机器学习常见知识汇总之模型评估

田晓彬
xiaobin9652@163.com

模型评估

1.过拟合与欠拟合

当模型把训练样本学的太好时，可能会把训练样本自身的一些特点当作了所有潜在样本共有的特点，这样会导致模型的泛化性能下降。即模型在训练集上表现很好，但是在测试集或者新数据上表现较差。这种现象被称为**过拟合**。

与过拟合相对应的是**欠拟合**，指训练样本的一般特点并未学习完全。即模型在训练和测试集上表现都不好。

下面介绍几种解决过拟合和欠拟合的方法。

降低过拟合风险

1. 获取更多的训练数据是解决过拟合最有效的手段，通常可以通过一定的规则来扩充数据集；
2. 降低模型的复杂度；
3. 正则化方法；
4. 引入集成学习。

降低欠拟合风险

1. 添加新特征；
2. 增加模型的复杂度；
3. 减少正则化系数。

2.评价方法

通常，我们需要对模型的泛化能力进行评估，但是我们往往仅有一个包含若干个数据集。这时我们可以通过对数据集划分，产生训练集和测试集。下面介绍几种常用的数据集划分方法。

留出法

留出法是最简单但也最直接的验证方法，它将原始的数据集划分成训练集和测试集两部分。留出法的缺点很明显，即在测试集上计算出来的评价结果和数据集的划分有很大的关系。为了消除这个缺点，可以引入交叉验证方法。

交叉验证

首先将全部样本划分成 k 个大小相等的样本子集；依次遍历这 k 个子集，每次把当前子集作为测试集，其余所有子集当作训练集进行模型的评估；最后将k次评估结果进行平均，作为模型最终的评价指标。

假定数据集中包含 D 个样本，若 $k=D$,可以得到交叉验证的一个特例，留一法。但当数据集样本数过多时，留一法会导致训练的计算开销过大。

自助法

当数据集规模较大时，交叉验证可以很好的对模型进行评估；但是当数据集的规模较小时，将数据集进行划分会进一步减少可训练样本的个数，这可能会影响最终模型的评价结果。而自助伐可以较好的解决这个问题。

对于样本总数为 n 的数据集，进行 n 次有放会的随机抽样，得到大小为 n 的训练集。在采样过程中，因为有放会，所以会导致有的样本会被重复取样，但是有的样本并没有没抽出过，将这些未被抽出的样本作为验证集进行模型评估。若样本数接近无穷大，未被采样的样本个数为 36.8% 。

3.评价指标

对模型的评估，需要衡量模型泛化能力的评价标准，即评价指标。不同的任务需要，通常对应着不同的模型评价指标，这意味着模型的好坏是相对的，需要取决于最终的任务需求。

3.1 准确率

准确率是分类任务中最常用的一种性能度量，可以适用于二分类和多分类任务。准确率又叫做精度，是被分类正确的样本数占样本总数的比例，即

$$Accuracy=\frac{1}{n}\sum_{i=1}^{n}I(f(\boldsymbol{x}_i)=y_i)$$

其中 n 为样本总数。

准确率虽然是分类问题中最简单也最直观的评价指标，但是却有着一个明确的缺陷。当不同类别的样本比例非常不平衡时，占比大的类别往往会成为决定准确率的主要因素。例如，当数据某一类样本占比 99%，即使分类器将所有样本都预测为这一类，还是有着 99%的高准确率。

3.2 精确率和召回率

对于二分类问题，我们通常可以根据其真实类别和模型预测的类别组合划分为真正例（TP）、假正例（FP）、真反例（TN）和假反例（FN）。很显然，$TP+FP+TN+FN=n$。分类结果的混淆矩阵如下表所示：

	预测结果为正	预测结果为负
真实标签为正	TP	FN

	预测结果为正	预测结果为负
真实标签为负	FP	TN

精确率又叫查准率，是指分类正确的正样本个数占模型预测为正样本的个数的比例；**召回率**又叫查全率，是指分类正确的正样本个数占真正的正样本个数的比例。通过混淆矩阵，我们可以很容易得出精确率和召回率的计算公式

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

$$Recall = \frac{TP}{TP + FN}$$

准确率和召回率是两个即矛盾有同意的指标，通常为了提高精确率，模型会在更有把握的时候才把样本预测为正样本，但这时会遗漏很多没有把握的正样本，导致召回率下降。为了全面的考虑模型的好坏，引入**P-R曲线**对模型进行评估。

在很多情况下，我们都可以对模型的预测结果进行排序，越靠前的样本被认为越可能是正样本。按照这个排序依次将样本进行划分，每次划分都可以得到一组精确率和召回率。以精确率作为纵轴，召回率作为横轴，就可以得到P-R曲线。通过P-R曲线的整体表现，可以更好的对模型进行全面的评估。若一个模型的P-R曲线被另一个模型完全包住，则可以断言后者的性能优于前者。当两个模型的曲线发生了交叉，可以引入**平衡点(BEP)**来度量，平衡点是精确率=召回率时的取值，这个值越大的模型的性能越好。

但是使用BEP也过于简单，所以更多的时候会使用$F1$度量：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

更一般的，我们使用 F_β 度量：

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall}$$

F_β 度量可以表示出对精确率和召回率的不同偏好。

3.3 ROC和AUC

ROC曲线的中文名称叫做“受试者工作特征曲线”。ROC曲线的横坐标为假阳性率（FPR）；纵坐标为真阳性率（TPR）。FPR和TPR的计算方法分别为：

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

$$TPR = \frac{TP}{TP + FN}$$

在一般的二值分类模型中，模型的输出一般都是预测样本为正例的概率。我们需要将这个概率转换成二值结果，这是就需要确定一个分类阈值。概率大于这个阈值的样本被认为是正例，概率小于这个阈值的样本被预测为负例。若动态的调整阈值，则每一个阈值都可以对应一个FPR和TPR，在ROC图上绘制出每个阈值对应的点，在将所有的点连接就得到最终的ROC曲线。

还有一种更直观的绘制ROC曲线的方法。首先根据样本标签统计出正负样本的数量，假设正样本数量为 P，负样本数量为 N。接下来把横轴的刻度间隔设置为 $1/N$，纵轴的刻度间隔设置为 $1/P$。在根据模型的输出概率将样本从高到低进行排序。依次遍历样本，同时从零点开始绘制ROC曲线，每遇到一个正样本就沿纵轴方向绘制一个刻度间隔的曲线；每遇到一个负样本就沿横轴方向绘制一个刻度间隔的曲线。

AUC指得是ROC曲线下的面积的大小，该值可以量化的反映基于ROC曲线确定的模型的性能。AUC越大，说明模型越把真正的正样本排在前面，模型的分类性能越好。下面介绍两种计算AUC值的方法。

1. 在有 P 个正样本，N个负样本的数据集中，一共有 $P * N$ 对样本，则

$$AUC = \frac{\sum I(P_P, P_N)}{P \times N}$$

其中

$$I(P_P, P_N) = \begin{cases} 1, & P_P > P_N \\ 0.5, & P_P = P_N \\ 0, & P_P < P_N \end{cases}$$

例如：

ID	label	pro
A	0	0.1
B	0	0.4
C	1	0.35
D	1	0.8

则 $AUC = \frac{1+1+1+0}{4} = 0.75$

2. 第二种方法首先将所有的样本按照概率大小进行排序，得到每个正样本的排序后序号后，利用下面的公式进行计算

$$AUC = \frac{\sum_{i \in P} rank_i - \frac{P \times (P+1)}{2}}{P \times N}$$

其中 $rank_i$ 代表第 i个正样本的排序后的序号。

如上表的样本所示，排序后的样本集为：

ID	label	pro	rank
A	0	0.1	1
C	1	0.35	2
B	0	0.4	3
D	1	0.8	4

按照这个排序结果，我们可以计算 $AUC = \frac{(4+2) - \frac{2 \times (2+1)}{2}}{2 \times 2} = 0.75$