笔记丨统计评估指标AUC 详解

素质云笔记 3月3日

---

1 AUC的两种解读视角：

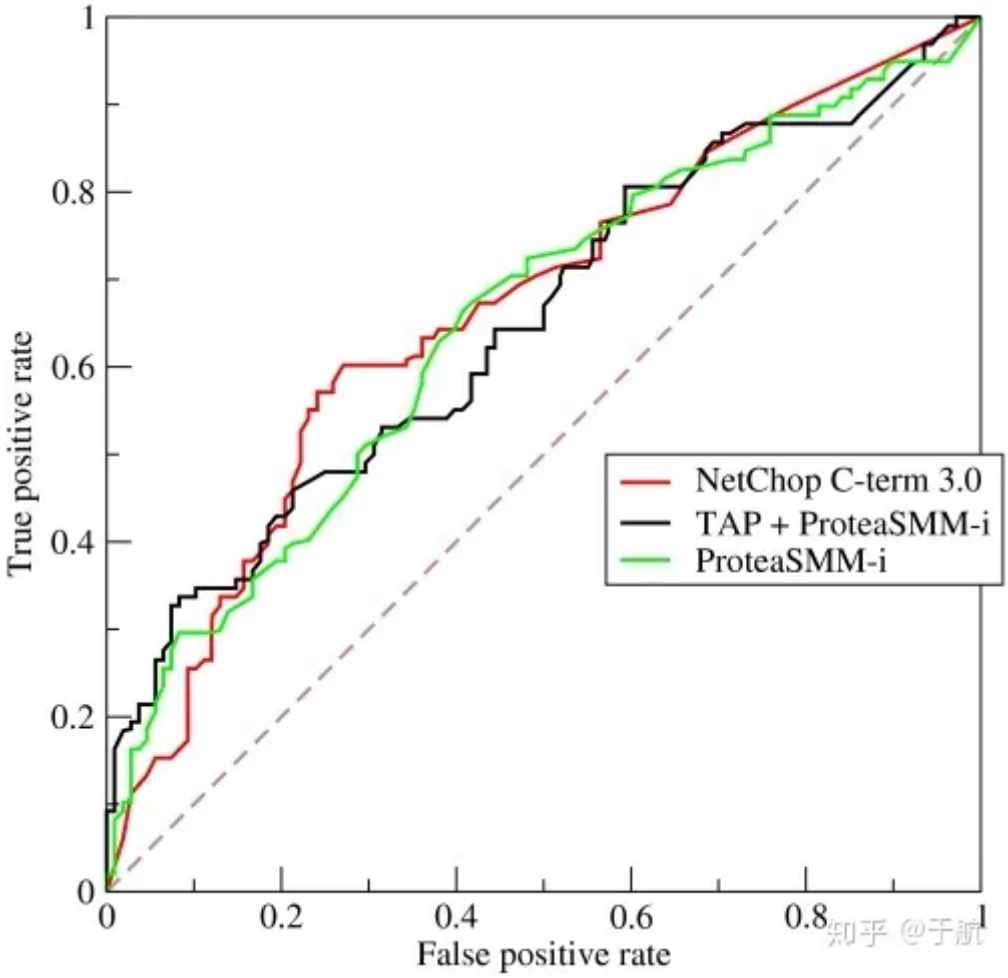
1.1 ROC曲线与坐标轴形成面积

AUC 的全称是 AreaUnderRoc 即 Roc 曲线与坐标轴形成的面积，取值范围 [0, 1].

Roc (Receiver operating characteristic) 曲线是一种二元分类模型分类效果的分析工具。首先需要知道如下定义：

- TPR: 在所有实际为阳性的样本中，被正确地判断为阳性之比率TPR = TP/P = TP/(TP+FN)
- FPR: 在所有实际为阴性的样本中，被错误地判定为阳性之比率 FPR = FP/N = FP/(FP + TN)

给定一个二元分类模型和它的阈值，就能从所有样本的（阳性 / 阴性）真实值和预测值计算出一个 (X=FPR, Y=TPR) 座标点。



从 (0, 0) 到 (1,1) 的对角线将ROC空间划分为左上 / 右下两个区域，在这条线的以上的点代表了一个好的分类结果（胜过随机分类），而在这条线以下的点代表了差的分类结果（劣于随机分类）。

1.2 古典概率模型——求导AUC

文章【最浅显易懂的图解AUC和GAUC】有提及：

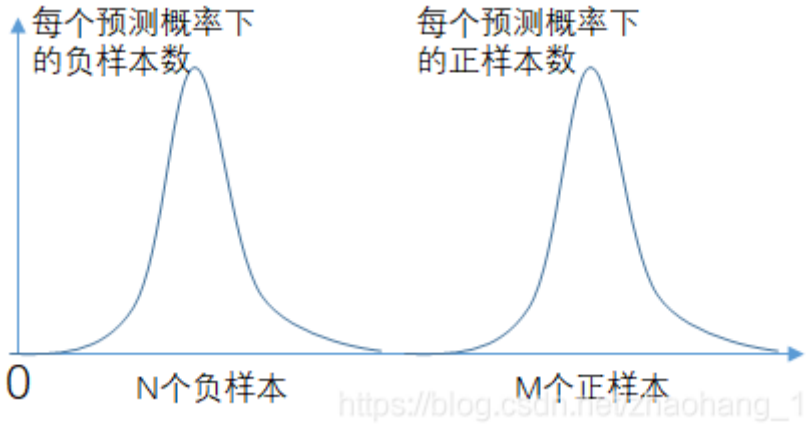
> 另一种定义更常用，分别随机从正负样本集中抽取一个正样本，一个负样本，正样本的预测值大于负样本的概率。

按照定义分别随机从政府样本集中抽取一个正负样本，正样本的预测值大于负样本的概率。

每个预测为正的样本，能比多少个负样本大 积分所在的区域是啥呢？

实际是正样本和负样本的交叉，也即 正样本数*负样本数

这里我们可以设想一种理想状态：

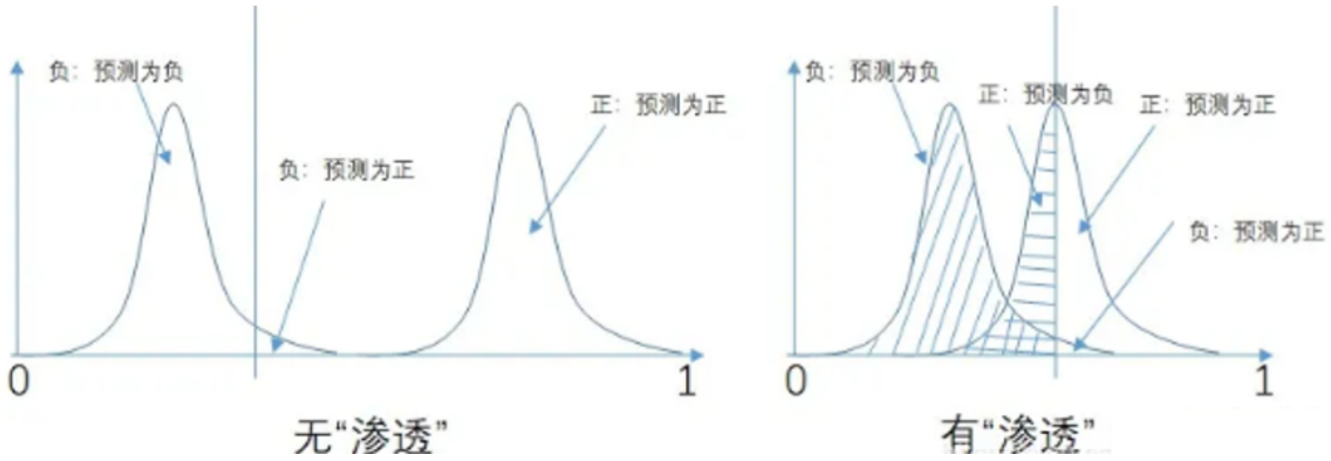


正样本和负样本是两个互不纠缠的正态分布，其中有M个正样本，N个负样本。

阈值如果遍历所有正样本，则每个正样本都比N个负样本大，因此，积分下来，就是N+N+N+...+N = M * N，而积分的区域是M * N，因此，这种理想状态下，得到了AUC为

$$AUC = \frac{\sum_1^N N}{M * N} = 1$$

因此这里我们可用通过这种方式重新计算第一节中的AUC



第一节中，原始有五个正样本：

- p=0.9的真实正样本，它在所有5个负样本前面，因此记为5
- p=0.8的真实正样本，它在所有5个负样本前面，因此记为5
- p=0.7的真实正样本，它在所有5个负样本前面，因此记为5
- p=0.6的真实正样本，它在4个负样本前面，因此记为4
- p=0.4的真实正样本，它在3个负样本前面，因此记为3

交叉区域记为5*5=25

因此最终的AUC记为

$$AUC = \frac{5+5+5+4+3}{5*5} = 0.88$$

此时的一个通俗易懂的解读：

> 例如0.7的AUC，其含义可以大概理解为：给定一个正样本和一个负样本，在70%的情况下，模型对正样本的打分高于对负样本的打分。
> 可以看出在这个解释下，我们关心的只有正负样本之间的分数高低，而具体的分值则无关紧要。

---

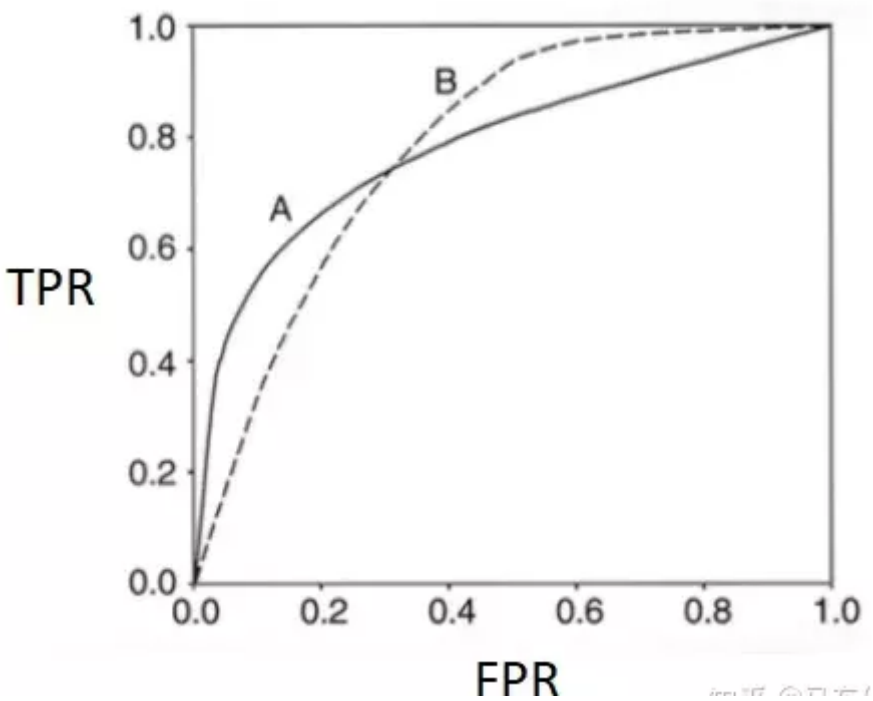
2 AUC的特性与优劣

（1）AUC指标本身和模型预测score绝对值无关，只关注排序效果，因此特别适合排序业务。

当然也会带来问题：

- AUC 反应了太过笼统的信息。无法反应召回率、精确率等在实际业务中经常关心的指标。

- 还有，AUC 的 misleading 的问题：



modelA 和 modelB 的 ROC 曲线下面积 AUC 是相等的，但是两个模型在不同区域的预测能力是不相同的，所以我们不能单纯根据 AUC 的大小来判断模型的好坏。

（2）AUC对分值本身不敏感，故常见的正负样本采样，并不会导致auc的变化。

比如在点击率预估中，处于计算资源的考虑，有时候会对负样本做负采样，但由于采样完后并不影响正负样本的顺序分布。

即假设采样是随机的，采样完成后，给定一条正样本，模型预测为score1，由于采样随机，则大于score1的负样本和小于score1的负样本的比例不会发生变化。
但如果采样不是均匀的，比如采用word2vec的negative sample，其负样本更偏向于从热门样本中采样，则会发现auc值发生剧烈变化。

（3）auc非常适合评价样本不平衡中的分类器性能

---

3 AUC多大才算好？

多高的AUC才算高？

首先AUC = 1可能吗？

难度非常大，样本数据中本身就会存在大量的歧义样本，即特征集合完全一致，但label却不同。

当我们拿到样本数据时，第一步应该看看有多少样本是特征重复，但label不同，这部分的比率越大，代表其“必须犯的错误”越多。学术上称它们为Bayes Error Rate，也可以从不可优化的角度去理解。

影响Max AUC的主要因素：样本的不确定性。
AUC的高低只和特征取值的多样性有关。
其实不只是Max AUC，在真实问题中使用更多好的特征也会提高AUC（起码是训练集AUC），本质上是因为更多的特征可以组合出更多的特征取值来，从而提高特征的区分能力。

---

4 线上、线下AUC差异较大成因分析

4.1 业务场景使用AUC：点击模型与购买模型的差异

> AUC大小 => 点击模型 < 购买转化率模型 => 购买model 更好预测
>
> 线下、线上差异 = > 点击模型< 购买转化率模型 => 购买model 线上与线下差异较大

我们在实际业务中，常常会发现点击率模型的auc要低于购买转化率模型的auc。

正如前文所提，AUC代表模型预估样本之间的排序关系，即正负样本之间预测的gap越大，auc越大。

通常，点击行为的成本要低于购买行为，从业务上理解，点击率模型中正负样本的差别要小于购买力模型，

即购买转化模型的正样本通常更容易被预测准。

AUC毕竟是线下离线评估指标，与线上真实业务指标有差别。差别越小则AUC的参考性越高。

比如上文提到的点击率模型和购买转化率模型，虽然购买转化率模型的AUC会高于点击率模型，但往往都是点击率模型更容易做，线上效果更好。

购买决策比点击决策过程长、成本重，且用户购买决策受很多场外因素影响，比如预算不够、在别的平台找到更便宜的了、知乎上看了评测觉得不好等等原因，这部分信息无法收集到，导致最终样本包含的信息缺少较大，模型的离线AUC与线上业务指标差异变大。

总结起来，样本数据包含的信息越接近线上，则离线指标与线上指标gap越小。而决策链路越长，信息丢失就越多，则更难做到线下线上一致。

4.2 线上、线下AUC有差异

> 情况一：新模型比老模型auc高

代表新模型对正负样本的排序能力比老模型好。
理论上，这个时候上线abtest，应该能看到ctr之类的线上指标增长。

实际上经常会发生不一致，首先，我们得排除一些低级错误：

- 排除bug，线上线下模型predict的结果要符合预期。
- 谨防样本穿越。比如样本中有时间序类的特征，但train、test的数据切分没有考虑时间因子，则容易造成穿越。

> 情况二：离线auc涨了不少，上线一看效果ctr和cpm反而下降

几种可能的原因和解决办法:

1. 特征/数据出现穿越

一般就是使用了和label强相关的特征导致的数据泄漏。这种问题一般相对好查，很多时候在离线阶段就能发现。

明显的表现就是训练集和测试集差异比较大

2. 线上线下特征不一致

据我所知，这种情况是导致离线涨在线跌或者没效果的最常见情况。
首先是代码不一致，例如，离线对用户特征的加工处理采用scala/python处理，抽取用户最近的50个行为，在线特征抽取用c++实现只用了30个。

只要离线和在线用不同的代码抽取就很容易存在这种代码带来的不一致。
另外一种线上线下不一致，是由于数据的不一致导致。

这在离线拼接样本和特征的pipeline中比较常见。

一般离线特征都是按照天处理的，考虑各种数据pipeline的流程，处理时间一般都会有延迟，离线特征处理完之后导到线上供线上模型预估时请求使用。

要严格保证线上线下的特征一致性，最根本的方法就是同一套代码和数据源抽取特征，业内目前通用的方法就是，在线实时请求打分的时候落地实时特征，训练的时候就没有特征拼接的流程，只需要关联label，生成正负样本即可

3. 数据分布的不一致

如果仔细排查，既不存在数据泄漏，也没有出现不一致的问题，
离线auc明明就是涨了很多，线上就是下降，而且是离线涨的越多，线上下降越多，还有一种可能就是数据的不一致，也就是数据的“冰山效应”

----离线训练用的是有偏的冰山上的数据，而在线上预估的时候，需要预测的是整个冰山的数据，包括大量冰面以下的数据！

这里给下两个在我们这还比较有效的经验：

（1）对无偏数据进行上采样

这里的无偏是相对的，可以是随机/探索流量产生的样本，也可以是新模型产生的样本。大概意思，就是尽可能利用这些对新模型有利的样本

（2）线上线下模型融合

比较trick的方法，没有太多方法论，但是确实能work。

5 AUC逻辑升级 - GAUC

AUC作为排序的评价指标本身具有一定的局限性，它衡量的是整体样本间的排序能力，对于计算广告领域来说，它衡量的是不同用户对不同广告之间的排序能力，而线上环境往往需要关注同一个用户的不同广告之间的排序能力。

线上会出现新样本，在线下没有见过，造成AUC不足。
这部分更多是采用online learning的方式去缓解，AUC本身可改进的不多。

线上的排序发生在一个用户的session下，而线下计算全集AUC，即把user1点击的正样本排序高于user2未点击的负样本是没有实际意义的，但线下auc计算的时候考虑了它。

所以，阿里在 Deep Interest Network中提到一种改进版本的 AUC 指标，用户加权平均AUC（gAUC）更能反映线上真实环境的排序能力。

$$gAUC = \frac{\sum_{i=1}^{n} impression_i * AUC_i}{\sum_{i=1}^{n} impression_i}$$

即以user为group，在对user的impression做加权平均。私以为，只是对用户做group还不够，应该是基于session去做group。

参考文献

1 AUC的理解与计算
2 最浅显易懂的图解AUC和GAUC
3 为什么搜索与推荐场景用AUC评价模型好坏？
4 如何理解机器学习和统计中的AUC？
5 多高的AUC才算高？
6 线下AUC提升为什么不能带来线上效果提升？--测试和评估的一些真相
7 精确率、召回率、F1 值、ROC、AUC 各自的优缺点是什么？
8 如何解决离线和线上auc和线上点击率不一致的问题？