机器学习 - 周志华 - 模型评估与选择

经验误差与过拟合

> **定义：经验误差、泛化误差**
>
> 模型在训练集上的误差称为**经验误差**，或**训练误差**. 模型在新样本（测试集）上的误差称为**泛化误差**.

一般而言，误差的度量方式有很多种，分类任务一般考虑分类正确率，设 m 个样本有 a 个分类错误，则

$$E = 1 - \frac{a}{m}$$

为分类正确率. 回归任务的度量方法就更为多样了，有均方误差 MSE ，均方根误差 RMSE 等等. 距离（范数）的选择也会更多样.

- 欠拟合：在训练集上就具有较大的训练误差.
- 过拟合：训练误差较小，但泛化性能差（泛化误差）大.

如果认为 $\mathrm{P} = \mathrm{NP}$ ，那么对于一般的机器学习任务，都可以通过多项式复杂度的算法避免过拟合，但是现实来看似乎并不太可能.

评估法

留出法

> **方法：留出法**
>
> 留出法是一种训练集 D 划分为两个互斥的集合 S, T ：
>
> $$D = S \cup T, S \cap T = \varnothing$$
>
> 再在 S 上进行训练，T 作为测试集的方法.

这种方法主要用来评估模型的效用：如果我们只有一个数据集 D ，但是要通过验证来防止过拟合，那么这就是绝佳的方法.

编程实现考虑 `sklearn` 的 `train_test_split` 方法.

交叉验证法

这个方法和留出法类似，但是它是划分为 k 个不相交子集：

$$D = \bigsqcup_{i=1}^{k} D_i$$

其中 $\sqcup$ 表示不交并，然后每次取一个为测试集，剩余为训练集：

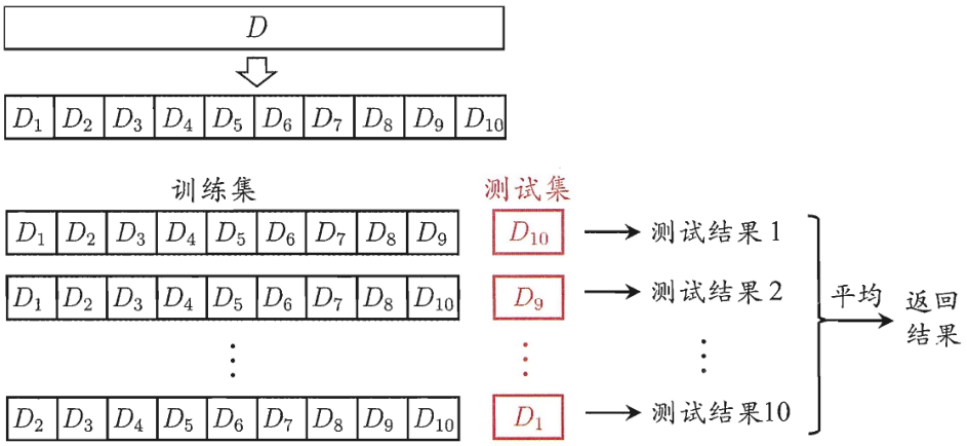

图 2.2 10 折交叉验证示意图

这种方法称为 k **折交叉验证法**. 如果 k 就等于样本数，就称**留一法**.

Bootstrap 法

Bootstrap 法是一种等概率采样方法，它的操作流程如下：（D 中共有 m 个样本）

- 首先建立一个新的空数据集 D' ，每次随机从 D 中挑选一个样本放入 D' .
- 重复 m 次操作，得到的 D' 和 D 一样大，只不过有的样本可能在 D' 中重复出现，有的没出现.
- 以 D' 为训练集，$D \setminus D'$ 为测试集.

一个样本从来没被取到的概率为 $(1-\frac{1}{m})^m$ ，我们可以在大样本情形下有

$$\left(1-\frac{1}{m}\right)^m \to \frac{1}{\mathrm{e}} \approx 0.368$$

Bootstrap 法是一个高效的方法，且因为其随机性，减小了样本划分造成的影响.

性能度量

错误率与精度

定义：错误率、精度

错误率定义为：

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

精度定义为：

$$\mathrm{acc}(f;D)=1-E(f;D)$$

准确率、召回率、$F1$ 度量

错误率和精度都是针对全体样本的分类结果的，假设一车西瓜已经判定完成，我们从中取出所有判定的好瓜，那么这里面真的是好瓜的比例是多少呢？这个比例称为**准确率**，和精度的差别在于它只关注其中我们判定为好瓜的这些内容.

我们不妨从理论上解析这里的内容，我们根据真实类别和判定类别建立一个矩阵，称为**混淆矩阵**：

真实情况/预测结果正		反
正	TP （真正例）	FN （假反例）
反	FP （假正例）	TN （真反例）

定义：准确率（查准率）

正例当中真的比例为**准确率**：

$$P=\frac{TP}{TP+FP}$$

此外，我们还可能关心有多少好瓜被我们找出来了，此时就有召回率.

> **定义：召回率（查全率）**
>
> 真正例当中被实际预测准确的比率为**召回率**：
>
> $$R = \frac{TP}{TP + FN}$$

准确率和召回率是矛盾的量，一般情况下我们会通过绘制 P-R 曲线来进行研究. 想要达到双高并不简单，取一个平衡点是我们常用的做法.

> **定义：平衡点 (Break-Event Poin,BEP)**
>
> 准确率和召回率相等的取值点称为**平衡点**.

但是很多时候，这种说法过于模糊，我们需要的是更为精确的度量，因此我们可以引入 $F1$ 度量：

> **定义：$F1$ 度量**
>
> 准确率和召回率的调和平均称为 $F1$ **度量**：
>
> $$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

如果对两者的偏好程度不同，我们可以对其进行加权.

> **定义：加权 $F1$ 度量**
>
> $$F_\beta = \frac{(1 + \beta^2) \cdot P \cdot R}{(\beta^2 \cdot P) + R}$$

其中的 β 即为权：即为召回率相对准确率的重要性. $\beta > 1$ 时召回率有更大影响，$\beta < 1$ 时准确率有更大影响.

ROC 和 AUC

我们这里将召回率用 TPR 表示，同时引入误判率（假正例率）：

$$\text{FPR} = \frac{FP}{TN + FP}$$

ROC 曲线实际上就是以 FPR 为横轴，TPR 为纵轴绘制的曲线，如果一个模型的 ROC 曲线比另一个模型的曲线更高，那么前者就是更好的模型.

但是很多时候高低不是绝对的，此时就需要比对曲线下方的面积，定义为 **AUC** (Area Under ROC Curve). 对于有限个样本点形成的曲线，可简单估算为：

$$\text{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)(y_i + y_{i+1})$$

代价敏感错误率和代价曲线

有时，分类错误的代价不一定会是均等的，我们在医院当中将一个病人的症状判轻或判重所造成的后果是不一样的，因此我们需要精确化出一个代价矩阵：

真实类别/预测类别	0	1
0	0	c_1
1	c_2	0

> **定义：代价敏感错误率**
>
> 代价敏感错误率定义为如下内容：
>
> $$E(f; D; c) = \frac{1}{m}\left(\sum_{\boldsymbol{x}_i \in D^+} \mathbb{1}_{(f(\boldsymbol{x}_i) \neq y_i)} c_1 + \sum_{\boldsymbol{x}_i \in D^-} \mathbb{1}_{(f(\boldsymbol{x}_i) \neq y_i)} c_2\right)$$

因此，我们仿照 ROC 曲线的想法，可以将代价加权到其中，得到代价曲线. 其横轴为：

定义：正例概率代价

$$P(+)c = \frac{p \cdot c_1}{p \cdot c_1 + (1-p) \cdot c_2}$$

其中 p 为样例为正例的概率，纵轴为 $[0,1]$ 的归一化代价：

定义：归一化代价

$$c_{\text{norm}} = \frac{\text{FPR} \cdot (1-p) \cdot c_2 + (1-\text{TPR}) \cdot p \cdot c_1}{p \cdot c_1 + (1-p) \cdot c_2}$$