

机器学习评估指标

小虎仔

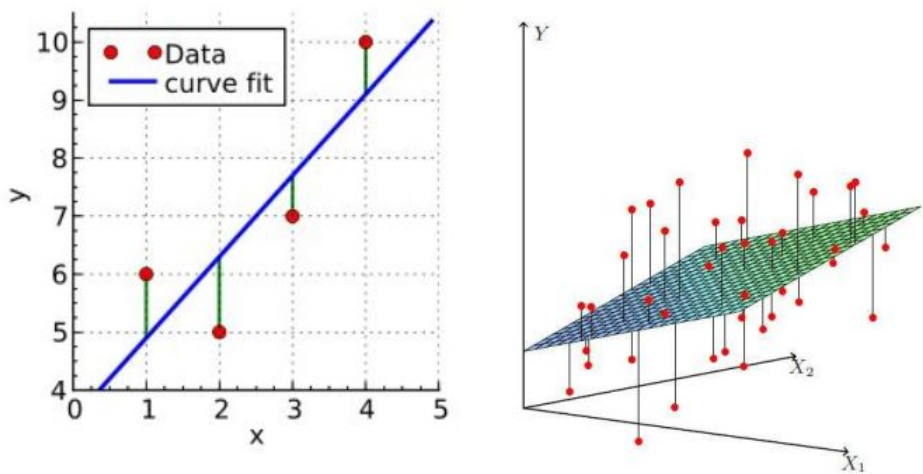
215 人赞同了该文章

机器学习评估指标

1、回归 (Regression) 算法指标

- Mean Absolute Error 平均绝对误差
- Mean Squared Error 均方误差
- Root Mean Squared Error: 均方根误差
- Coefficient of determination 决定系数

以下为一元变量和二元变量的线性回归示意图:



怎样来衡量回归模型的好坏呢？我们第一眼自然而然会想到采用残差（实际值与预测值差值）的均值来衡量，即：

$$\text{residual}(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^n (y_i - \hat{y}_i)$$

问题 1：用残差的均值合理吗？

当实际值分布在拟合曲线两侧时，对于不同样本而言有正有负，相互抵消，因此我们想到采用预测值和真实值之间的距离来衡量。

1.1 平均绝对误差 MAE

平均绝对误差MAE（Mean Absolute Error）又被称为 L1范数损失。

$$\text{MAE}(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^n |y_i - \hat{y}_i|$$

问题 2：MAE有哪些不足？

MAE虽能较好衡量回归模型的好坏，但是绝对值的存在导致函数不光滑，在某些点上不能求导，可以考虑将绝对值改为残差的平方，这就是均方误差。

1.2 均方误差 MSE

均方误差MSE（Mean Squared Error）又被称为 L2范数损失。

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

问题 3：还有没有比MSE更合理一些的指标？

由于MSE与我们的目标变量的量纲不一致，为了保证量纲一致性，我们需要对MSE进行开方。

1.3 均方根误差 RMSE

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

问题 4：RMSE有没有不足的地方？有没有规范化（无量纲化的指标）？

上面的几种衡量标准的取值大小与具体的应用场景有关系，很难定义统一的规则来衡量模型的好坏。比如说利用机器学习算法预测上海的房价RMSE在2000元，我们是可以接受的，但是当四五线城市的房价RMSE为2000元，我们还可以接受吗？下面介绍的决定系数就是一个无量纲化的指标。

1.4 决定系数 R^2

变量之所以有价值，就是因为变量是变化的。什么意思呢？比如说一组因变量为[0, 0, 0, 0, 0]，显然该因变量的结果是一个常数0，我们也没有必要建模对该因变量进行预测。假如一组的因变量为[1, 3, 7, 10, 12]，该因变量是变化的，也就是有变异，因此需要通过建立回归模型进行预测。这里

▲ 赞同 215

● 9 条评论

➦ 分享

❤ 喜欢

★ 收藏

📄 申请转载

...

赞同 215

分享

$$SST = \sum_i^m (y_i - \bar{y})^2 \quad SST = \text{total sum of squares}$$

$$SSR = \sum_i^m (\hat{y}_i - \bar{y})^2 \quad SSR = \text{sum of due to regression}$$

$$SSE = \sum_i^m (\hat{y}_i - y_i)^2 \quad SSE = \text{sum of due to errors}$$

$$SST = SSR + SSE$$

$$R^2(y, \hat{y}) = \frac{SSR}{SST}$$

如果结果是0，就说明模型预测不能预测因变量。如果结果是1，就说明是函数关系。如果结果是0-1之间的数，就是我们模型的好坏程度。化简上面的公式，分子就变成了我们的均方误差MSE，下面分母就变成了方差：

$$R^2(y, \hat{y}) = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2 / m}{\sum_{i=1}^m (y_i - \bar{y})^2 / m} = 1 - \frac{MSE(\hat{y}, y)}{\text{Var}(y)}$$

问题 5：以上评估指标有没有缺陷，如果有，该怎样改进？

以上的评估指标是基于误差的均值进行评估的，均值对异常点（outliers）较敏感，如果样本中有一些异常值出现，会对以上指标的值有较大影响，即均值是非鲁棒的。

1.5 解决评估指标鲁棒性问题

我们通常用一下两种方法解决评估指标的鲁棒性问题：

- 剔除异常值
- 设定一个相对误差，当该值超过一定的阈值时，则认为其是一个异常点，剔除这个异常点，将异常点剔除之后，再计算平均误差来对模型进行评价。
- 使用误差的分位数来代替
- 如利用中位数来代替平均数。例如 MAPE：

$$MAPE = \text{median}(|y_i - \hat{y}_i| / y_i)$$

MAPE是一个相对误差的中位数，当然也可以使用别的分位数。

2、分类 (Classification) 算法指标

- 精度 Accuracy
- 混淆矩阵 Confusion Matrix
- 准确率 (查准率) Precision
- 召回率 (查全率) Recall
- Fβ Score
- AUC Area Under Curve
- KS Kolmogorov-Smirnov

2.1 精度 Acc

预测正确的样本的占总样本的比例，取值范围为[0,1]，取值越大，模型预测能力越好。

其中：

$$sign(\hat{y}_i, y_i) = \begin{cases} 1 & \hat{y}_i = y_i \\ 0 & \hat{y}_i \neq y_i \end{cases}$$

精度评价指标对平等对待每个类别，即每一个样本判对 (0) 和判错 (1) 的代价都是一样的。

问题 6：精度有什么缺陷？什么时候精度指标会失效？

- 对于有倾向性的问题，往往不能用精度指标来衡量。
- 比如，判断空中的飞行物是导弹还是其他飞行物，很显然为了减少损失，我们更倾向于相信是导弹而采用相应的防护措施。此时判断为导弹实际上是其他飞行物与判断为其他飞行物实际上是导弹这两种情况的重要性是不一样的；
- 对于样本类别数量严重不均衡的情况，也不能用精度指标来衡量。
- 比如银行客户样本中好客户990个，坏客户10个。如果一个模型直接把所有客户都判断为好客户，得到精度为99%，但这显然是没有意义的。

对于以上两种情况，单纯根据Accuracy来衡量算法的优劣已经失效。这个时候就需要对目标变量的真实值和预测值做更深入的分析。

2.2 混淆矩阵 Confusion Matrix

	预测值 正	预测值 负
真实值 正	TP	FN
真实值 负	FP	TN

这里牵扯到三个方面：真实值，预测值，预测值和真实值之间的关系，其中任意两个方面都可以确定第三个。

通常取预测值和真实值之间的关系、预测值对矩阵进行划分：

- True positive (TP)
- 真实值为Positive，预测正确（预测值为Positive）
- True negative (TN)
- 真实值为Negative，预测正确（预测值为Negative）
- False positive (FP)
- 真实值为Negative，预测错误（预测值为Positive），第一类错误，Type I error。
- False negative (FN)
- 真实值为Positive，预测错误（预测值为 Negative），第二类错误，Type II error。

2.3 准确率（查准率） Precision

Precision 是分类器预测的正样本中预测正确的比例，取值范围为[0,1]，取值越大，模型预测能力越好。

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

2.4 召回率（查全率） Recall

赞同 215

分享

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

应用场景：

1. 地震的预测 对于地震的预测，我们希望的是Recall非常高，也就是说每次地震我们都希望预测出来。这个时候我们可以牺牲Precision。情愿发出1000次警报，把10次地震都预测正确了；也不要预测100次对了8次漏了两次。

- “宁错拿一万，不放过一个”，分类阈值较低

1. 嫌疑人定罪 基于不错怪一个好人的原则，对于嫌疑人的定罪我们希望是非常准确的。即使有时候放过了一些罪犯，但也是值得的。因此我们希望有较高的Precision值，可以合理地牺牲Recall。

- “宁放过一万，不错拿一个”，“疑罪从无”，分类阈值较高

问题 7：某一家互联网金融公司风控部门的主要工作是利用机器模型抓取坏客户。互联网金融公司要扩大业务量，尽量多的吸引好客户，此时风控部门该怎样调整Recall和Precision？如果公司坏账扩大，公司收紧业务，尽可能抓住更多的坏客户，此时风控部门该怎样调整Recall和Precision？

如果互联网公司要扩大业务量，为了减少好客户的误抓率，保证吸引更多的好客户，风控部门就会提高阈值，从而提高模型的查准率Precision，同时，导致查全率Recall下降。如果公司要收紧业务，尽可能抓住更多的坏客户，风控部门就会降低阈值，从而提高模型的查全率Recall，但是这样会导致一部分好客户误抓，从而降低模型的查准率 Precision。

根据以上几个案，我们知道随着阈值的变化Recall和Precision往往会向着反方向变化，这种规律很难满足我们的期望，即Recall和Precision同时增大。

问题 8：有没有什么方法权衡Recall和Precision 的矛盾？

我们可以用一个指标来统一Recall和Precision的矛盾，即利用Recall和Precision的加权调和平均值作为衡量标准。

2.5 F_β Score

Precision和Recall 是互相影响的，理想情况下肯定是做到两者都高，但是一般情况下Precision高、Recall 就低，Recall 高、Precision就低。为了均衡两个指标，我们可以采用Precision和Recall的加权调和平均（weighted harmonic mean）来衡量，即F_β Score，公式如下：

$$F_{\beta} = (1 + \beta^2) \times \frac{P \times R}{\beta^2 \times P + R}$$

β表示权重：

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

由于Fβ Score 无法直观反映数据的情况，同时业务含义相对较弱，实际工作用到的不多。

2.6 ROC 和 AUC

AUC是一种模型分类指标，且仅仅是二分类模型的评价指标。AUC是Area Under Curve的简称，那么Curve就是 ROC（Receiver Operating Characteristic），翻译为"接受者操作特性曲线"。也就是说ROC是一条曲线，AUC是一个面积值。

2.6.1 ROC

ROC曲线为 FPR 与 TPR 之间的关系曲线，这个组合以 FPR 对 TPR，即是以代价 (costs) 对收益 (benefits)，显然收益越高，代价越低，模型的性能就越好。

- x 轴为假阳性率（FPR）：在所有的负样本中，分类器预测错误的比例

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

- y 轴为真阳性率（TPR）：在所有的正样本中，分类器预测正确的比例（等于Recall）

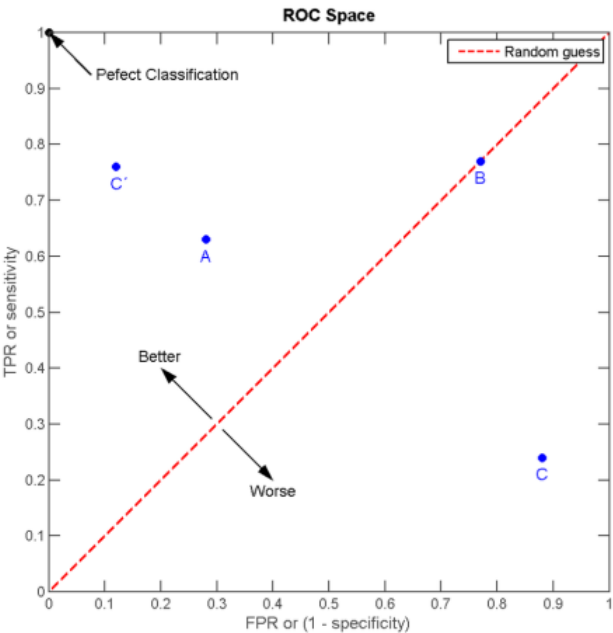
$$TPR = \frac{TP}{TP+FN}$$

为了更好地理解ROC曲线，我们使用具体的实例来说明：

如在医学诊断的主要任务是尽量把生病的人群都找出来，也就是TPR越高越好。而尽量降低没病误诊为有病的人数，也就是FPR越低越好。

不难发现，这两个指标之间是相互制约的。如果某个医生对于有病的症状比较敏感，稍微的小症状都判断为有病，那么他的TPR应该会很很高，但是FPR也就相应地变高。最极端的情况下，他把所有的样本都看做有病，那么TPR达到1，FPR也为1。

我们以FPR为横轴，TPR为纵轴，得到如下ROC空间：

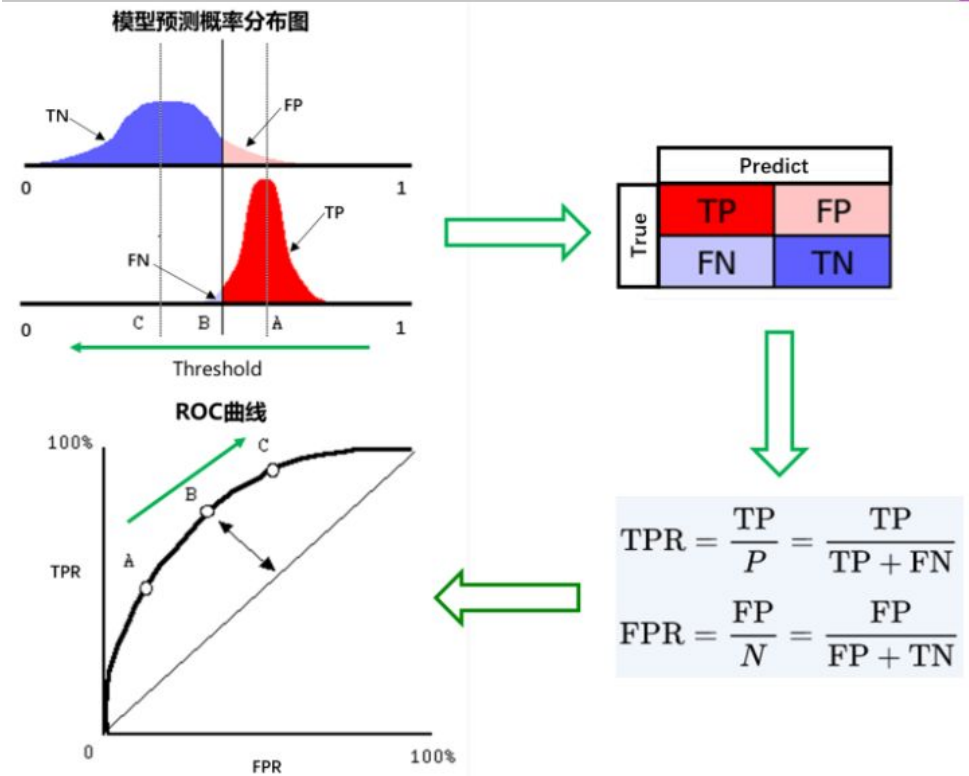


我们可以看出，左上角的点(TPR=1，FPR=0)，为完美分类，也就是这个医生医术高明，诊断全对。点A(TPR>FPR),医生A的判断大体是正确的。中线上的点B(TPR=FPR),也就是医生B全都是蒙的，蒙对一半，蒙错一半；下半平面的点C(TPR<FPR)，这个医生说你有病，那么你可能没有

遍低于得病人群的输出概率值（即正常人诊断出疾病的概率小于得病人群诊断出疾病的概率）。

竖线代表阈值。显然，图中给出了某个阈值对应的混淆矩阵，通过改变不同的阈值，得到一系列的混淆矩阵，进而得到一系列的TPR和FPR，绘制出ROC曲线。

阈值为1时，不管你什么症状，医生均未诊断出疾病（预测值都为N），此时，位于左下。阈值为0时，不管你什么症状，医生都诊断结果都是得病（预测值都为P），此时，位于右上。



2.6.2 AUC

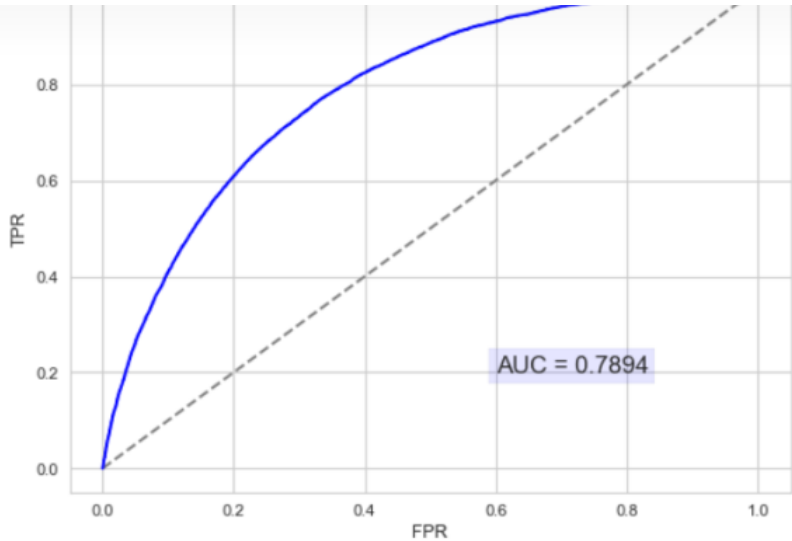
AUC定义：

- AUC 值为 ROC 曲线所覆盖的**区域面积**，显然，AUC越大，分类器分类效果越好。
- AUC = 1，是完美分类器。
- 0.5 < AUC < 1，优于随机猜测。有预测价值。
- AUC = 0.5，跟随机猜测一样（例：丢铜板），没有预测价值。
- AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

注：对于AUC小于0.5的模型，我们可以考虑取反（模型预测为positive，我们就取negative），这样就可以保证模型的性能不可能比随机猜测差。

以下为ROC曲线和AUC值的实例：

赞同 215



AUC的物理意义 AUC的物理意义正样本的预测结果大于负样本的预测结果的概率。所以AUC反应的是分类器对样本的排序能力。另外值得注意的是，AUC对样本类别是否均衡并不敏感，这也是不均衡样本通常用AUC评价分类器性能的一个原因。

问题 13：小明一家四口，小明5岁，姐姐10岁，爸爸35岁，妈妈33岁，建立一个逻辑回归分类器，来预测小明家人为成年人概率。

以下为三种模型的输出结果，求三种模型的 AUC：

	小明	姐姐	妈妈	爸爸
a	0.12	0.35	0.76	0.85
b	0.12	0.35	0.44	0.49
c	0.52	0.65	0.76	0.85

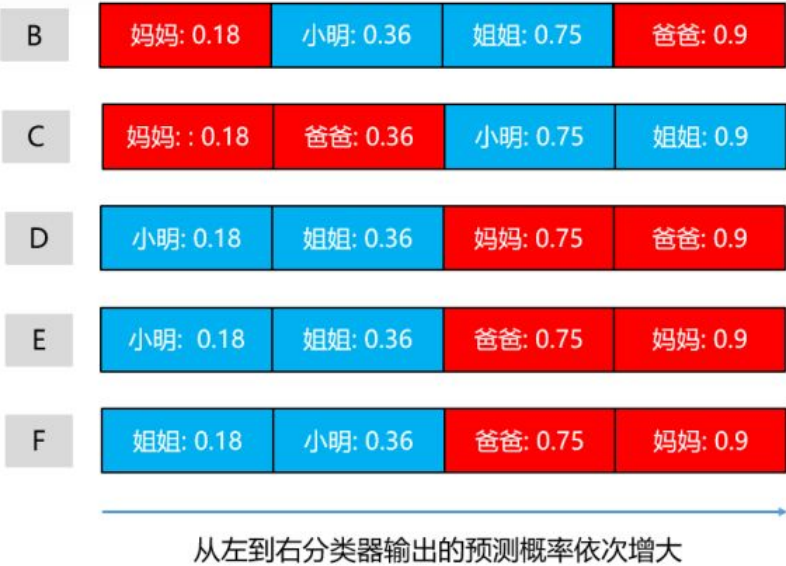
1. AUC更多的是关注对计算概率的排序，关注的是概率值的相对大小，与阈值和概率值的绝对大小没有关系 例子中并不关注小明是不是成人，而关注的是，预测为成人的概率的排序。
2. AUC只关注正负样本之间的排序，并不关心正样本内部，或者负样本内部的排序。这也体现了AUC的本质：任意个正样本的概率都大于负样本的概率的能力。

例子中AUC只需要保证（小明和姐姐）（爸爸和妈妈），小明和姐姐在前2个排序，爸爸和妈妈在后2个排序，而不会考虑小明和姐姐谁在前，或者爸爸和妈妈谁在前。AUC只与概率的相对大小（概率排序）有关，和绝对大小没关系。由于三个模型概率排序的前两位都是未成年人（小明，姐姐），后两位都是成年人（妈妈，爸爸），因此三个模型的AUC都等于1。

问题 14：以下已经对分类器输出概率从小到大进行了排列，哪些情况的AUC等于1，情况的AUC为0（其中背景色表示True value，红色表示成年人，蓝色表示未成年人）。

赞同 215

分享



D 模型, E模型和F模型的AUC值为1, C 模型的AUC值为0 (爸妈为成年人的概率小于小明和姐姐, 显然这个模型预测反了)。

AUC的计算

- 法1: AUC为ROC曲线下的面积, 那我们直接计算面积可得。面积为一个个小的梯形面积 (曲线) 之和。计算的精度与阈值的精度有关。
- 法2: 根据AUC的物理意义, 我们计算正样本预测结果大于负样本预测结果的概率。取n1*n0(n1为正样本数, n0为负样本数)个二元组, 每个二元组比较正样本和负样本的预测结果, 正样本预测结果高于负样本预测结果则为预测正确, 预测正确的二元组占总二元组的比率就是最后得到的AUC。时间复杂度为O(N* M)。
- 法3: 我们首先把所有样本按照score排序, 依次用rank表示他们, 如最大score的样本, rank=n (n=n0+n1, 其中n0为负样本个数, n1为正样本个数), 其次为n-1。那么对于正样本中rank最大的样本, rank_max, 有n1-1个其他正样本比他score小,那么就有(rank_max-1)-(n1-1)个负样本比他score小。其次为(rank_second-1)-(n1-2)。最后我们得到正样本大于负样本的概率为:

$$AUC = \frac{\sum_{\text{正样本}} rank(score) - \frac{n_1*(n_1+1)}{2}}{n_0 * n_1}$$

其计算复杂度为O(N+M)。

下面有一个简单的例子:

真实标签为 (1, 0, 0, 1, 0) 预测结果1 (0.9, 0.3, 0.2, 0.7, 0.5) 预测结果2 (0.9, 0.3, 0.2, 0.7, 0.8))

分别对两个预测结果进行排序, 并提取他们的序号 结果1 (5, 2, 1, 4, 3) 结果2 (5, 2, 1, 3, 4)

对正分类序号累加 结果1: SUM正样本 (rank(score))=5+4=9 结果2: SUM正样本 (rank(score))=5+3=8

计算两个结果的AUC: 结果1: AUC= (9-2*3/2)/6=1 结果2: AUC= (8-2*3/2)/6=0.833

问题 15: 为什么说 ROC 和AUC都能应用于非均衡的分类问题?

ROC曲线只与横坐标 (FPR) 和 纵坐标 (TPR) 有关系。我们可以发现TPR只是正样本中预测正确的概率, 而FPR是负样本中预测错误的概率, 和正负样本的比例没有关系, 因此 ROC 曲线与正负

2.7 KS Kolmogorov-Smirnov

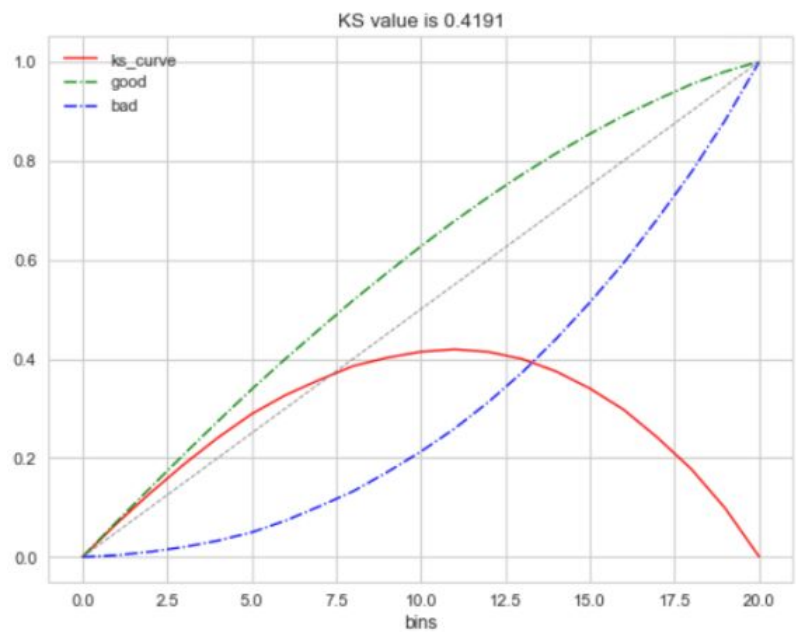
KS值是在模型中用于**区分预测正负样本分隔程度**的评价指标，一般应用于金融风控领域。

与ROC曲线相似，ROC是以FPR作为横坐标，TPR作为纵坐标，通过改变不同阈值，从而得到ROC曲线。

而在KS曲线中，则是以阈值作为横坐标，以FPR和TPR作为纵坐标，ks曲线则为TPR-FPR，ks曲线的最大值通常为ks值。

为什么这样求KS值呢？我们知道，当阈值减小时，TPR和FPR会同时减小，当阈值增大时，TPR和FPR会同时增大。而在实际工程中，我们希望TPR更大一些，FPR更小一些，即TPR-FPR越大越好，即ks值越大越好。



可以理解TPR是收益，FPR是代价，ks值是收益最大。图中绿色线是TPR、蓝色线是FPR。



编辑于 2018-05-01 18:56

机器学习 深度学习 (Deep Learning) 数据挖掘

文章被以下专栏收录

- **Python**
Python的一些应用
- **通俗易懂的机器学习和深度学习**
主要包括机器学习，深度学习，自然语言处理，cv等

推荐阅读


▲ 赞同 215 ▼ ● 9 条评论 ➦ 分享 ♥ 喜欢 ★ 收藏 📄 申请转载 ...

知乎

首发于Python


无碍

写文章




机器学习评估指标

呈序员阿德



机器学习篇-评估机器学习的模型

眼睛流产



机器学习中的性能度量

yyHaker

我们面临的问题 最近发现，有不少做风控的小伙伴，脑子里有互金机器学习、数据分析的概念，但缺少整体做风控架构、系统、策略的思维，具体可以表现为对机器学习评分卡了如指掌，对专家评分...
正阳

9 条评论

切换为时间排序

写下你的评论...



竹露清响

2021-04-06

很全面

赞



中华好少年

2020-11-30

您好，这篇文章写的太好了，但是我有疑问，我想请教一下，为什么有的时候测出来一个分类器ACC，AUC的值比另一个大，而f1-score值却比另一个分类器小呢？这三个指标之间的大小有什么必然联系吗？

赞



double dimension

2020-05-03

可有详细介绍评估指标的论文么，十分感谢

赞



果儿 回复 double dimension

2021-07-02

请问您找到了吗

赞



知乎用户

2020-04-14

MAPE的公式写错了

赞



知乎用户

2019-12-17

对roc auc理解很到位，面试很喜欢问这样的问题

赞



Elsie

2019-12-16

您好，FN对应的是真实值为positive，预测值为negative，您应该写错了。

赞



zakki

2019-07-10

学习了！

赞



晓伟

2019-05-10

写得真的很棒！

赞

赞同 215

9 条评论

分享

喜欢

收藏

申请转载

...