模型评估与选择

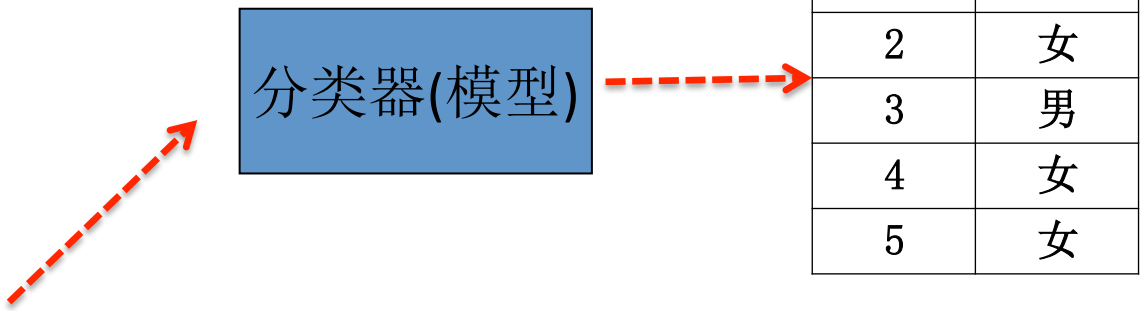
目录

经验误差与过拟合

真实值与预测值

样本ID	发型	喉结	胡须	性别
1	0.2	1	0.5	男
2	0.7	0	0.1	女
3	0.1	1	0.6	男
4	0.6	1	0.2	男
5	0.2	0	0.2	女

样本ID	发型	喉结	胡须
1	0.2	1	0.5
2	0.7	0	0.1
3	0.1	1	0.6
4	0.6	1	0.2
5	0.2	0	0.2



样本ID	性别
1	男
2	女
3	男
4	女
5	女

经验误差与过拟合

真实值与预测值

- 错误率：分类错误样本数占总样本数比例
- 精度：1 - 错误率
- 误差：模型输出与样本真实值之间的差异
- 训练误差 / 经验误差：模型在训练集上的误差
- 泛化误差：模型在新样本上的误差

样本ID	real	pre
1	男	男
2	女	女
3	男	男
4	男	女
5	女	女

经验误差与过拟合

真实值与预测值

- 目标：得到泛化误差小的模型 / 学习器
- 实际：新样本未知

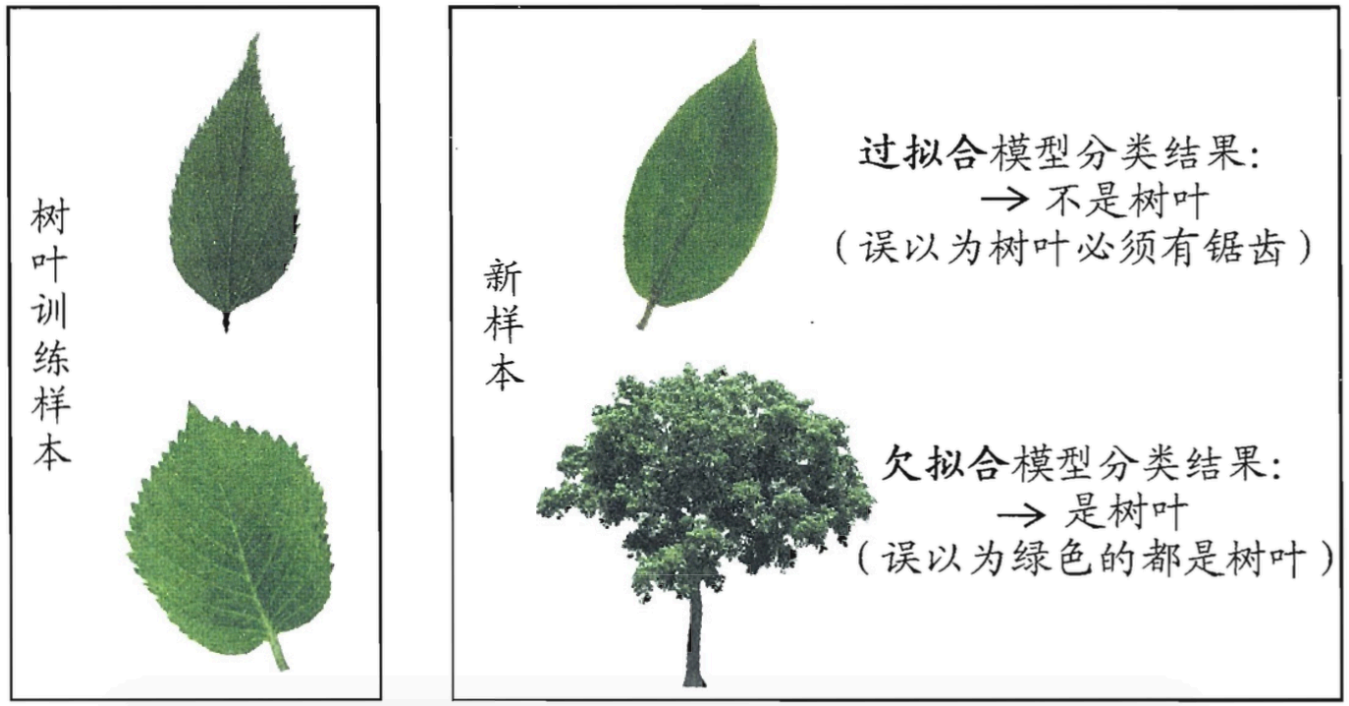
以经验误差代表泛化误差

- 模型从训练样本中学得适用于所有潜在样本的“普遍规律”

经验误差与过拟合

“过”与“不及”

- 过拟合：用力过猛
- 欠拟合：用力不足



目录

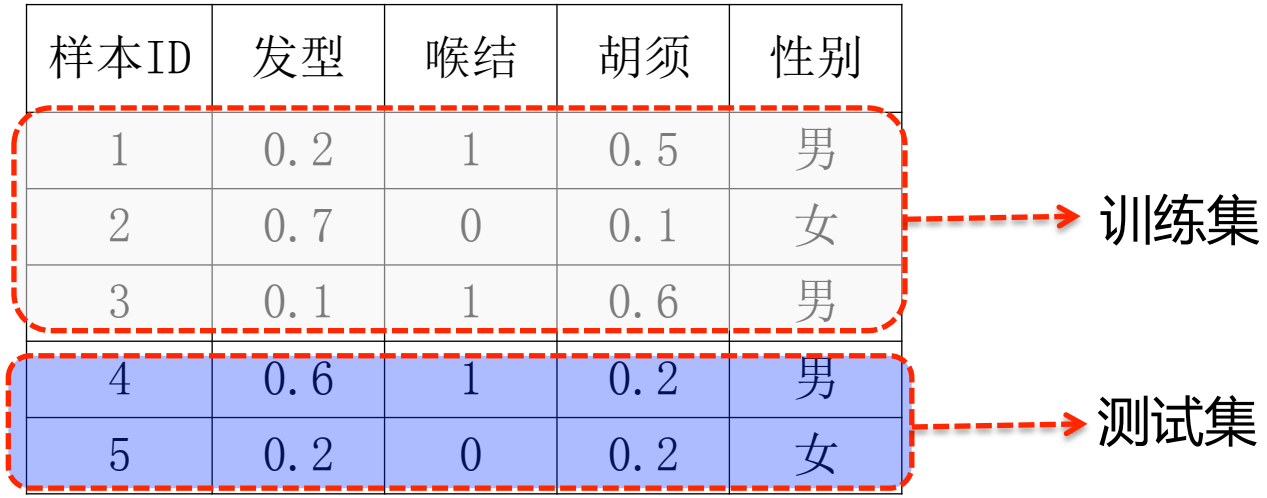
评估方法

训练集与测试集

目标：

- 对于模型 / 学习器的泛化误差进行评估
- 专家样本：训练集 + 测试集
- 训练集：训练误差
- 测试集：测试误差
- 独立同分布&互斥
- 用测试误差近似表示泛化误差

样本ID	发型	喉结	胡须	性别
1	0.2	1	0.5	男
2	0.7	0	0.1	女
3	0.1	1	0.6	男
4	0.6	1	0.2	男
5	0.2	0	0.2	女



评估方法

测试误差与泛化误差

- 留出法
- 交叉验证
- 自助法

评估方法

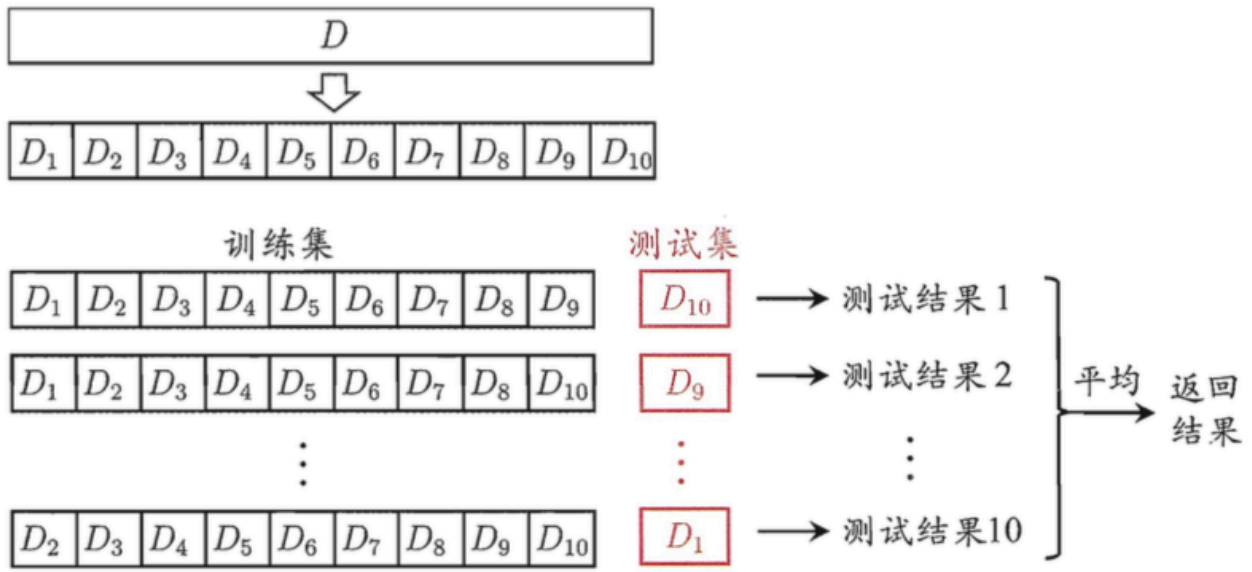
留出法

- 训练集 + 测试集：互斥互补
- 训练集训练模型，测试集测试模型
- 合理划分、保持比例
- 单次留出与多次留出
- 多次留出法：如对专家样本随机进行100次训练集 / 测试集划分，评估结果取平均

评估方法

交叉验证法

- K折交叉验证：将专家样本等份划分为K个数据集，轮流用K - 1个用于训练，1个用于测试
- P次K折交叉验证



评估方法

自助法

留出法与交叉验证法的训练集数据少于样本数据

- 给定m个样本的数据集D，从D中有放回随机取m次数据，形成训练集D’
- 用D中不包含D’ 的样本作为测试集
- D中某个样本不被抽到的概率： $\left(1-\frac{1}{m}\right)^m$
- 测试集数据量： $\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368$
- 缺点：改变了初始数据集的分布

目录

性能度量

评价方法与评价标准

- 回归任务的评价标准：均方误差

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2$$

性能度量

错误率与精度

- 错误率：分类错误样本数占总样本数比例

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

- 精度：1 - 错误率，分类正确样本数占总样本数比例

$$\begin{aligned}\operatorname{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

数据挖掘流程

查准率与查全率

- 查准率 / 准确率（precision）： F = TP/(TP+FP)
- 查全率 / 召回率 / 灵敏度（recall）：F = TP/(TP+FN)
- “1”代表正例，“0”代表反例

样本ID	real	pre
1	男	男
2	女	女
3	男	男
4	男	女
5	女	女

		预测值	
		1	0
实际值	1	TP	FN
	0	FP	TN

		预测值	
		1	0
实际值	1	2	1
	0	1	2

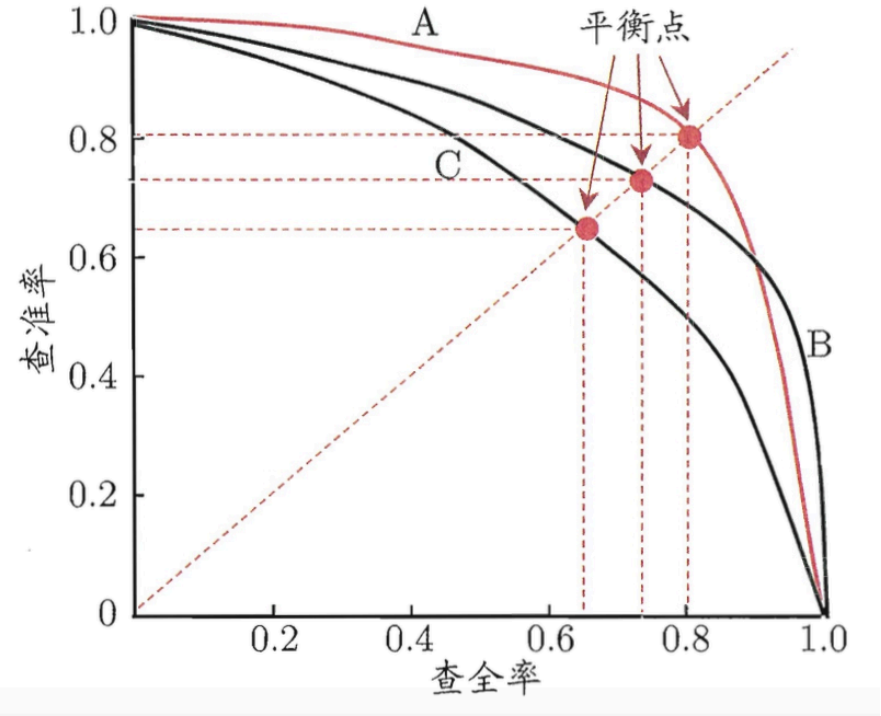
性能度量

查准率与查全率

- 查准率与查全率相互制约，且不同场景有不同要求

P-R曲线图

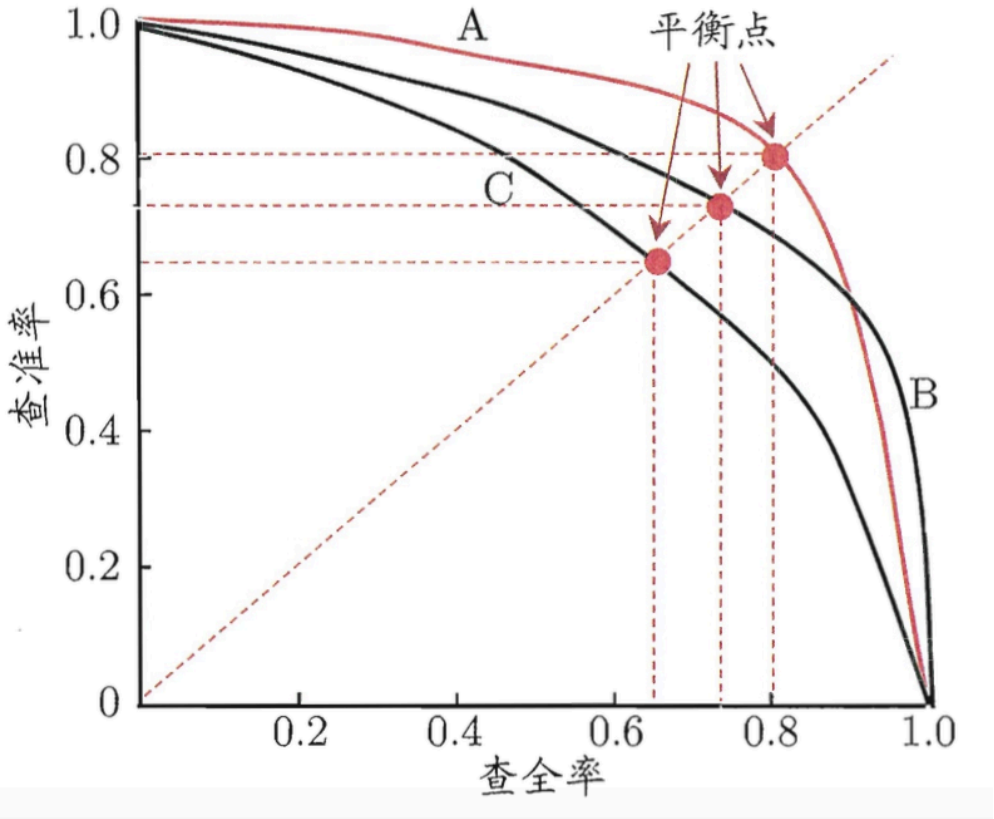
- 以二分类为例，模型输出各样本为正例的可能性列表，按可能性大小顺序逐个把样本预测为正例，则每次可计算出当前查全率与查准率，最后连线成图。
- 注：P-R图一般为非光滑非单调曲线
- 曲线下面积与平衡点（BEP）



性能度量

P-R曲线图

样本ID	1	2	3	4	5
男pr	0.9	0.3	0.7	0.6	0.3
pre1	男	女	女	女	女
pre2	男	女	男	女	女
pre3	男	女	男	男	女
pre4	男	男	男	男	女
pre5	男	男	男	男	男



性能度量

F1系数

- 综合查准率与查全率：

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{样例总数 + TP - TN}$$

- 更一般的形式：

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

- 其中 β 为正数，度量了查全率对查准率的相对重要性
- $\beta = 1$ ：标准的F1系数
- $\beta > 1$ ：查全率有更大影响
- $\beta < 1$ ：查准率有更大影响

性能度量

多次训练 / 测试时的F1系数

- 先分后总：先分别计算各混淆矩阵的查准率和查全率，再以均值汇总

$$\text{macro-}P = \frac{1}{n}\sum_{i=1}^{n} P_i \qquad \text{macro-}R = \frac{1}{n}\sum_{i=1}^{n} R_i \qquad \text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R}$$

- 先总后分：先将各混淆矩阵的对应元素（TP、FP、TN、FN）进行汇总平均，再求P、R、F1值

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \qquad \text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \qquad \text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R}$$

目录

比较检验与偏差方差

测试误差能代表泛化误差吗？

- 详见周志华：《机器学习》2.4比较检验

泛化错误率的构成：偏差 + 方差 + 噪声

- 偏差：模型输出与真实值的偏离程度，刻画了算法的拟合能力
- 方差：同样大小的训练集的变动导致的学习性能的变化，即数据扰动造成的影响
- 噪声：当前学习器所能达到的泛化误差的下限

- 偏差大：拟合不足 / 欠拟合；方差大：过拟合
- 详见周志华：《机器学习》2.5偏差与方差

Thank you!

泰迪科技：www.tipdm.com
热线电话：40068-40020