对AUC计算公式和几何意义的理解

对AUC的一些理解[1].

AUC（Area Under ROC Curve）是对机器学习模型的一种性能度量，其直观的意义是ROC（Receiver Operating Characteristic）围成的面积，然而很多文章对AUC的计算公式与其直观意义“面积”的对应描述不甚明了，于是我花了一些时间研究了AUC的计算公式和几何意义，全是我个人理解，如有不对，还望指出。
本文代码在
https://github.com/luo3300612/MachineLearningPy/blob/master/Intuition/auc.ipynb
如果看不了，就复制上面这行到
https://nbviewer.jupyter.org/
代码上我参考了
https://www.zhihu.com/people/Enuok/activities
知乎的一个用户的写法，我也把他的代码放在我的仓库里用来对比。

要想理解AUC，得从混淆矩阵说起。

混淆矩阵

对于一个**二分类**问题，每个样例有自己的**真实类别**和模型给出的**预测类别**，真实类别和预测类别都分别有两种，定义标记为1的为**正例**，标记为0的为**反例**，当我们在测试集上使用模型分类时，会有以下四种情况。

	标记为正例	**标记为反例**
实际为正例	真正例(True-Positive)	假反例(False-Negative)
实际为反例	假正例(False-Positive)	真反例(True-Negative)

下面我们用**TP、FN、FP、TN**分别表示四种结果的样本个数。

分类

在一个二分类问题中，往往我们的预测输出是一个0-1之间的数字，为了得到新的样本的类别，我们需要选择一个阈值来将这些预测结果进行分类，假设x_k是待分类的样本，$f(x)$是模型对x的预测值，$h(x)$通过一个阈值（这里以0.5为例）将x区分为正例或反例。

$$\hat{y}_k = f(x_k)$$

$$h(x_k) - \begin{cases} 1, & \text{if } \hat{y}_k >= 0.5 \\ 0, & \text{if } \hat{y}_k < 0.5 \end{cases}$$

ROC曲线

基本概念

当我们有了混淆矩阵的概念以及对测试集上样本分类的概念之后，我们就可以绘制ROC曲线，在测试集上，我们使用模型对测试集样本进行预测后，在某个阈值下进行分类得到**TP、FN、FP、TN**，计算真正例率TPR(True Positive Rate)和假正例率FPR(False Positive Rate)

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

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

TPR的直观理解就是在所有正例中被模型预测为正例的样本所占比率，FPR就是在所有反例中被模型误判为正例的样本所占比率，例如在所有10个正例样本，4个被判定为正例，6个被判定为反例，于是有4个真正例，则$TPR = 0.4$，在所有10个反例样本中，3个被判断为正例，7个被判断为反例，于是有3个假反例，则$FPR = 0.3$。如果你知道**召回率**（Recall）的话，其实TPR就是召回率。**注意到二者的分母对于同一个样本集是一个定值，分别是样本中真实标记为正例和真实标记为反例的个数。**
现在，针对不同的阈值k，我们都有一个点(FPR_k, TPR_k)，于是以FPR为横坐标，TPR为纵坐标，将所有的点画在坐标系上，我们就能得到**ROC曲线。**

ROC曲线的绘制

在绘制ROC曲线的时候，我们当然不会遍历所有的实数阈值k（事实上这也不可能），下面在有限样例的情况下绘制ROC曲线的实例，以样例数$m = 10$为例，10个样例的预测结果和真实类别如下。

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
预测结果	0.06	0.47	0.50	0.94	0.83	0.90	0.73	0.07	0.50	0.27
真实类别	1	1	1	0	1	0	0	0	1	1

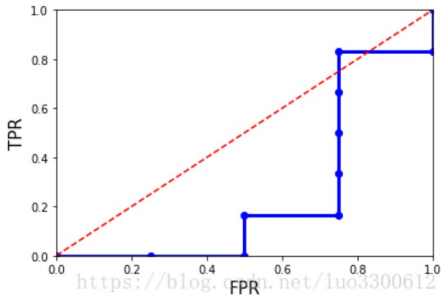
举个例子，如果阈值$k = 0.5$，则混淆矩阵为

	标记为正例	标记为反例
实际为正例	3	3
实际为反例	3	1

假设我们设置k从1开始，从大到小搜索k的值，我们发现，当我们取$k \in [0.08, 0.26]$时，混淆矩阵对于这个区间上的所有k是一样的，因为没有改变任何样本的预测类别，这是因为没有样本的预测结果在$[0.08, 0.26]$中。进而我们发现，只要我们k的变化没有使它从大于某个预测值到小于某个预测值时，我们的分类结果都不会发生改变，因此，**k只需要从大到小地遍历所有的预测值**，就可以得到ROC曲线上的所有点了。
有一个问题是，**为什么要从大到小遍历**，不可以随机地在所有样本中取**k**吗？其实，从大到小遍历是为了作图的方便，也是一种作图的规定。下面我们就能看到。
为了从大到小遍历k，将样本按照预测值从大到小排序，重新编号（因为编号不影响结果），得到

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
预测结果	0.94	0.90	0.83	0.73	0.50	0.50	0.47	0.27	0.07	0.06
真实类别	0	0	1	0	1	1	1	1	0	1

记正样本个数为m^+，负样本个数为m^-，取$k=1$，此时所有样例都被分为反例，因此真正例和假正例的比率均为0，对应原点$(0,0)$，接下来，根据上表的预测结果，从左到右依次取$k=0.94, 0.90, 0.83....$，过程等同于逐个地将样本判断为正例（$k=0.94$时，x_1就是正例，其余都是反例，$k=0.90$时，x_1, x_2是正例，其余是反例，以此类推），若前一个标记点是(x,y)，若下一个加入到正例中的样本真实类别是正例，说明我们多了一个TP，回顾真正例率的计算公式，正例样本不变，多了一个正例，则正例率相应提高$\frac{1}{m^+}$，对应标记点为$(x, y+\frac{1}{m^+})$，若下一个加入到正例中的样本真实类别是反例，则多了一个FP，假反利率相应提高$\frac{1}{m^-}$，则对应标记点为$(x+\frac{1}{m^-}, y)$，然后用线段连接相应的点即可得到ROC曲线。（**在这里我强烈建议读者能自己在草稿纸上尝试着画一下，以便后续的理解**）



图中蓝色的线条就是我们得到的ROC曲线，线上的每个点对应的k的一个取值下的TPR和FPR，曲线从$(0,0)$开始（$k=1$，所有样本都被标记为反例，TPR和FPR都是0），到$(1,1)$结束（$k=0$，所有样本都被标记为正例，TPR和FPR都是1），图中红色的线条表示通过瞎猜分类得到的ROC曲线。

AUC

AUC是ROC曲线与x轴围成的面积，越大说明分类器的效果越好，还是看上图，通常情况下，ROC曲线是覆盖红线的，因为一般机器学习得到的算法分类效果总会比瞎猜要好（好低的标准），蓝线是我随机生成的，在红线下面也情有可原。
AUC定义为ROC曲线与x轴围成的面积，计算公式为

$$AUC = 1 - \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(W\left(f(x^+)<f(x^-)\right)+\frac{1}{2}W\left(f(x^+)=f(x^-)\right)\right)$$

这是周志华老师《机器学习》上的公式，其中D^+为所有正例组成的集合，x^+是其中的一个正例，D^-为所有反例组成的集合，x^-是其中的一个反例，$f(x)$是模型对样本x的预测结果，在0-1之间，$W(x)$仅在x为真时取1，否则取0。
右边被减数表示任取一对正反例，正例的预测值小于反例的预测值的正反例对数，以及满足正例的预测值等于反例预测值的正反例对数的一半。我觉得这个公式和AUC的几何意义相差太大，因此，要想办法从这个公式中找到AUC的几何意义。
实际上，

$$\begin{aligned}&\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(W\left(f(x^+)<f(x^-)\right)+W\left(f(x^+)=f(x^-)\right)+W\left(f(x^+)>f(x^-)\right)\right)\\&=\sum_{x^+\in D^+}\sum_{x^-\in D^-}1=m^+m^-\end{aligned}$$

第一个等号成立是因为三个W有且仅有一个为1（不是大于就是小于就是等于，只有其一满足）
则

$$AUC=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(W\left(f(x^+)>f(x^-)\right)+\frac{1}{2}W\left(f(x^+)=f(x^-)\right)\right)$$

问题简化

为了方便我们建立公式定义和图像定义之间对应的**Intuition**，我们先假设不存在正反例对使得

$$f(x^+)=f(x^-)$$

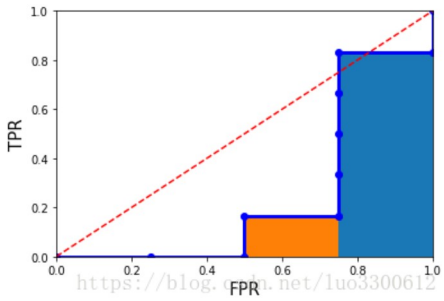
成立，从而

$$AUC = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\Big(W\big(f(x^+) > f(x^-)\big)\Big)$$

现在我们要结合之前画ROC曲线的步骤，令

$$g(t) = \text{当我们将第}t\text{个反例加入到判定为正例的集合中时，之前一共加入的样例数}$$

这个定义稍先复杂，但是其实含义十分简单，比如我依次将正例、反例、反例、正例.....加入到初始为空的正例集合中时，先加入正例，再加入反例，注意这是我们加入的第一个反例，但是却是我们加入的第二个样例，因此$g(1) = 2$，继续下去，加入反例，注意这是我们加入的第二个反例，却是我们加入的得3个样例，因此$g(2) = 3$，再加入正例.....为什么要定义这样的一个奇怪的函数？再次看到我们之前画的ROC图，



为了求ROC曲线和x轴围城的面积，我们将目标区域沿着x轴划分为一个个宽为$\frac{1}{m^-}$的矩形，注意到$\frac{1}{m^-}$是我们在作图中在得到的下一个样例是反例时，横坐标x向右移动的步长，划分后的矩形如图所示，因为TPR在$[0, 0.5]$上恒为0，因此这部分的矩形的面积是0，这个时候我们确定了矩形的宽，**那么矩形的长呢?**

通过思考矩形的生成过程，我们可以知道，横坐标x每向右移动一次会生成一个矩形，而横坐标x向右移动$\frac{1}{m^-}$，当且仅当我们下一个样例是反例，因此，第t个矩形是我们取到的第t个反例时，横坐标向右移动形成的。此时，根据g的定义，我们恰好取到了$g(t) - t$个正样例，因此对应的纵坐标就是$\frac{g(t)-t}{m^+}$，从而ROC与x轴围城的面积是

$$AUC = \sum_{i=1}^{m^-}\frac{1}{m^-}\frac{g(t)-t}{m^+} = \frac{1}{m^+m^-}\sum_{i=1}^{m^-}(g(t)-t)$$

因为这一共有$g(t) - t$个正例在第t个反例之前取到，因此这$g(t) - t$个正例的预测值均大于第t个反例的预测值，也即

$$g(t) - t = \sum_{x^+\in D^+}\Big(W\big(f(x^+) > f(x_t)\big)\Big)$$

代入，得

$$AUC = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{t=1}^{m^-}\Big(W\big(f(x^+) > f(x_t)\big)\Big)$$

用面积定义计算出的结果和之前的公式一致。

原问题

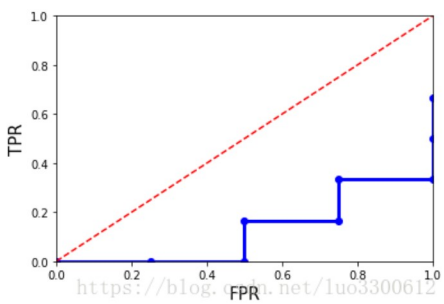
现在问题来了，那个二分之一的项究竟代表什么呢？为什么要乘二分之一，而不把它和某个不等式合并成小于等于或者大于等于呢?
再看AUC的公式

$$AUC = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\Big(W\big(f(x^+) > f(x^-)\big) + \frac{1}{2}W\big(f(x^+) = f(x^-)\big)\Big)$$

当存在x^+和x^-满足$f(x^+) = f(x^-)$时，回到我们画ROC图的步骤，记得在画图之前，我们需要对样例进行排序吗，如果存在$f(x^+) = f(x^-)$，我们该把哪个排在前面呢?
我们来考虑一个新的样本集合，其中存在一对(x^+, x^-)，满足$f(x^+) = f(x^-)$，一种排序方法如下

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
预测结果	0.94	0.90	0.83	0.73	0.50	0.50	0.47	0.27	0.07	0.06
真实类别	0	0	1	0	1	0	1	1	0	1

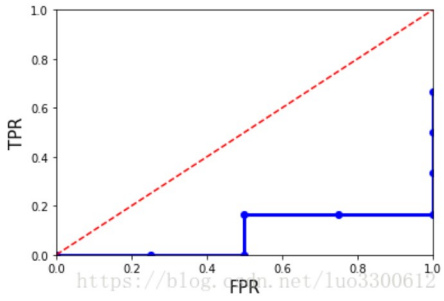
也就是我们把正例排在反例前面，结果得到的ROC图如下，



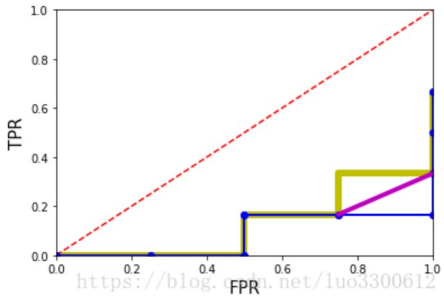
如果我们把反例排在正例的前面，结果如下

样例	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
预测结果	0.94	0.90	0.83	0.73	0.50	0.50	0.47	0.27	0.07	0.06
真实类别	0	0	1	0	0	1	1	1	0	1

画出的ROC曲线如下



显然，前者的AUC比后者的AUC大，这是容易理解的，如果我们把正例放在前面，不断加入样例的过程中，因为先遇到正例，曲线先向上，再向右，反之，把反例放在前面，因为先遇到反例，曲线先向右，再向上，因此这会导致AUC的不同，于是，当碰到这种情况时，AUC曲线取折中的方法，如图



AUC会在产生分歧的位置计算紫色线段下的梯形面积而非任何一个矩形面积，对于某个定值，有不止一对正负样本的预测值等于该定值时，也是采用这种策略。推导过程类似于之前的过程，有兴趣的读者可以推导一下，一种思路是假设ROC在绘画面临向右和向上的抉择时，永远选择前者（即将所有反例排在正例前面），然后证明AUC公式的第二项等同于缺失的部分面积即可。

---

1. 本文主要参考周志华老师的《机器学习》. ↩