1.什么是AUC?

面试的时候，一句话说明AUC的本质和计算规则：

AUC：一个正例，一个负例，预测为正的概率值比预测为负的概率值还要大的可能性。

所以根据定义：我们最直观的有两种计算AUC的方法

1：绘制ROC曲线，ROC曲线下面的面积就是AUC的值

2：假设总共有（m+n）个样本，其中正样本m个，负样本n个，总共有m*n个样本对，计数，正样本预测为正样本的概率值大于负样本预测为正样本的概率值记为1，累加计数，然后除以（m*n）就是AUC的值

PS：百度百科，随机挑选一个正样本以及一个负样本，当前的分类算法根据计算得到的Score值将这个正样本排在负样本前面的概率就是AUC值。这里的score值就是预测为正的概率的值，排在前面表示的是正样本的预测为正的概率值大于负样本的预测为正的概率值

AUC是一个模型评价指标，只能用于二分类模型的评价，对于二分类模型，还有很多其他评价指标，比如logloss，accuracy,precision。

标，比如logloss，accuracy，precision。如果你经常关注数据挖掘比赛，比如kaggle，那你会发现AUC和logloss基本是最常见的模型评价指标。为什么AUC和logloss比accuracy更常用呢？因为很多机器学习的模型对分类问题的预测结果都是概率，如果要计算accuracy，需要先把概率转化成类别，这就需要手动设置一个阈值，如果对一个样本的预测概率高于这个预测，就把这个样本放进一个类别里面，低于这个阈值，放进另一个类别里面。所以这个阈值很大程度上影响了accuracy的计算。使用AUC或者logloss可以避免把预测概率转换成类别。

AUC(Area under curve)，ROC曲线下的面积。ROC曲线是基于样本的真实类别和预测概率来画的。X轴是伪阳性率（false positive rate)，y轴是真阳性率（true positive rate)。

		真实类别	
		1(P)	0(N)
预测类别	1(P')	真阳性	伪阳性
	0(N')	伪阴性	真阴性

真阳性率=（真阳性的数量）/（真阳性的数量+伪阴性的数量）

假阳性率=（伪阳性的数量）/（伪阳性的数量+真阴性的数量）

有了上面两个公式，我们就可以计算真、伪阳性率了。但是如何根据预测的概率得到真伪阳性、阴性的数量。

我们来看一个具体例子，比如有5个样本：

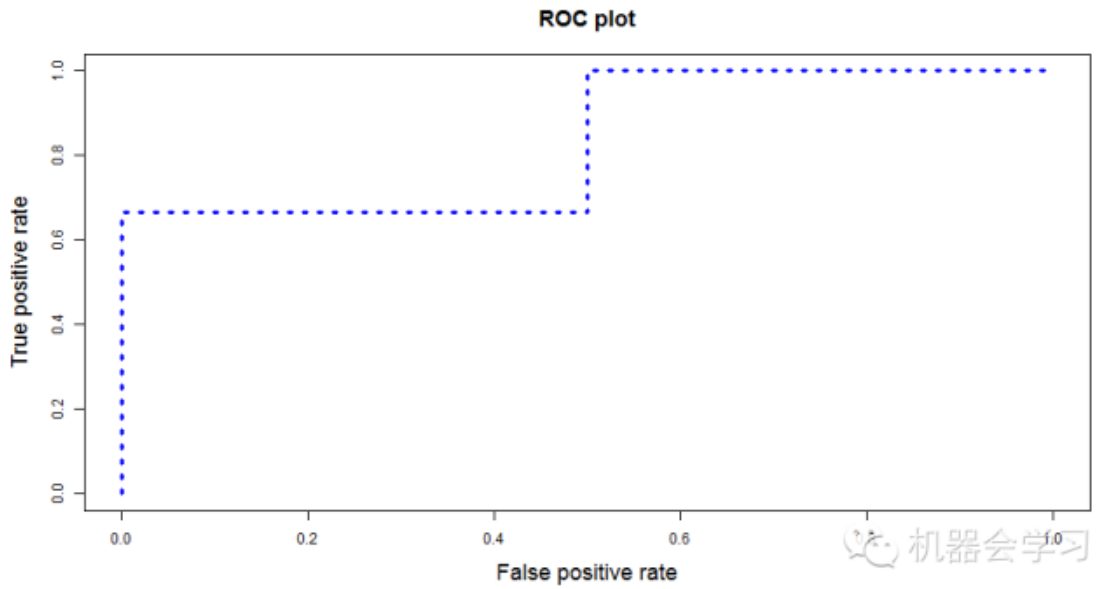
真实的类别（标签）是y=c(1,1,0,0,1)

一个分类器预测样本为1的概率是p=c(0.5,0.6,0.55,0.4,0.7)

如文章一开始所说，我们需要选定阈值才能把概率转化为类别，选定不同的阈值会得到不同的结果。如果我们选定的阈值为0.1，那5个样本被分进1的类别，如果选定0.3，结果仍然一样。如果选了0.45作为阈值，那么只有样本4被分进0，其余都进入1类。一旦得到了类别，我们就可以计算相应的真、伪阳性率，当我们把所有计算得到的不同真、伪阳性率连起来，就画出了ROC曲线，我们不需要手动做这个，因为很多程序包可以自动计算真、伪阳性率，比如在R里面，下面的代码可以计算以上例子的真、伪阳性率并且画出ROC曲线：

```
library(ROCR)
p=c(0.5,0.6,0.55,0.4,0.7)
y=c(1,1,0,0,1)
pred = prediction(p, y)
perf = performance(pred,"tpr","fpr")
plot(perf,col="blue",lty=3, lwd=3,cex.lab=1.5, cex.axis=2, cex.main=1.5,main="ROC plot'
```

上面代码可以画出下图：



得到了ROC曲线，那么曲线下面的面积也就可以算出来了，同样，我们可以通过程序得到面积：

Mann–Whitney U statistic的角度来解释，AUC就是从所有1样本中随机选取一个样本， 从所有0样本中随机选取一个样本，然后根据你的分类器对两个随机样本进行预测，把1样本预测为1的概率为p1，把0样本预测为1的概率为p0，p1>p0的概率就等于AUC。所以AUC反应的是分类器对样本的排序能力。根据这个解释，如果我们完全随机的对样本分类，那么AUC应该接近0.5。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。

有朋友用python，下面代码用于在python里面计算auc：

```
from sklearn import metrics
def aucfun(act,pred):
  fpr, tpr, thresholds = metrics.roc_curve(act, pred, pos_label=1)
  return metrics.auc(fpr, tpr)
```

【AUC的排序本质】

大部分分类器的输出是概率输出，如果要计算准确率，需要先把概率转化成类别，就需要手动设置一个阈值，而这个超参数的确定会对优化指标的计算产生过于敏感的影响

AUC从Mann–Whitney U statistic的角度来解释：随机从标签为1和标签为0的样本集中分别随机选择两个样本，同时分类器会输出两样本为1的概率，那么我们认为分类器对**“标签1样本的预测概率>对标签0样本的预测概率”的概率**等价于AUC。

因而AUC反应的是分类器对样本的排序能力，这样也可以理解AUC对不平衡样本不敏感的原因了。

【作为优化目标的各类指标】

最常用的分类器优化及评价指标是AUC和logloss，最主要的原因是：不同于accuracy，precision等，这两个指标不需要将概率输出转化为类别，而是可以直接使用概率进行计算。

顺便贴上logloss的公式

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}log(p_{i,j})$$

- N：样本数
- M：类别数，比如上面的多类别例子，M就为4
- yij：第i个样本属于分类j时为为1，否则为0
- pij：第i个样本被预测为第j类的概率

参考：https://www.zhihu.com/question/39840928