模型效果的评估方法

leolinuxer

July 20, 2020

Contents

1 评估目标的设定原则 [3]

以 YouTube 的推荐系统为例，推荐系统的终极优化目标应该包括两个维度：一个维度是用户体验的优化，另一个维度是满足公司的商业利益。对于 YouTube 公司而言，其优化用户体验结果的最直接体现就是用户观看时长的增加。而 YouTube 作为一家以广告位主要收入来源的公司，其商业利益也建立在用户观看时长的增长之上，因为总用户观看时长与广告的总曝光机会成正比。

所以，YouTube 推荐系统的优化目标就是用户观看时长，而不是传统系统看中的“点击率”。其大致推荐流程是：先通过构建深度学习模型，预测用户观看某候选视频的时长，再按照预测时长进行候选视频的排序，行成最终的推荐列表。

2 Confusion Matrix

Confusion Matrix 矩阵如下表所示：

预测值-实际值	True	False
True	True Positive(真阳性)	False Positive(假阳性)
False	False Negative(假阴性)	True Negative(真阴性)

Table 1: Confusion Matrix

3 各种率的定义

正确率（Precision）：

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

真阳性率 (True Positive Rate，TPR)，灵敏度 (Sensitivity)，召回率 (Recall):

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

真阴性率 (True Negative Rate，TNR)，特异度 (Specificity)：

$$Specificity = Recall = \frac{TN}{FP + TN}$$

假阴性率 (False Negatice Rate，FNR)，漏诊率 (= 1 - 灵敏度)：

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

假阳性率 (False Positice Rate，FPR)，误诊率 (= 1 - 特异度)：

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

4 主要评价指标 [1]

4.1 ROC

对于分类器，或者说分类算法，评价指标主要有 precision，recall，F-score 等，以及这里要讨论的 ROC 和 AUC。

ROC 曲线：接收者操作特征曲线（receiver operating characteristic curve），是反映敏感性和特异性连续变量的综合指标，ROC 曲线上每个点反映着对同一信号刺激的感受性。下图是一个 ROC 曲线的示例：

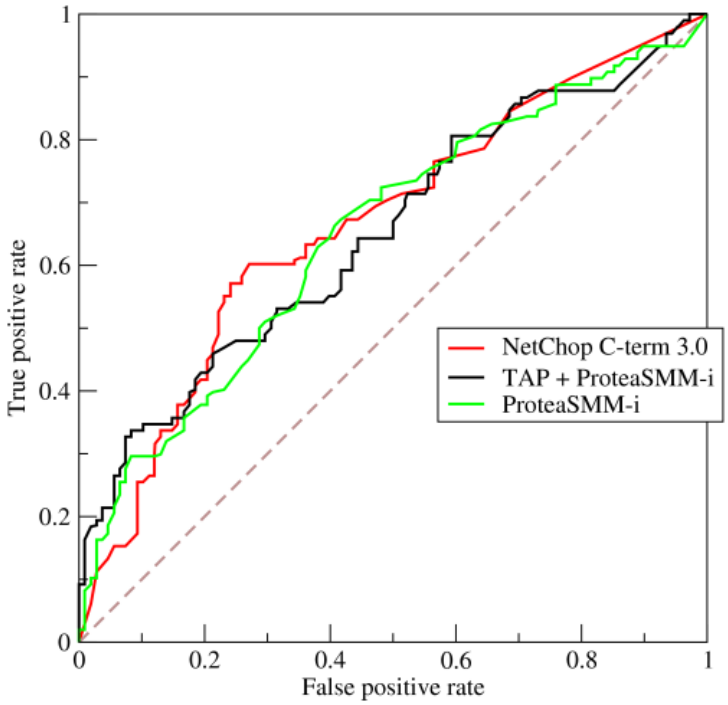

Figure 1: ROC 曲线示意

ROC 曲线的横纵坐标分别为：

横坐标：1-Specificity，伪正类率 (False positive rate，FPR)，预测为正但实际为负的样本占所有负例样本的比例 (负例中预测错了的比例)；

纵坐标：Sensitivity，真正类率 (True positive rate，TPR)，预测为正且实际为正的样本占所有正例样本的比例 (正例中预测对了的比例)。

在一个二分类模型中，假设采用逻辑回归分类器，其给出针对每个实例为正类的概率，那么通过设定一个阈值如 0.6，概率大于等于 0.6 的为正类，小于 0.6 的为负类。对应的就可以算出一组 (FPR,TPR)，在平面中得到对应坐标点。随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样

也掺杂着真正的负实例，即 TPR 和 FPR 会同时增大。阈值最大时，对应坐标点为 (0,0)，阈值最小时，对应坐标点 (1,1)。

4.2 AUC(Area Under Curve)

AUC (Area Under Curve) 被定义为 ROC 曲线下的面积，显然这个面积的数值不会大于 1。又由于 ROC 曲线一般都处于 y=x 这条直线的上方（如果不是，那么可以交换阈值上下对应的分类，即可得到更好的分类结果），所以 AUC 的取值范围一般在 0.5 和 1 之间。使用 AUC 值作为评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。

4.3 为什么使用 ROC 曲线

既然已经这么多评价标准 (如 precision-recall 等)，为什么还要使用 ROC 和 AUC 呢? 因为 ROC 曲线有个很好的特性：当测试集中的正负样本的分布变化的时候，ROC 曲线能够保持不变。在实际的数据集中经常会出现类不平衡（class imbalance）现象，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间变化。下图是 ROC 曲线和 Precision-Recall 曲线的对比：

在上图中，(a) 和 (c) 为 ROC 曲线，(b) 和 (d) 为 Precision-Recall 曲线。(a) 和 (b) 展示的是分类其在原始测试集（正负样本分布平衡）的结果，(c) 和 (d) 是将测试集中负样本的数量增加到原来的 10 倍后，分类器的结果。可以明显的看出，ROC 曲线基本保持原貌，而 Precision-Recall 曲线则变化较大。

4.4 平均精度均值（mAP, Mean Average Precision）

平均精度均值是对平均精度（Average Precision）的再次平均。

要计算 mAP 必须先绘出各类别 PR 曲线，计算出 AP。假定推荐系统对某一用户测试集的排序结果如下表所示：

推荐序列	N = 1	N= 2	N=3	N= 4	N= 5	N= 6
真实标签	1	0	0	1	1	1

其中，1 代表正样本，0 代表负样本。

在排序模型中，通常没有一个确定的阈值把预测结果直接判定为正样本或负样本，而是采用 TopN 排序结果的精确率（precision@N）和召回率 (recall@N) 来衡量排序模型的性能，即认为模型排序的 TopN 的结果就是模型判定的正样本，然后计算 precision@N 和 recall@N。

接下来，计算上述序列中每个位置上的 precision@N：

AP 的计算只取正样本处的 precision 进行平均，即 AP = (1/1 + 2/4 + 3/5 + 4/6) = 0.6917，那么什么是 mAP 呢?

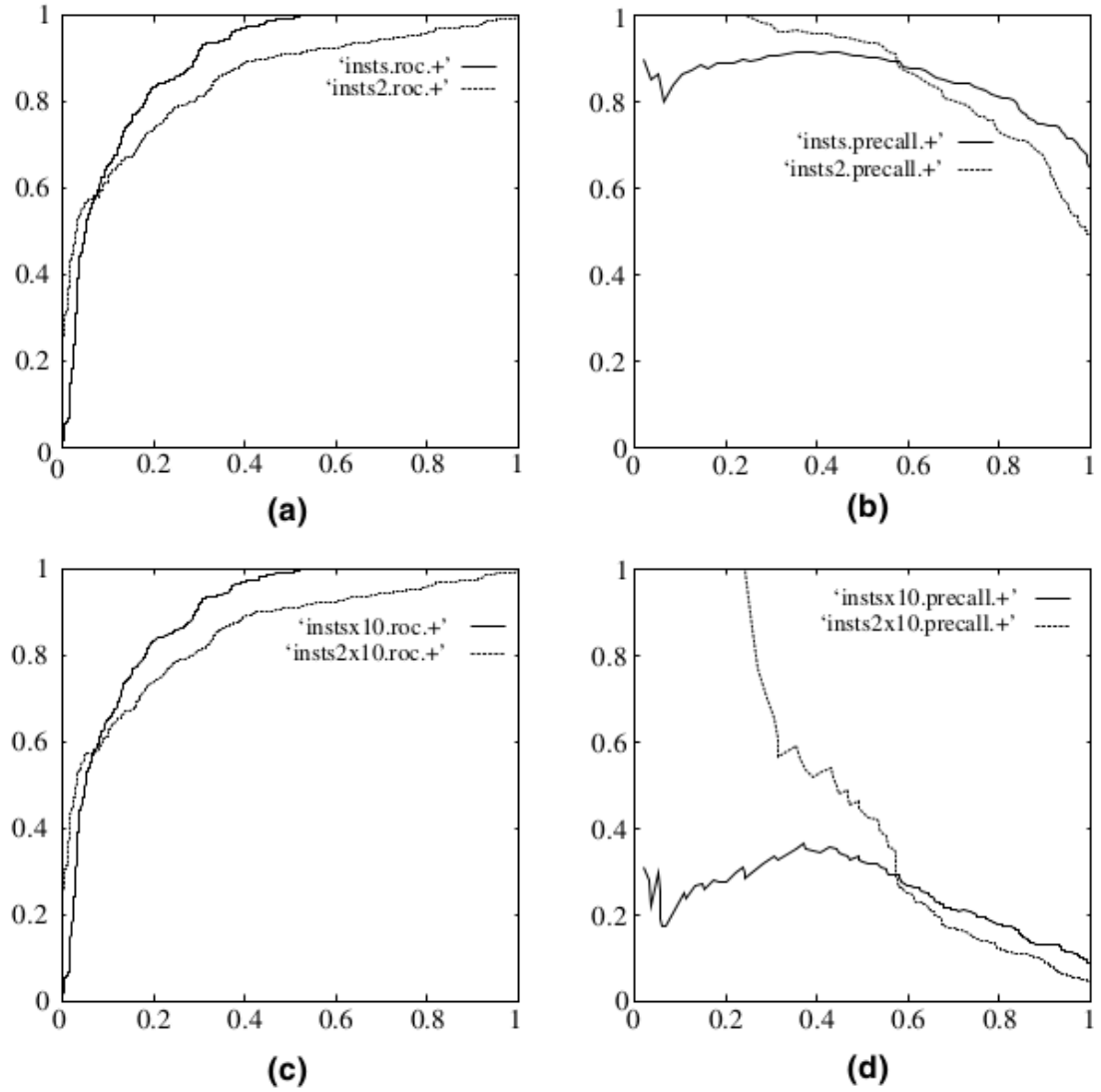

Figure 2: ROC vs Precision-Recall

推荐序列	N = 1	N= 2	N=3	N= 4	N= 5	N= 6
真实标签	1	0	0	1	1	1
precision@N	1/1	1/2	1/3	2/4	3/5	4/6

如果推荐系统对测试集中的每个用户都进行样本排序，那么每个用户都会计算出一个 AP 值，然后再对所有用户的 AP 值进行平均，就得到了 mAP。

值得注意的是，mAP 的计算方法和 P-R 曲线、ROC 曲线的计算方法完全不同，因为 mAP 需要对每个用户的样本进行分用户排序，而 P-R 曲线和 ROC 曲线均是对全量测试样本进行排序。

4.5 精确率、准召率、F1 值各自的优缺点 [2]

4.5.1 精确率 Accuracy

Accuracy 是最常见也是最基本的 evaluation metric。但在 binary classification 且正反例不平衡的情况下，尤其是我们对 minority class 更感兴趣的时候，accuracy 评价基本没有参考价值。什么 fraud detection（欺诈检测），癌症检测，都符合这种情况。举个栗子：在测试集里，有 100 个 sample，99 个

反例，只有 1 个正例。如果我的模型不分青红皂白对任意一个 sample 都预测是反例，那么我的模型的 accuracy 是正确的个数 / 总个数 = 99/100 = 99%。你拿着这个 accuracy 高达 99% 的模型屁颠儿屁颠儿的去预测新 sample 了，而它一个正例都分不出来，有意思么……也有人管这叫 accuracy paradox。

4.5.2 precision 和 recall

准招率是比 Accuracy 更有用的 metric。

recall 是相对真实的答案而言：true positive / golden set 。假设测试集里面有 100 个正例，你的模型能预测覆盖到多少，如果你的模型预测到了 40 个正例，那你的 recall 就是 40%。

precision 是相对你自己的模型预测而言：true positive / retrieved set。假设你的模型一共预测了 100 个正例，而其中 80 个是对的正例，那么你的 precision 就是 80%。我们可以把 precision 也理解为，当你的模型作出一个新的预测时，它的 confidence score 是多少，或者它做的这个预测是对的的可能性是多少。

一般来说，鱼与熊掌不可兼得。如果你的模型很贪婪，想要覆盖更多的 sample，那么它就更有可能犯错。在这种情况下，你会有很高的 recall，但是较低的 precision。如果你的模型很保守，只对它很 sure 的 sample 作出预测，那么你的 precision 会很高，但是 recall 会相对低。

这样一来，我们可以选择只看我们感兴趣的 class，就是 minority class 的 precision，recall 来评价模型的好坏。

4.5.3 F1-score

F1-score 就是一个综合考虑 precision 和 recall 的 metric：

$$
\begin{aligned}
F1 - score &= \frac{2}{1/precision + 1/recall} \\
&= 2 * precision * recall/(precision + recall)
\end{aligned}
$$

可以看出，F1-score 是 precision 和 recall 的调和平均数（调和平均数（harmonic mean）又称倒数平均数，是总体各统计变量倒数的算术平均数的倒数。$H_n = n/\sum_{i=1}^{n}\frac{1}{x_i} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$）

如果两个模型，一个 precision 特别高，recall 特别低，另一个 recall 特别高，precision 特别低的时候，F1-score 可能是差不多的，也不能基于此来作出选择。

5 离线评估的主要方法 [3]

5.1 Holdout 检验

将原始样本随机划分为测试集和验证集两部分；比如，可以将样本按照 70%-30% 的比例随机分成两部分，70% 的样本用于模型的训练，30% 的样本用于模型的评估

Holdout 检验的缺点：在验证集上计算出来的评估指标与训练集和验证集的划分有很大关系，如果仅仅进行少量 holdout 检验，则得到的结论存在较大的随机性。为了消除这种随机性，“交叉检验” 的思想被提出。

5.2 交叉检验

5.2.1 k-fold 交叉检验

先将全部样本划分为 k 个大小相等的样本子集；依次遍历这 k 个子集，每次都把当前子集作为验证集，其余所有子集作为训练集，进行模型的训练和评估；最后将所有 k 次的评估指标的平均值作为最终的评估指标。在实际实验中，k 经常取 10。

5.2.2 留一验证

每次留下一个样本作为验证集，其余所有样本作为训练集。样本总数为 n，依次遍历所有 n 个样本，进行 n 次验证，再讲评估指标求平均得到最终各指标。在样本总数较多的情况下，留一验证法的时间开销极大。事实上，留一验证是留 p 验证的特例。留 p 验证是指每次留下 p 个样本作为验证集，而从 n 个元素中选择 p 个元素有 C_n^p 中可能，因此它的时间开销远远大于留一验证，因此很少在实际工程中应用。

5.2.3 自助法（bootstrap）

不管是 holdout 检验还是交叉检验，都是基于划分样本机和验证集的方式进行模型评估的。然而，当样本规模比较小时，将样本机进行划分会让训练集进一步减小，这可能会影响模型的训练效果。

bootstrap 法能在一定程度上维持训练集样本的规模。是基于自助采样法的验证方法：对于总数为 n 的样本集合，进行 n 次有放回的随机抽样，得到大小为 n 的训练集。在 n 次采样过程中，有的样本会被重复采样，有的样本没有被抽出过，将这些没有被抽出的样本作为验证集进行模型验证，就是自助法的验证过程。

References

[1] “机器学习之分类性能度量指标: Roc 曲线、auc 值、正确率、召回率.” [Online]. Available: https://www.jianshu.com/p/c61ae11cc5f6

[2] “精确率、召回率、f1 值、roc、auc 各自的优缺点是什么.” [Online]. Available: https://www.zhihu.com/question/30643044/answer/224360465

[3] 王🄵, 深度学习推荐系统.