08 | 模型的评估指标

2018-06-21 王天一

讲述：王天一
时长 17:51 大小 8.52M

用训练数据集拟合出备选模型的参数，再用验证数据集选出最优模型后，接下来就到了是骡子是马牵出来溜溜，也就是模型评估的阶段了。模型评估中使用的是测试数据集，通过衡量模型在从未出现过的数据上的性能来估计模型的泛化特性。为简便起见，我将以二分类任务为例来说明度量模型性能的不同指标。

二分类任务是最重要也最基础的机器学习任务，其最直观的性能度量指标就是**分类的准确率**。给定一组训练数据，算法不可能完全正确地划分所有实例，而是会将一部分正例误判为反例，也会将一部分反例误判为正例。**分类正确的样本占样本总数的比例是精度（accuracy），分类错误的样本占样本总数的比例是错误率（error rate），两者之和等于 1。**

在现实生活中，二分类任务的一个实际应用就是疾病的诊断。你可以回忆一下在“贝叶斯视角下的机器学习”中提到的例子：“Jo 去进行某种疾病的检查。已知检查的准确率是 95%，也就是此病患者的检查结果 95% 会出现阳性，非此病患者的检查结果 95% 会出现阴性，同时在 Jo 的类似人群中，此病的发病率是 1%。如果 Jo 的检查结果呈阳性，那么她患病的概率是多大呢？”

这个例子就是一个典型的二分类问题。根据之前的分析结果，即使 Jo 的检查结果呈现阳性，她患病的概率也只有 16%，如果一个庸医完全按照检查结果判定的话，每 6 个病人里他就要误诊 5 个！（这又是频率主义直观的看法）但是需要注意的是，错误的分类不仅包括假阳性这一种情况，假阴性也要考虑在内——也就是确实生病的患者没有被检查出来的情形，假阳性和假阴性共同构成所有的误分类结果。

那么在 Jo 的例子中，出现假阴性的可能性有多大呢？同样令随机变量 a 表示 Jo 的真实健康状况，$a=1$ 表示 Jo 生病，$a=0$ 表示 Jo 没病；令随机变量 b 表示 Jo 的检查结果，$b=1$ 表示阳性，$b=0$ 表示阴性。由此可以计算出 Jo 的检查结果呈阴性，但是她患病的概率

$$
\begin{aligned}
p(a=1|b=0) &= \frac{p(b=0|a=1)\cdot p(a=1)}{p(b=0|a=1)\cdot p(a=1)+p(b=0|a=0)\cdot p(a=0)} \\
&= \frac{0.05\times 0.01}{0.05\times 0.01+0.95\times 0.99} = 0.053\%
\end{aligned}
$$

可以看出，虽然这个检查容易把没病的人误诊成有病，但把有病的人误诊成没病的概率是极低的。这符合我们一贯的认知：在现实中，假阳性无非就是给患者带来一些不必要的精神压力，通常不会产生更加严重的后果；可假阴性却可能让患者错过最佳的治疗时机，一旦发现便为时已晚。因此，在医学检查中本着“宁可错判，不能放过”的原则，对假阴性的要求比对假阳性的要求更加严格。

不光是在医学中，在很多情况下将正例误判为反例和将反例误判为正例的代价都是不同的，这也是数理统计将分类错误分为一类错误和二类错误的原因。为了更清楚地体现出不同的错误类型的影响，**机器学习采用了混淆矩阵（confusion matrix），也叫列联表（contingency table）来对不同的划分结果加以区分。**

		真实值		总数
		p	*n*	
预测输出	*p′*	真阳性 (TP)	伪阳性 (FP)	P′
	n′	伪阴性 (FN)	真阴性 (TN)	N′
总数		P	N	

混淆矩阵（图片来自维基百科）

如上图所示，**在混淆矩阵中，所有测试样例被分为真正例（true positive, TP）、假正例（false positive, FP）、假反例（false negative, FN）、真反例（true negative, TN）四大类。**真正例和真反例容易理解，假正例指的是样例本身是反例而预测结果是正例，也就是假阳性；假反例指的是样例本身是正例而预测结果是反例，也就是假阴性。

这样的分类能够对机器学习模型的性能做出更加精细的刻画，**查准率**（precision）和**查全率**（recall）就是两个具体的刻画指标。

查准率 P 也叫正例预测值（positive predictive value），表示的是真正例占所有预测结果为正例的样例的比值，也就是模型预测结果的准确程度，写成数学表达式是

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

查全率 R 也叫真正例率（true positive rate, TPR），表示的是真正例占所有真实情况为正例的样例的比值，也就是模型对真实正例的判断能力，写成数学表达式是

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

通俗地说，查准率要求把尽可能少的真实负例判定为预测正例，查全率则要求把尽可能少的真实正例判定为预测负例。

一般情况下，查准率和查全率是鱼和熊掌不可兼得的一对指标。使用比较严苛的判定标准可以提高查准率，比如医学上对青光眼的诊断主要依赖于眼压值，将诊断阈值设定得较高可以

保证所有被诊断的患者都是真正的病人，从而得到较高的查准率。可这样做会将症状不那么明显的初期患例都划分为正常范畴，从而导致查全率的大幅下降。

反过来，如果将眼压的诊断阈值设定得较低，稍有症状的患者都会被诊断为病人。这样做固然可以保证真正的病人都被确诊，使查全率接近于 100%，但确诊的病例中也会包含大量的疑似患者，指标稍高的健康人也会被误诊为病人，从而导致查准率的大幅下降。

将查准率和查全率画在同一个平面直角坐标系内，得到的就是 P-R 曲线，它表示了模型可以同时达到的查准率和查全率。如果一个模型的 P-R 曲线能够完全包住另一个模型的曲线，就意味着前者全面地优于后者。可是更普遍的情况是有些模型查全性能较优，而另一些模型查准性能较优，这就需要根据任务本身的特点来加以选择了。

除了 P-R 曲线外，另一个对机器学习模型性能进行可视化的方式是**受试者工作特征曲线**（receiver operating characteristic curve），简称**ROC 曲线。**ROC 这个名字来源于曲线的原始用途：判断雷达接收到的信号到底是敌机还是干扰。在机器学习中，这样的场景就演化为所有的样例共同符合一个混合分布，这个混合分布由正例和反例各自服从的单独概率分布叠加组成。此时二分类模型的任务就是确定新来的样本究竟来源于哪个分布。数据中的随机变化在分类器中体现为阈值动态取值的随机变化，分类器的性能则取决于两个概率分布之间的分离程度。

ROC 曲线描述的是真正例率和假正例率之间的关系，也就是收益（真正例）与代价（假正例）之间的关系。所谓的假正例率（false positive rate, FPR）等于假正例和所有真实反例之间的比值，其数学表达式为

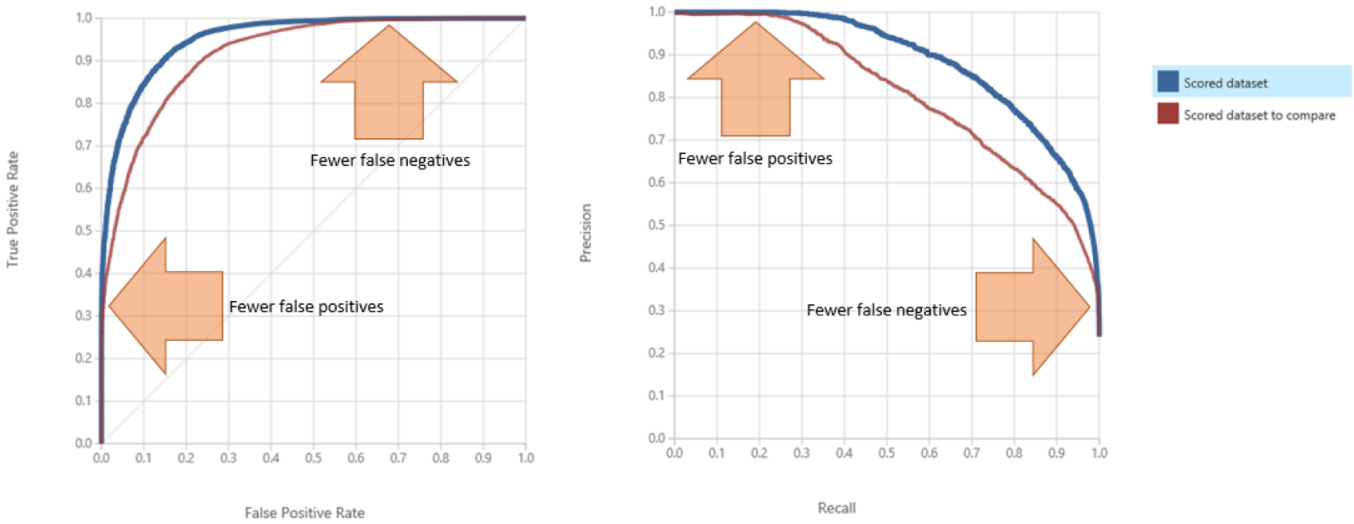
$$FPR = \frac{FP}{FP + TN}$$

ROC 空间将 FPR 定义为 X 轴，TPR 定义为 Y 轴。给定一个二元分类模型和它的阈值，就能计算出模型的 FPR 和 TPR，并映射成由 (0, 0)、(0, 1)、(1, 0)、(1, 1) 四个点围成的正方形里。在这个正方形里，从 (0, 0) 到 (1, 1) 的对角线代表了一条分界线，叫作**无识别率线**，它将 ROC 空间划分为左上 / 右下两个区域。

无识别率线描述的是随机猜测的模型，以 0.5 的概率将新来的实例判定为正例，这种模型的 TPR 和 FPR 是处处相等的。在无识别率线左上方，所有点的 TPR 都大于 FPR，意味着分类结果优于二选一的随机猜测；而在无识别率线右下方，所有点的 TPR 都小于 FPR，意味着分类结果劣于随机猜测。**完美的模型体现在 ROC 空间上的 (0, 1) 点：**$FPR = 0$ **意**

味着没有假正例，没有负例被掺入；$TPR = 1$ 意味着没有假负例，没有正例被遗漏。也就是说，不管分类器输出结果是正例还是反例，都是 100% 完全正确。

不同类型的模型具有不同的 ROC 曲线。决策树这类模型会直接输出样例对应的类别，也就是硬分类结果，其 ROC 曲线就退化为 ROC 空间上的单个点。相比之下，朴素贝叶斯这类输出软分类结果，也就是属于每个类别概率的模型就没有这么简单了。将软分类概率转换成硬分类结果需要选择合适的阈值，每个不同的阈值都对应着 ROC 空间上的一个点，因此整个模型的性能就是由多个离散点连成的折线。下图给出了 ROC 曲线和 P-R 曲线的示意图，你可以直观感受一下两者的区别。



典型的 ROC 曲线（左）与 P-R 曲线（右）

（图片来自https://blogs.msdn.microsoft.com/andreasderuiter/2015/02/09/using-roc-plots-and-the-auc-measure-in-azure-ml/）

ROC 曲线可以用来衡量习得模型的性能。模型的 ROC 曲线越靠近左上方，其性能就越好。和 P-R 曲线一样，如果一个模型的 ROC 曲线能够完全包住另一个模型的曲线，那么前者的性能就优于后者。但大多数情况下，模型之间并不存在全方位的碾压性优势，自然不会出现 ROC 曲线完全包含的情形。这时要评估不同模型性能的话，就需要 ROC 曲线下面积的概念。

ROC 曲线下面积（Area Under ROC Curve）简称 AUC。由于 AUC 的计算是在 1×1 的方格里求面积，因此其取值必然在 0 到 1 之间。对于完全靠蒙的无识别率线来说，其 AUC 等于 0.5，这样的模型完全没有预测价值。一般来说，通过调整模型的阈值，可以让

模型的最优 AUC 大于 0.5，达到比随机猜测更好的判别效果。如果模型的 AUC 比 0.5 还小，这样的模型可以通过求解其镜像，也就是将分类结果反转来获得优于随机猜测的结果。

但 ROC 曲线的意义不仅限于求解面积，它还可以提供其他的信息。不同性能的算法对应着 ROC 空间上不同的点，如果能够确定所有样例中真实正例的比例 pos 和真实负例的比例 $1-pos$，那么模型的精度就可以表示为 $pos \cdot TPR + (1-pos) \cdot (1-FPR)$。根据这个数量关系可以得出，虽然不同的模型具有不同的 TPR 和 FPR，但它们的精度是可以相等的。在 ROC 空间上，这些精度相同的模型都落在同一条斜率为 $(1-pos)/pos$，也就是负例与正例比值的直线上，这样的直线就被称为**等精度线**（iso-accuracy lines）。

由此，正例和负例的比例就可以作为已知的先验信息指导模型的选择。如果正例和负例的比例约为 2:1，那就可以在 ROC 空间上作一条斜率为 1/2 且经过 (0, 1) 的直线，并向右下方平行移动。当平移的直线与 ROC 曲线相交时，交点所对应的模型就是适用于这个先验信息的最优模型。此时最优模型的精度是多少呢？就是交点所在直线的截距，也就是和 TPR 轴的交点。

今天我和你分享了对机器学习模型不同的性能度量方法，其要点如下：

在二分类任务中，模型性能度量的基本指标是精度和错误率，两者之和为 1；

混淆矩阵是个 2×2 的性能度量矩阵，其元素分别是真正例、假正例、假反例和真反例的数目；

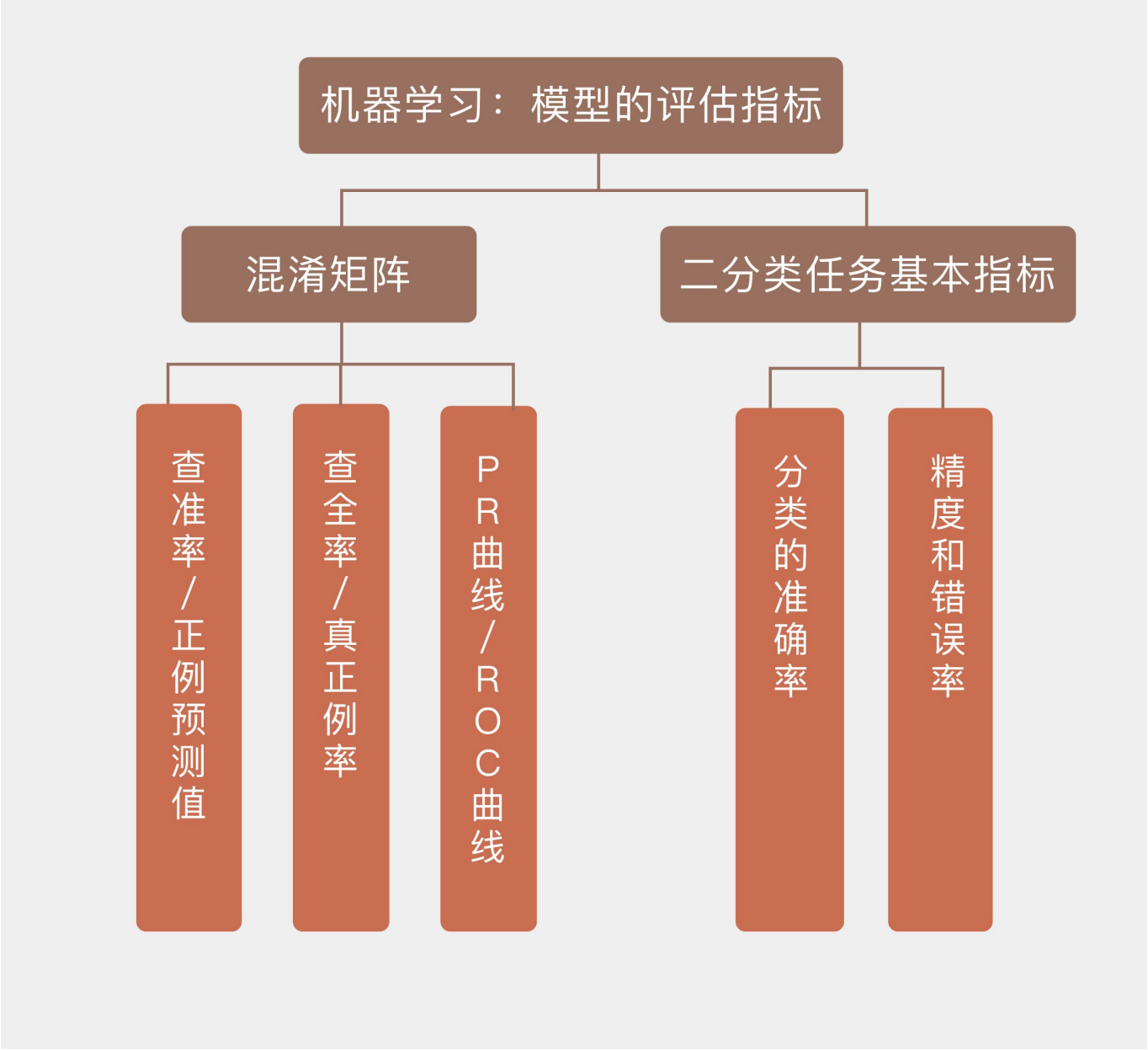
P-R 曲线表示的是查准率和查全率之间的关系，曲线在点 (1, 1) 上达到最优性能；

ROC 曲线表示的是真正例率和假正例率之间的关系，曲线在点 (0, 1) 上达到最优性能。

关于模型性能的评估我想给你推荐一位学者，他就是英国布里斯托尔大学的彼得·弗拉克（Peter Flach）。这位教授在模型评估研究，尤其是 ROC 曲线分析上的造诣颇深，你可以在他的著作《机器学习》（Machine Learning）第 2 章和论文中领会模型评估中蕴藏的信息，这一定会让你受益匪浅。

对比 P-R 图和 ROC 曲线会发现一个有趣的现象，那就是当类别的平衡性，也就是数据中正例和负例的比例发生改变时，这种变化不会给 ROC 曲线带来变化，却会让 P-R 曲线产生明显的改变，为什么会出现这种现象呢？

你可以思考一下，并在这里分享你的观点。



上一篇 07 | 模型的验证方法

下一篇 09 | 实验设计

精选留言 (7)

写留言

冬瓜
2018-10-11

5

我的经验是，正负样本比例不平衡时，不能通过ROC曲线来评估模型，可能会出现ROC很好看，但业务上无法使用的情况。

具体来说就是，roc很高，但是PR很低，查准率和查全率都无法满足使用要求。
...
展开

作者回复: 您的例子非常好这就体现出样例平衡的作用了，脱离实际情况空谈指标有时会误事。

KingZone
2018-07-25

2

正样本10000个，负样本3个，那么查全率（即召回率）是不是很低？！

展开

作者回复: 查全率取决于模型的精确度，也就是正例有多少判定正确。但样本不平衡会导致对模型精确性判断的偏差，即使这3个负样本全部分类错误也说明不了问题，因为数量太少了。

林彦
2018-06-21

1

ROC曲线中的TPR的分子和分母里的TP，FN都来自正例，FPR的分子，分母里的FP，TN都来自负例。PR曲线中的Precision的分母里的TP和FP则会同时受到正例和负例的影响。

当样本的正负例比例发生较大变化时，原来同一类型的样本点在PR曲线受到的影响由于Precision值的明显变化，会比ROC曲线要大。这只是我的粗浅推测。怎么推导还是不明...

展开

作者回复: 其实你已经找到点了。从混淆矩阵看，ROC的两个指标计算分别对应两个列，也就是不同类别真实输出上的准确率。只要算法不发生变化，那准确率就不会受到样本数的影响。
反过来，PR的两个指标在混淆矩阵里是一行一列，一个考察真实输出的准确率，一个考察预测输出的准确率。当数据类别不平衡导致各类真假正负例的数目改变时，这一行一列在计算比例时就没法保证相同的变化尺度，导致PR曲线变形。

TranQ
2019-04-08

老师，除了分类模型之外，其他模型也可以用精度，ROC曲线等这类工具进行评估吗？例如回归模型。如果可以，那么大致思路是什么呢？

code-arti...
2019-01-27

对这几个概念和对应的计算公式很难有一个直觉的把握。需要在实际项目上练出这种直觉吗？

paradox
2018-08-09

老师，您好
文中：
此时最优模型的精度是多少呢？就是交点所在直线的截距，也就是和 TPR轴的交点。

精度的数值是不是应该是 截距*pos?...
展开

作者回复: 精度的数值取决于等精度线和ROC曲线交点的位置

张权
2018-06-21

机器学习课程只通过语音讲授，效果不太好。好多东西很难解释的很清楚。
展开

作者回复: 是的，图片和公式还是要看文本。