mathbb{I}(f(x) \neq y_i) \times cost_{10} \right).$$

似然曲线下. $P(\text{f}) \approx \frac{P \times \text{cost}^{\alpha}}{P \times \text{cost}^{\alpha} + (1-p) \times \text{cost}^{\beta}}$ 从 $(0, PPR)$ 到 $(1, FNR)$ 线段.

似然比似然: $\text{costnorm} = \frac{FNR \times p \times \text{cost}^{\alpha} + FPR \times (1-p) \times \text{cost}^{\beta}}{p \times \text{cost}^{\alpha} + (1-p) \times \text{cost}^{\beta}}$

比较检验.

1. 单边检验.

Σ : 混化错误率. 学习器在一批样本上犯错概率.

$\hat{\Sigma}$: m 个测试样本有 $\hat{\Sigma}_{xm}$ 错误分类.

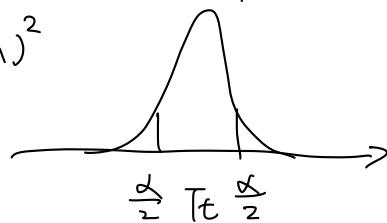
恰好得 $\hat{\Sigma}_{xm}$ 个样本错误概率: $P(\hat{\Sigma}, \varepsilon) = \binom{m}{\hat{\Sigma}_{xm}} \varepsilon^{\hat{\Sigma}_{xm}} (1-\varepsilon)^{m-\hat{\Sigma}_{xm}}$.

二项分布. $\varepsilon = \hat{\varepsilon}$ 最大.

$$\text{错误率 } \mu = \frac{1}{k} \sum_{i=1}^k \hat{\varepsilon}_i \quad k: k \text{ 个采样}$$

$$\text{方差 } \delta^2 = \frac{1}{k-1} \sum_{i=1}^k (\hat{\varepsilon}_i - \mu)^2$$

$$Tt = \frac{\sqrt{k}(\mu - \varepsilon_0)}{\delta}$$



若 $|\mu - \varepsilon_0|$ 在 $[t - \frac{t}{2}, t + \frac{t}{2}]$ 内, 则不能拒绝 $\mu = \varepsilon_0$. 混化错误率为 ε_0 . 置信度 $1-\alpha$.

否则可拒绝假设. 混化错误率 $> \varepsilon_0$.

2. 双边检验与 t 检验.

学习器 A. $\varepsilon_1^A, \varepsilon_2^A, \dots, \varepsilon_k^A$

学习器 B. $\varepsilon_1^B, \varepsilon_2^B, \dots, \varepsilon_k^B$

$\Delta_i = \varepsilon_i^A - \varepsilon_i^B \rightarrow \Delta_1, \dots, \Delta_k$ 对学习器 A, B
特征相同假设检验.

$\rightarrow Tt = \left| \frac{\bar{\Delta}_M}{\delta} \right| < t_{\frac{\alpha}{2}, k-1}$. 则无明显差异. 否则有.

e.g. 5x2 双边检验. $\mu = \frac{1}{2}(\Delta_1^1 + \Delta_1^2)$.

$$\delta_1^2 = (\Delta_1^1 - \frac{\Delta_1^1 + \Delta_1^2}{2})^2 + (\Delta_1^2 - \frac{\Delta_1^1 + \Delta_1^2}{2})^2.$$

$$Tt = \frac{\mu}{\sqrt{0.2 \sum_{i=1}^k \delta_i^2}} \quad \text{自由度 } S.$$

3. McNemar 检验.

| A | V | X |
|---|----------|----------|
| B | | |
| V | e_{00} | e_{01} |
| X | e_{10} | e_{11} |

假设两次学习结果相同, $e_{01}=e_{10}$, 则 $|e_{01}-e_{10}|$ 服从 χ^2 .

均值为1, 方差 $e_{01}+e_{10}$. 均值为1; 每个样本 $(-1, 0, 1)$.

$$\bar{F}x^2 = \frac{(e_{01}-e_{10}-1)^2}{e_{10}+e_{01}} > \chi^2 \text{ 有显著差异}$$

4. Friedman 检验与 Nemenyi 后续检验.

$$T_F = \frac{(N-1)Tx^2}{N(k-1)-Tx^2}, \quad T_F \text{ 服从自由度为 } k-1 \text{ 和 } (k-1)(N-1) \text{ 的 F 分布.}$$

$$Tx^2 = \frac{12N}{k(k+1)} \left(\sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4} \right), \quad N: N \text{ 个数据集, } k: k \text{ 种方法, } r_i: \text{平均强度.}$$

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$$

5. 偏差 - 方差.

测试样本 x , 数据集中标记 y_D , 真实标记 y , $f(x; D)$ 为训练集中模型在 x 上的输出.

学习算法期望预测 $\bar{f}(x) = E_D[f(x; D)]$.

方差 $Var(x) = E_D[(f(x; D) - \bar{f}(x))^2]$.

噪声 $\varepsilon^2 = E_D[(y_D - y)^2]$.

期望与真实差别为偏差 $bias(x) = (\bar{f}(x) - y)^2$.

期望泛化误差: $E(f; D) = bias^2(x) + var(x) + \varepsilon^2$.