评估分类模型

分类模型相对是比较好评估的,一般来说我们在训练时会将训练数据分为 训练集,验证集和测试集 三份或者 训练集,测试集 两份.当模型训练完成,基于模型在测试集上的预测结果和真实结果,计算一些指标来评估模型的泛化能力.这些指标主要包括准确度(accuracy),精确度(precision),召回率(recall),F1,AUC.理论上来看,指标越大,模型的泛化能力越强.
这些指标按照内在逻辑可以分为两种

1. 基于样例的.这类指标基于测试集每个样例的预测结果和真实结果,每个样例的对比结果是独立的.包括准确度,精确度,召回率和F1
2. 基于排序的.考虑到大多数分类模型的输出是属于各个类的概率分布,比如线性回归和逻辑回归.我们可以对比测试集整体按照概率排序的预测结果和真实情况来作为对模型的评价.比如AUC,NDCG和Rank correlation.

基于样例的指标

在介绍指标之前,不妨假设这是二分类问题,类别有正/负两种(或者1/0两种),先给出以下四种定义,用来划分测试集:

- 预测为正且实际为正的 叫做TP(True Positive).
- 预测为正且实际为负的 叫做FP(False Positive).
- 预测为负且实际为正的 叫做FN(False Negative).
- 预测为负且实际为负的 叫做TN(True Negative).

测试集中的样本一定会符合以上四种定义之一.

横轴预测,纵轴真值	预测为正	预测为负	不论预测
实际为正	TP	FN	P
实际为负	FP	TN	N

如果没有特殊说明,这六种符号均表示符合对应定义的样本的数量.

准确度(accuracy)

准确度反映的是预测值与真值一致的程度.其分子为预测对的样本的数量,分母为测试集总的样本数量.

$$Accurcy = \frac{TP + TN}{TP + FP + TN + FN}$$

精确度(precision)

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

准确度是预测为正的样本中真值为正的百分比.

召回率(recall)

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

召回率描述的是真值为正的样本中被预测为正的比例.也叫TPR(True Positive Rate).

F1_score

我们不妨想象,一个分类器把所有样本都预测为正,则召回率很高;如果它只把样本中最有可能为正的一个预测为正,其他都为负,则准确度很高.所以精确度和召回率都不能独自完全描述一个模型的准确程度.而F1是准确度和召回率的调和平均,是综合准确度和召回率的一种指标,反映的是预测对的正样本和预测错的样本数的比例.

$$F1 = \frac{1}{\frac{\frac{1}{precision} + \frac{1}{recall}}{2}} = 2 * \frac{precision * recall}{precision + recall} = \frac{2 * TP}{2 * TP + FN + FP}$$

混淆矩阵(confusion matrix)

混淆矩阵是更加细致的一个模型效果评估工具,它的纵轴为真实标签,横轴为预测标签,其中的数字则是命中的个数.通过这个可以比较清晰地看出模型预测的偏向性,以此可以作为模型调整的依据.

比如我们有如下的混淆矩阵,就可以看出把真值非1的样本预测为1的比率很高,背后原因可能是因为训练集中1类的样本很多.

横轴预测,纵轴真值	预测为1	预测为2	预测为3
实际为1	5	2	0
实际为2	5	3	0
实际为3	10	0	2

基于排序的指标

AUC(Area Under Curve)

除了precision, recall, F1等基于样例的评价指标,还有AUC(Area Under Curve)这种关于样本总体顺序的指标.
AUC是ROC曲线 (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)和y=0直线和x=1直线围的面积,换句话说就是ROC曲线下的面积.

ROC(Receiver Operating Characteristic curve)

ROC是输出概率分布的二分类器分类能力的一种图形化展示,通过改变阈值(样例的预测概率大于阈值则预测样例为正,否则为负.比如概率大于0.5则预测为正,否则为负,这里0.5就是阈值)画出召回率相对错误正类率的图像.
实际操作中就是

1. 我们从所有样例的预测概率和比最小概率略小的数组成的集合中取阈值,
2. 计算该阈值下的召回率和错误正类率作为y坐标和x坐标画在二维空间上,
   - 当阈值取-0.1时,所有样例都被预测为正,则召回率为1,FPR也为1.
   - 当阈值取1时,所有样例都被预测为负,则召回率为0,FPR也是0.
   - 阈值越小时,样例被错误地预测为正的风险越大,也就是FPR越接近1,召回真值为正的样例越容易,也就是召回率越接近1.
3. 将这些点连起来.连成的这条曲线就是ROC曲线(折线).

错误正类率(False Positive Rate, FPR)

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

FPR反映的是真值为负的样例中有多少被错误地预测为正的.

理解AUC为什么反映的是正负样本相对顺序

为什么说ROC反映的是样本中样例顺序呢?因为ROC的每个点对应一个阈值,而这个阈值对应了按照预测概率升序降序后样本(我们叫做sorted sample)的一种分类方式(排名第几之前的样例全部预测为正).

ROC的绘制

通过sorted sample,**ROC的绘制**可以描述为

1. sorted sample从前往后遍历,每遇到一个真值为负的样例(称为负样例),
2. 该负样例和之前的样例预测为正,之后的预测为负.换句话说,阈值就是该样例后一个样例的概率(或者该样例的概率和前一个样例的概率之间的某个实数作为阈值,实数稠密性保证阈值必定存在.如果该负样例已经是sorted sample最后一个,则用-0.1作为阈值).这时我们在ROC上绘制一个点.
3. 重复1和2,直到最后一个负样例
   这样绘制的点是从左往右的,直到(1,1)点.

AUC计算公式

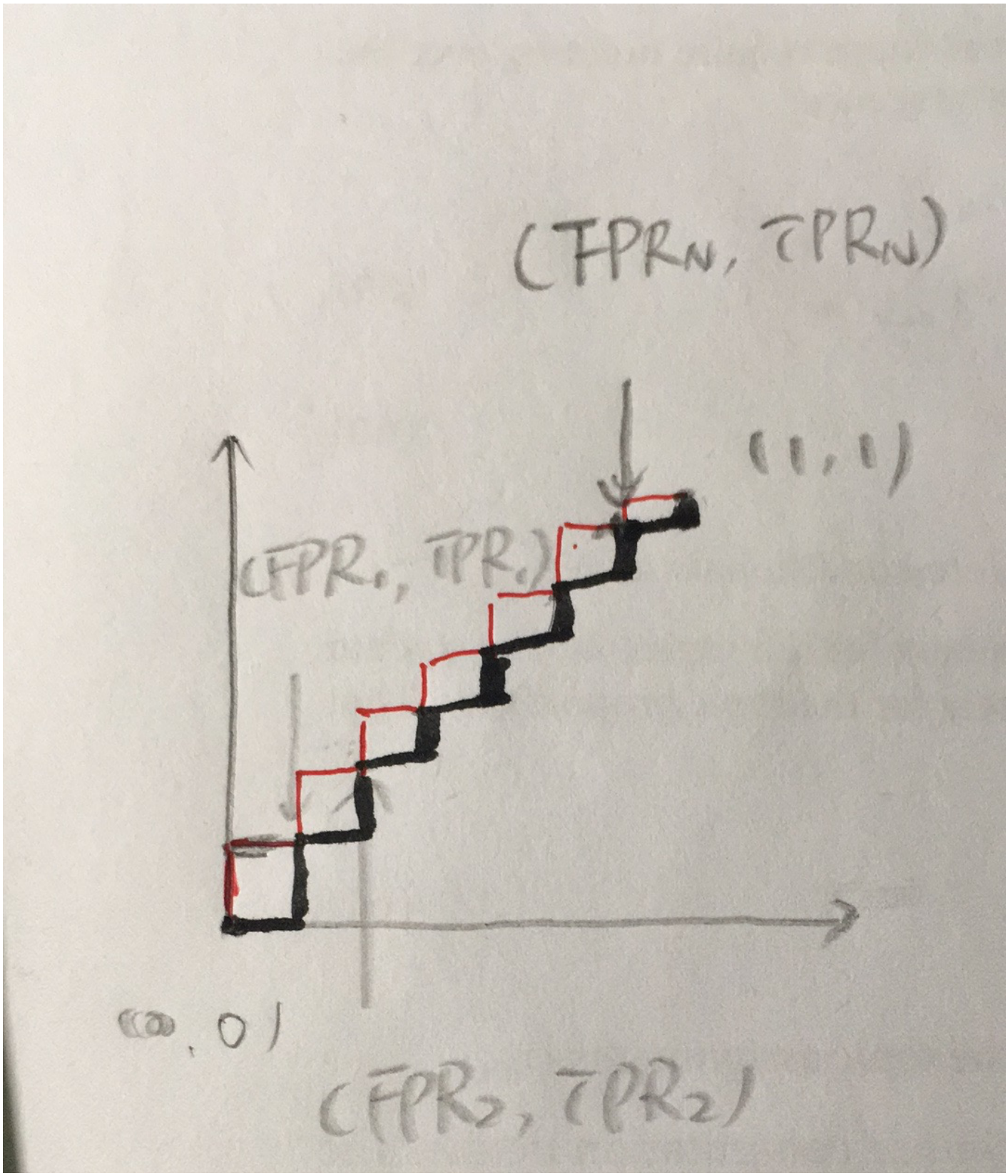
我们记1中sorted sample第i个负样例出现的位置为pos_i,那它之前的正样例数为$pos_i - i$.这样ROC上的点从原点开始,分别是

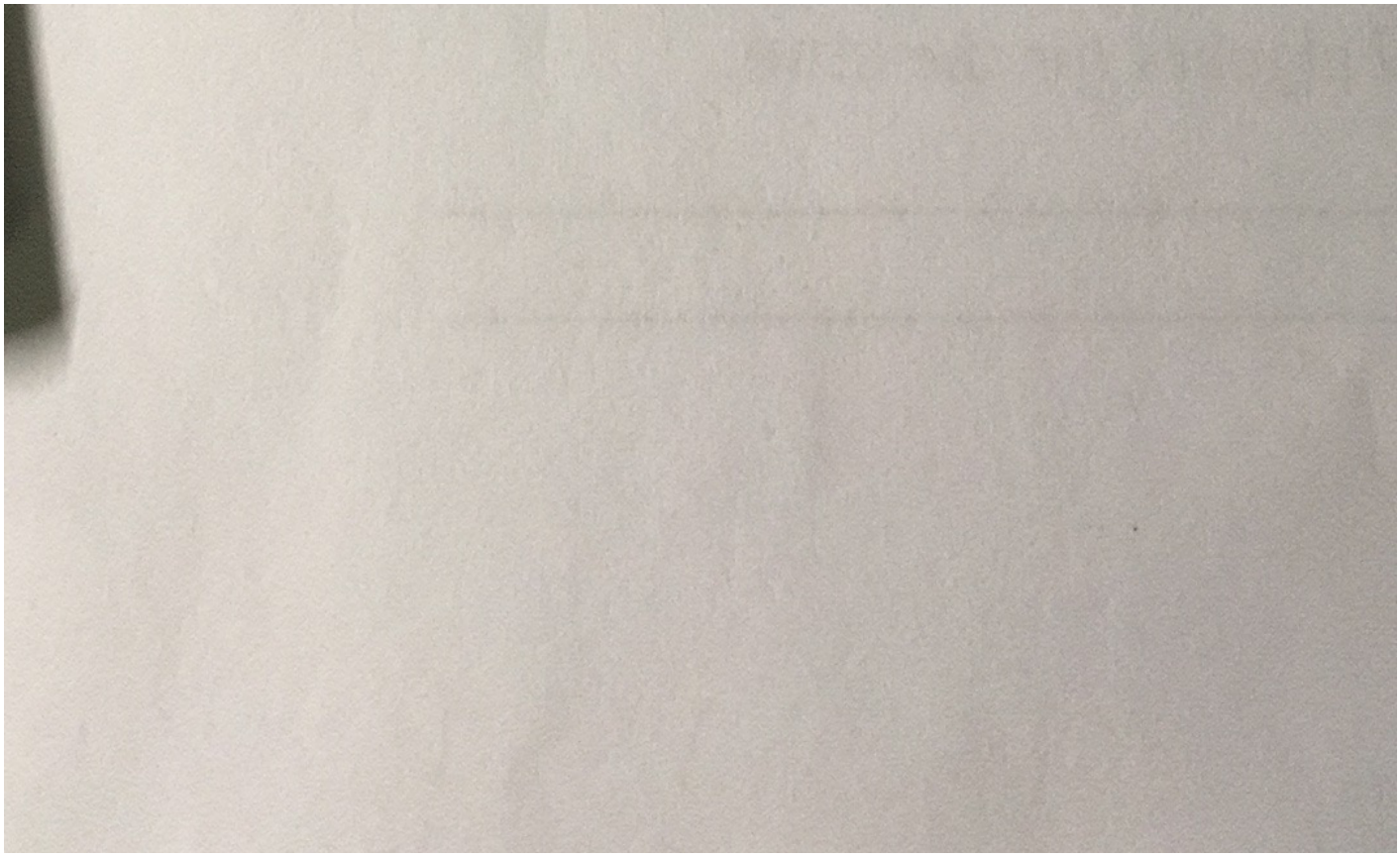
$$(0,0),(FPR_1,TPR_1),(FPR_2,TPR_2),\ldots,(FPR_N,TPR_N),(1,1)$$

等于

$$(0,0),(\frac{1}{N},\frac{pos_1-1}{T}),(\frac{2}{N},\frac{pos_2-2}{T}),\ldots,(\frac{N}{N},\frac{pos_N-N}{T}),(1,1)$$

因为ROC曲线是通过连接这些点绘制的,所以ROC曲线一定介于





中红线和绿线之间.AUC也就介于红线下的面积和绿线下的面积之间. 为了简便,我们不妨设pos_N = N+T,也就是说预测概率最小的那个样例是负样例.这时红线下的面积是多少呢?我们算一下

$$\frac{1}{N}\sum_{i=1}^{N} TPR_i = \frac{1}{N}\sum_{i=1}^{N}\frac{pos_i - i}{T} = -\frac{N+1}{2T} + \frac{1}{NT}\sum_{i=1}^{N} pos_i$$

而绿线下的面积呢?

$$\frac{1}{N}\sum_{i=1}^{N-1} TPR_i = \frac{1}{N}\sum_{i=1}^{N-1}\frac{pos_i - i}{T} = -\frac{N-1}{2T} + \frac{1}{NT}\sum_{i=1}^{N-1} pos_i = -\frac{N+1}{2T} + \frac{1}{NT}\sum_{i=1}^{N} pos_i + \frac{1}{N}$$

AUC可以体现样本中正负样例的相对顺序

我们不妨想象

- 理想的二分类模型是sorted sample中所有正样例排在所有负样例之前,ROC曲线就是连接$(0,0),(1,1)$的曲线.换句话说$pos_i - i = T, \forall i \in 1,2,..,N$.AUC就介于1和$1-\frac{1}{N}$.当N(负样例数)无限大时,AUC就是1.
- 最糟糕的二分类器就是sorted sample所有负样例排在所有正样例之前,ROC曲线就是连接$(0,0),(\frac{N-1}{N},0)$和(1,1)的折线.换句话说,$pos_i - i = 0, \forall i \in 1,2,..,N$ AUC就介于0和$\frac{1}{N}$.当N(负样例数)无限大时,AUC就是0.
- sorted sample中第一个负样例出现的位置越靠前,召回率越低,这个点(FPR_1, TPR_1)离x轴越近,ROC曲线下的面积越小.同样,第二个负样例出现的位置越靠前,召回率越低,ROC曲线第二个点下面的面积越小.推而广之,全部负样例在sorted sample中出现的位置越靠前,ROC曲线覆盖的面积越小,也就是AUC越小.全部负样例在sorted sample中位置和AUC的具体关系看AUC计算公式.

NDCG (Normalized Discounted Cumulative Gain)

可以参考刘铁岩老师的learning2rank (https://github.com/HAOzj/Reading-and-Summary/blob/master/learning_to_rank.pdf).

NDCG和AUC的不同之处在于NDCG中一般用$log_2 pos_i + 1$的discount来惩罚负样例拍在正样例前,因为该discount函数是凸函数,所以NDCG对小的pos_i,也就是负样例的预测概率比太多的正样例的预测概率都高,惩罚更强.

Rank Correlation

Rank Correlation是基于样例对之间的相对顺序的.同样可以参考刘铁岩老师的learning2rank (https://github.com/HAOzj/Reading-and-Summary/blob/master/learning_to_rank.pdf)

使用sklearn做模型评估

除了上面的这些指标,sklearn还提供了一些其他接口来做分类模型的评估

接口	说明
metrics.accuracy_score(y_true, y_pred[, …])	模型准确度(Accuracy)
metrics.auc(x, y[, reorder])	使用梯形法则计算曲线下面积(AUC)
metrics.average_precision_score(y_true, y_score)	计算平均精确率(AP)
metrics.brier_score_loss(y_true, y_prob[, …])	计算Brier得分
metrics.classification_report(y_true, y_pred)	构建主要分类指标的文本报告
metrics.cohen_kappa_score(y1, y2[, labels, …])	Cohen's kappa: 一个衡量内部注释者协议的统计量
metrics.confusion_matrix(y_true, y_pred[, …])	计算混淆矩阵来评估分类的准确性
metrics.f1_score(y_true, y_pred[, labels, …])	计算F1得分
metrics.fbeta_score(y_true, y_pred, beta[, …])	计算F-beta得分
metrics.hamming_loss(y_true, y_pred[, …])	计算平均海明损失
metrics.hinge_loss(y_true, pred_decision[, …])	计算平均hinge损失
metrics.jaccard_similarity_score(y_true, y_pred)	Jaccard相似系数评分
metrics.log_loss(y_true, y_pred[, eps, …])	对数损失,又名逻辑损失或交叉熵损失
metrics.matthews_corrcoef(y_true, y_pred[, …])	计算马修斯相关系数(MCC)
metrics.precision_recall_curve(y_true, …)	针对不同的概率阈值计算精度(precision),召回率对
metrics.precision_recall_fscore_support(…)	计算精度，召回率，f1,support对
metrics.precision_score(y_true, y_pred[, …])	计算精度
metrics.recall_score(y_true, y_pred[, …])	计算召回
metrics.roc_auc_score(y_true, y_score[, …])	计算特征曲线（ROC AUC）下预测分数的计算区域
metrics.roc_curve(y_true, y_score[, …])	计算ROC
metrics.zero_one_loss(y_true, y_pred[, …])	0-1分类器损失

指标的选取

使用哪种或哪些指标来评估模型时,需要考虑模型的应用.

- 如果我们更关注模型对单个样例的预测,比如预测一个用户的性别而不关心两个用户之间相对的性别倾向时,基于样例的指标就更合适;
  - 如果模型的目标是尽量地找到某类样例,不在乎找错,也就是“宁可错杀一千也不放过一个”,recall更合适
  - 如果模型的目标是尽量准确地找到某类样例,不想错把其他类别归为该类,precision更合适
  - 如果模型的目标介于两者之间,F1-score更合适
- 如果我们侧重的是正负样例的相对关系,比如做推荐时做点击率预测,我们更加侧重不同物品被点击的相对大小,这个时候AUC和NDCG就更有意义.
- 如果我们侧重的是某些样例对之间的关系,Rank correlation可能更合适.

In [ ]:

In [ ]: