模型评估与选择

2017 年 3 月 8 日

目录

1 经验误差与过拟合 **1**

2 评估方法 **2**

2.1 留出法 2

2.2 交叉验证法 (k 折交叉验证) 3

2.3 自助法 3

2.4 调参与最终模型 4

3 性能度量 **5**

3.1 错误率与精度 5

3.2 查准率、查全率和 F1 5

3.3 ROC 与 AUC 8

4 比较检验 **9**

5 偏差与方差 **9**

1 经验误差与过拟合

1. 概念

- 错误率：分类错误样本数占样本总数的比率

- 精度：1-错误率
- 误差：学习器的实际预测输出与样本的真实输出差异称为“误差”
- 泛化误差：学习器在训练集上的误差称为“训练误差”或“经验误差”，在新样本上的误差称为“泛化误差”

2. 过拟合与欠拟合

- 过拟合：最常见的情况是由于学习能力过于强大，以至于把训练样本所包含的不太一般的特性都学到了
- 欠拟合：通常由于学习能力低下造成

2 评估方法

2.1 留出法

直接将数据集 D 划分为两个互斥的集合，其中一个集合作为训练集 S, 另一个作为测试集 T，在 S 上训练出模型后，用 T 来估计其测试误差，作为对泛化误差的估计。

- 注意点：
  1. 训练/测试集划分要尽可能保持数据分布一致性。
  2. 即便给定训练集/测试集样本比例后，仍存在多种方式对初始数据集进行划分，不同划分方式得到的泛化误差也会有差异。
  3. 单次使用留出法将导致评估结果不够稳定可靠，在使用留出法时，一般要曹勇若干次随机划分，重复进行实验评估后取平均值作为留出法的评估结果。
  4. 留出法的训练集如果比较大，会比较接近用 D 训练出的模型，但是由于 T 比较小，评估结果可能不够准确稳定，如果测试集 T 比较大的话，训练集 S 与 D 的差别会更大，这个缺陷没有完美解决方案，常见做法是将大约 $2/3 \approx 4/5$ 的样本用于训练，剩余样本用于测试。

2.2 交叉验证法 (k 折交叉验证)

先将数据集 D 划分为 k 个大小相似的互斥子集，即 $D = D_1 \cup D_2 \cup \ldots \cup D_k, D_i \cap D_j = \emptyset (i \neq j)$. 每个子集 D_i 都尽可能保持数据分布一致性，即从 D 中通过分层采样得到。然后，每次用 $k-1$ 个子集的并集作为训练集，余下的那个子集作为测试集；这样，得到 k 组训练/测试集，从而可进行 k 次训练与测试，最终返回的是这 k 个测试结果的均值。

- 注意点：

  1. 与留出法类似，将数据集 D 划分为 k 个子集同样存在多种方式，为减小样本划分不同引入的差别，k 折验证法通常要随机使用不同的划分方式 p 次，最终这 p 次 k 折验证法结果的均值为最终的评估结果。
  2. 留一法：假定数据集 D 中包含 m 个样本，若令 k=m，则得到了交叉验证法的特例，留一法。
     – 留一法不收随机样本划分方式影响
     – 使用的训练集与初始数据集相比仅差了一个样本，使得训练结果与实际评估的模型很相似
     – 数据集比较大的情况下，计算量会很大

2.3 自助法

自助法直接以自助采样法为基础，给定 m 个样本的数据集 D，我们对其进行采样产生数据集 D' ：每次随机从 D 中挑选一个样本，将其拷贝放入 D', 然后再将该样本放回初始数据集 D 中，使得该样本在下次采样中仍有可能被采到；这个过程重复 m 次后，我们得到包含 m 个样本的数据集 D', 这就是自助采样的结果。

- 说明：

  1. 我们希望评估的是用 D 训练出的模型，但在留出法和交叉验证法中，由于保留了一部分样本用于测试，因此实际评估的模型所使

用的数据集比 D 小，这必然会引入因训练样本规模不同而导致的估计偏差，而留一法则在样本规模比较大的情况下，计算量太大。自助法在数据量比较小，难以有效划分训练/测试集时很有用。

2. 显然，D 中有一部分数据会在 D' 中多次出现，另一部分数据则不会出现，因此，可以做一个简单估计，样本在 m 次采样中始终不被采集到的概率是 $(1-\frac{1}{m})^m$, 取极限有：

$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368 \tag{1}$$

即通过自助采样，初始数据集 D 中约有 36.8% 的样本未出现在采样数据集 D' 中，于是我们可以将 D' 作为训练集，将 D' 作为测试集，这样，实际评估的模型与期望评估的模型都使用了 m 个训练样本，而我们仍有数据总量的 1/3 的、没在训练集中出现的样本用于测试，这样的测试结果，亦称“包外估计”(`out-of-bag estimate`)

3. 自助法能从初始数据集中产生多个不同的训练集，这对集成学习等方法有很大好处。

4. **然而，自助法产生的数据集改变了初始数据集的分布，这会引入估计偏差。** 因此，在初始数据量足够的情况下，留出法和交叉验证法更常用。

2.4 调参与最终模型

给定包含 m 个样本的数据集 D，在模型评估与选择过程中，由于需要留出一部分数据进行评估测试，事实上我们仅使用了一部分数据训练模型。因此，在模型选择完成后，学习算法和参数配置已经选定，此时应该用数据集 D 重新训练模型，这个模型在训练过程中使用了所有的 m 个样本，这才是我们最终提交给用户的模型。

3 性能度量

在预测任务中，给定样例集 $D=\{(x_1,y_1),(x_2,y_1),\ldots,(x_m,y_m)\}$, 其中 y_i 是标记示例 x_i 的真实标记。要评估学习器 f 的性能，就要把学习器预测结果 $f(x)$ 与真实标记 y 进行比较。

回归任务最常用的性能度量是“均方误差”(`mean square error`)：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2 \tag{2}$$

更一般的，对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$, 均方误差可描述为：

$$E(f;\mathcal{D})=\int_{x\sim\mathcal{D}}(f(x-y))^2p(x)\mathrm{d}x \tag{3}$$

3.1 错误率与精度

对样例集 D, 分类错误率定义为：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}I(f(x_i)\neq y_i) \tag{4}$$

精度定义为：

$$\begin{aligned} acc(f;D) &= \tfrac{1}{m}\sum_{i=1}^{m}I(f(x_i)=y_i) \\ &= 1-E(f;D) \end{aligned} \tag{5}$$

3.2 查准率、查全率和 F1

- 混淆矩阵：对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例 (TP), 真反例 (TN), 真反例 (TN), 假反例 (FN)，分类结果可以列出“混淆矩阵”

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

- 查准率：学习器学习出来的正例中正确的正例所占的比例

$$P = \frac{TP}{TP + FP} \tag{6}$$

- 查全率：学习器学习出来的正例占整个测试样本的比例

$$R = \frac{TP}{TP + FN} \tag{7}$$

- P-R 曲线：横坐标为查全率 (Recall)，纵坐标查准率 (Precision)

  在很多情形下，我们可根据学习器预测结果对样例进行排序，排在前面的是学习器认为“最可能”是正例的样本，排在最后的则是学习器认为“最不可能”是正例的样本，按此顺序逐个将样本作为正例进行预测，则每次都可以得到一组查全率，查准率。以查全率为横轴，查准率为纵轴，可以画出“P-R 曲线”，显示该曲线的图称为“P-R 图”。

- 平衡点 (BEP) 如果一个学习器 A 的“P-R”曲线将另外一个学习器 B 的“P-R 曲线”完全包住，可认为学习器 A 的性能比较好，如果 A 不能完成包住 B，两个学习器有交叉，可以度量两个学习器在“P-R 图”上围住的面积，但是计算并不方便，此时，可以取两个学习器的“平衡点”(BEP)，即查准率与查全率相等的点，看哪个值更大。

- F1: 基于查准率与查全率的调和平均

  BEP 还是过于简单，更常用的是 F1，基于查准率与查全率的调和平均

$$\begin{aligned} F1 &= \tfrac{1}{2}\tfrac{1}{1/P+1/R} \\ &= \tfrac{2PR}{P+R} \end{aligned} \tag{8}$$

- F_β: 基于查准率与查全率的调和平均

  F1 隐含了查准率与查全率重要性是一样的，为了表达出对查准率、查全率不同程度的偏好，可以引入加权后的调和平均

$$\begin{aligned} F_\beta &= \tfrac{1}{1+\beta^2}\tfrac{1}{1/P+\beta^2/R} \\ &= \tfrac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \end{aligned} \tag{9}$$

- 宏-查准率，宏-查全率，微-查准率，微-查全率

  很多时候，我们有很多混淆矩阵，我们希望在 n 个二分类混淆矩阵上综合考察查准率和查全率

  - 宏-查准率 (macro-P)，宏-查全率 (macro-R)

$$macro-P = \frac{1}{n}\sum_{i=1}^{n} P_i, \tag{10}$$

$$macro-R = \frac{1}{n}\sum_{i=1}^{n} R-i, \tag{11}$$

$$macro-F1 = \frac{2\times macro-P\times macro-R}{macro-P+macro-R}. \tag{12}$$

  - 微查准率 (micro-P)，微-查全率 (micro-R)

    可以将所有混淆矩阵对应元素进行平均，得到 TP, NP, TN, FN 的平均值，记为 $T\bar{P}, F\bar{P}, T\bar{N}, F\bar{N}$, 然后可以计算微-查准率和微-查全率。

$$micro-P = \frac{T\bar{P}}{T\bar{P}+F\bar{P}}, \tag{13}$$

$$micro-R = \frac{T\bar{P}}{T\bar{P}+F\bar{N}}, \tag{14}$$

$$micro-F1 = \frac{2\times micro-P\times micro-R}{micro-P+micro-R}. \tag{15}$$

3.3 ROC 与 AUC

- ROC 曲线

  我们根据学习器的预测结果对样例进行排序，按此顺序逐个将样本作为正例进行预测，每次计算出两个重要量的值，分别以它们为横轴、纵轴作图，就得到了“ROC 曲线”，其中，横坐标为“真正例率”，纵坐标为“假正例率”，定义如下：

$$\begin{aligned} TPR &= \tfrac{TP}{TP+FN}, \\ FPR &= \tfrac{FP}{TN+FP}. \end{aligned} \tag{16}$$

  显示“ROC 曲线”的图叫做“ROC 图”，“ROC 曲线”下的面积称为“AUC(Area Under ROC Curve)”。

- 不同 ROC 曲线的比较：与 P-R 曲线类似，若一个学习器的 ROC 被另一个学习器曲线完成包住，则可以断言，后者性能会更好，如果两者有交叉，则可以比较 AUC.

- **AUC 面积计算** (难点)

  可以通过对 ROC 曲线下各部分的面积求和而得。假定 ROC 曲线是由坐标 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ 的点按序连接而成 $(x_1 = 0, x_m = 1)$. 则 AUC 估算为：

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i) \cdot (y_i + y_{i+1}). \tag{17}$$

- 排序损失函数 (loss function)

$$_{rank} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(I(f(x^+) < f(x^-)) + \frac{1}{2}I(f(x^+) = f(x^-))\right) \tag{18}$$

4 比较检验

5 偏差与方差