一份非常全面的机器学习分类与回归算法的评估指标汇总

原创 我是王老湿 AI派 2018-10-09

> 本文是《机器学习宝典》第 3 篇，读完本文你能够掌握分类与回归算法的评估指标。

PS：文末附有练习题

读完**机器学习算法常识**之后，你已经知道了什么是欠拟合和过拟合、偏差和方差以及贝叶斯误差。在这篇给大家介绍一些机器学习中离线评估模型性能的一些指标。

当我们训练得到了多个模型之后，如何衡量这几个模型的性能呢？也就是说我们需要一个能够衡量模型“好坏”的标准，我们称之为评估指标。在对比不同的模型效果时，使用不同的评估指标往往会导致不同的结论，这也就是说模型的效果好坏是相对的。

针对不同类型的学习任务，我们有不同的评估指标，这里我们来介绍最常见的分类与回归算法的一些评估指标。

分类指标

生活中大多数的分类问题都属于二分类问题，所以这里以二分类为例，来说明下分类相关的一些指标。

正式介绍指标之前，先来普及一些基本概念：有时候“阳性”、“真”、“正类”、“1” 指的是一回事，“阴性”、“假”、“负类”、“0”指的也是一回事。例如模型对这个样本的预测结果为 1，可以认为模型对这个样本的预测结果为真、或者为正类、或者为阳性，其实说的都是一个意思。

混淆矩阵

混淆矩阵（confusion matrix）是一个评估分类问题常用的工具，对于 k 元分类，其实它就是一个k x k的表格，用来记录分类器的预测结果。对于常见的二分类，它的混淆矩阵是 2x2 的。

在二分类中，可以将样本根据其真实结果和模型的预测结果的组合划分为真阳性（true positive，TP）、真阴性（true negative，TN）、假阳性（false positive，FP）、假阴性（false negative，FN）。根据 TP、TN、FP、FN 即可得到二分类的混淆矩阵。

Confusion Matrix		Predict	
		正类	负类
Real	正类	TP(真阳性)	TN(真阴性)
	负类	FP(假阳性)	FN(假阴性)

准确度

准确率（accuracy）是指模型预测正确（包括预测为真正确和预测为假正确）的样本数量占总样本数量的比例，即

$$Accuracy = \frac{m_{correct}}{m_{total}}$$

其中，$m_{correct}$ 表示模型正确分类的样本个数，m_{total} 表示所有的样本个数。

在二分类中，准确率可以通过下面的计算公式得到。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

准确率是分类问题中的一个最简单也最直观的评估指标，但是准确率存在一些局限性。比如，在二分类中，当负样本占比 99 %时，如果模型把所有样本都预测为负样本也能获得 99% 的准确率。虽然准确率看起来很高，但是其实这个模型时没有用，因为它找不出一个正样本。

精确率

精确率（precision）是指模型预测为真，实际也为真的样本数量占模型预测所有为真的样本数量的比例，即

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

举例来说明，比如警察要抓小偷，抓了 10 个人，其中有 6 个人是小偷，那么精确率就是 6/10 = 0.6。

召回率

召回率（recall）有时候也叫查全率，是指模型预测为真，实际也为真的样本数量占实际所有为真的样本数量的比例，即

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

举例来说明，还是上面的警察抓小偷的例子，抓了 10 个人，其中 6 个人是小偷，还有另外 3 个小偷逃之夭夭，那么召回率就是 6 / (6 + 3) ≈ 0.67。

F1值/Fα值

一般来说，精确率和召回率是互斥的，也就是说精确率高的话，召回率会变低；召回率高的话，精确率会变低。所以设计了一个同时考虑精确率和召回率的指标 F1值。F1值是精确率和召回率的调和平均，即

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

在某些场景下，我们对精确率与召回率的关注程度不一样，这时候，F1值更一般的形式Fα值就能够满足。Fα值定义如下

$$F_\alpha = \frac{(1 + \alpha^2) \cdot Precision \cdot Recall}{\alpha^2 \cdot Precision + Recall}$$

其中，α 的大小表示召回率对精确率的相对重要程度。

多分类的情况

很多时候我们遇到的是多分类问题，这就意味着每两两类别的组合都对应一个二元的混淆矩阵。假设得到了 n 个二分类的混淆矩阵，那如何来平均这 n 个结果呢?

宏平均

第一种办法就是先在各个混淆矩阵中分别计算出结果，再计算平均值，这种方式称为“宏平均”。

$$Precision_{macro} = \frac{1}{n} \sum_{i=1}^{n} Precision_i$$

$$Recall_{macro} = \frac{1}{n} \sum_{i=1}^{n} Recall_i$$

$$F_{1-macro} = \frac{2}{\frac{1}{Precision_{macro}} + \frac{1}{Recall_{macro}}}$$

微平均

除了上面的宏平均之外，我们也可以将二元混淆矩阵的对应的元素进行平均，得到 TP、TN、FP、FN 的平均值，然后再根据这些平均值来计算，这种方式称为“微平均”。

$$Precision_{micro} = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}$$

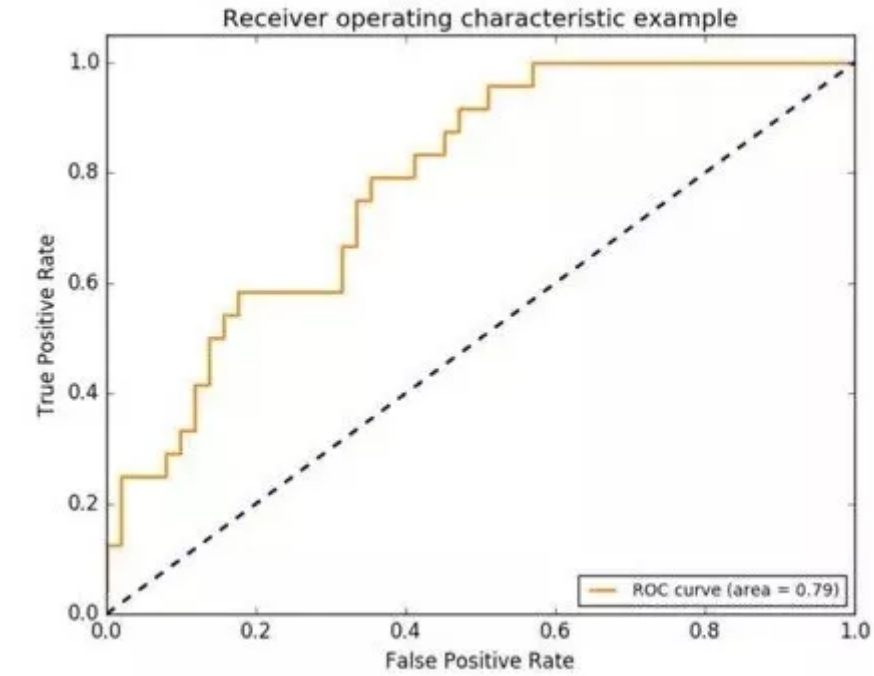
$$Recall_{micro} = \frac{\overline{TP}}{\overline{TP} + \overline{FN}}$$

$$F_{1-micro} = \frac{2}{\frac{1}{Precision_{micro}} + \frac{1}{Recall_{micro}}}$$

ROC

在前面介绍的这些指标中（如准确率、精确率、召回率等）都需要得到模型预测的结果（正类或负类），对很多模型来说，预测得到的是一个属于正类的概率值，所以就需要指定一个阈值，阈值以上的为正类，否则为负类。这个与之的大小直接决定了模型的泛化能力。

有一个评估指标叫受试者工作特征（Receiver Operating Characteristic，ROC）曲线，这种评估指标可以不用指定阈值。ROC曲线的纵轴是真阳率（TPR），横轴是假阳率（FPR）。



真阳率和假阳率的计算公式如下：

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

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

可以发现，TPR和Recall的计算公式是一样的。那么如何绘制ROC曲线呢？可以看到，ROC曲线是由一系列 (FPR, TPR)点构成的，但一个特定的模型，只得到一个分类结果，即只有一组 (FPR, TPR)，对应ROC曲线上的一个点，如何得到多个呢？

我们将模型对所有样本的预测值（属于正类的概率值）降序排列，然后依次将预测的概率值作为阈值，每次得到该阈值下模型预测结果为正类、负类的样本数，然后生成一组 (FPR, TPR) 值，这样就可以得到ROC曲线上的一点，最后将所有的点连接起来就出现了ROC曲线。很明显，阈值设置的次数越多，就会生成更多的 (FPR, TPR) 值，画出的ROC曲线也就越光滑。也就是说 **ROC曲线的光滑程度与阈值设置次数的多少有绝对的关系，与样本数量没有必然联系。**现实中，我们画出的 ROC 曲线多数都是不光滑的。

ROC曲线越靠近左上角，表示效果越好。左上角坐标为（0,1），即 FPR = 0，TPR = 1，这意味着 FP（假阳性）=0， FN（假阴性）=0，这就是一个完美的模型，因为能够对所有的样本正确分类。ROC曲线中的对角线（y=x）上的所有的点都表示模型的区分能力与随机猜测没有差别。

AUC

AUC（Area Under Curve）被定义为ROC曲线下的面积，很明显，AUC的结果不会超过 1，通常ROC曲线都在 y = x 这条直线上面，所以，AUC的值一般在 0.5 ~ 1 之间。

如何理解AUC的作用呢？随机挑选一个正样本（P）和负样本（N），模型对这两个样本进行预测得到每个样本属于正类的概率值，根据概率值对样本进行排序后，正样本排在负样本前面的概率就是AUC值。

AUC可以通过下面的公式计算得到。

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

其中，rank为将模型对样本预测后的概率值从小到大排序后的正样本的序号（排序从1开始），|P|为正样本数，|N|为负样本数。

需要注意的是，如果多个样本被模型预测的概率值一样，那么求rank的时候只需要将这些原始rank加起来求平均即可。所以说**相等概率得分的样本，无论正负，谁在前，谁在后无所谓。**

对数损失

对数损失（Logistic Loss，logloss）是对预测概率的似然估计，其标准形式为：

$$logloss = -logP(Y|X)$$

对数损失最小化本质是上利用样本中的已知分布，求解导致这种分布的最佳模型参数，使这种分布出现概率最大。

对数损失对应的二分类的计算公式为：

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y \cdot log\hat{y}_i + (1-y) \cdot log\hat{y}_i)$$

其中，N为样本数，$y \in \{0,1\}$，$\hat{y}_i$ 为第i个样本预测为1的概率。

对数损失在多分类问题中也可以使用，其计算公式为：

$$logloss = -\frac{1}{N}\cdot\frac{1}{C}\cdot\sum_{i=1}^{N}\sum_{j=1}^{C}y_{ij}\cdot log\hat{y}_{ij}$$

其中，N为样本数，C为类别数，y_{ij} 表示第i个样本的类别为j，$\hat{y}_{ij}$ 为第i个样本属于类别j的概率。

logloss衡量的是预测概率分布和真实概率分布的差异性，取值越小越好。

回归指标

在回归学习任务中，我们也有一些评估指标，一起来看看吧！

平均绝对误差

平均绝对误差（Mean Absolute Error，MAE）公式为：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

其中，N为样本数，y_i 为第i个样本的真实值，$\hat{y}_i$ 为第i个样本的预测值。

均方误差

均方误差（Mean Squared Error，MSE）公式为：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

平均绝对百分误差

平均绝对百分误差（Mean Absolute Percentage Error，MAPE）公式为：

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}|\frac{y_i - \hat{y}_i}{y_i}|, y_i \neq 0$$

MAPE通过计算绝对误差百分比来表示预测效果，其取值越小越好。如果MAPE=10，这表明预测平均偏离真实值10%。

由于MAPE计算与量纲无关，因此在特定场景下不同问题具有一定可比性。不过MAPE的缺点也比较明显，在 $y_i = 0$ 处无定义。另外需要注意的是，MAPE对负值误差的惩罚大于正值误差，比如预测一个酒店消费是200元，真实值是150元的会比真实值是250的MAPE大。

均方根误差

均方根误差（Root Mean Squared Error）的公式为：

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

RMSE代表的是预测值和真实值差值的样本标准差。和MAE相比，RMSE对大误差样本有更大的惩罚。不过RMSE有一个缺点就是对离群点敏感，这样会导致RMSE结果非常大。

基于RMSE也有一个常用的变种评估指标叫**均方根对数误差**（Root Mean Squared Logarithmic Error，RMSLE），其公式为：

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(y_i + 1) - log(\hat{y}_i + 1))^2}$$

RMSLE对预测值偏小的样本惩罚比预测值偏大的样本惩罚更大，比如一个酒店消费均价是200元，预测成150元的惩罚会比预测成250的大。

R2

R2（R-Square）的公式为：

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

R2用于度量因变量的变异中可由自变量解释部分所占的比例，一般取值范围是 0~1，R2越接近1,表明回归平方和占总平方和的比例越大,回归线与各观测点越接近,用x的变化来解释y值变差的部分就越多,回归的拟合程度就越好。

练习题

看完这篇文章，我们来做几道**练习题**来检验下学习成果：

1. 为什么说ROC曲线的光滑程度与样本数量没有绝对的关系呢？
2. 如果一个模型的AUC小于0.5，可能是因为什么原因造成的呢？
3. 在一个预测流量的场景中，尝试了多种回归模型，但是得到的 RMSE 指标都非常高，考虑下可能是因为什么原因造成的呢？
4. 在一个二分类问题中，15个样本的真实结果为[0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0]，模型的预测结果为[1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1]，计算准确率、精确率、召回率以及F1值。
5. 在一个二分类问题中，7个样本[A, B, C, D, E, F, G]的真实结果为[1, 1, 0, 0, 1, 1, 0]，模型的预测概率为[0.8, 0.7, 0.5, 0.5, 0.5, 0.5, 0.3]，计算AUC值。

以上所有的练习题答案我都会公布在我的知识星球中，方便后续做一个知识沉淀；另外，关于文章有任何疑问或者想要深入学习与交流，都可以加入我的知识星球来交流**（加入方式：扫描下方二维码或者点击“阅读原文”）。**

参考：

[1] 周志华.机器学习.第二章第三节（性能度量）
[2] 美团算法团队.美团机器学习实战.第一章第一节（评估指标）
[3] https://blog.csdn.net/qq_22238533/article/details/78666436
[4] https://blog.csdn.net/u013704227/article/details/77604500

历史推荐

人人都是数据分析师，人人都能玩转Pandas | Numpy 精品系列教程汇总 | 我是如何入门机器学习的呢 | 谷歌机器学习43条黄金法则