机器学习模型性能评估（二）：P-R曲线和ROC曲线

原创 石头 机器学习算法那些事 2018-09-23

上文简要介绍了机器学习模型性能评估的四种方法以及应用场景，并详细介绍了错误率与精度的性能评估方法。本文承接上文，继续介绍模型性能评估方法：P-R曲线和ROC曲线。

2.2 查准率、查全率与F1

错误率和精度虽然常用，但是不能满足特定的任务的需求。以西瓜问题为例，假定瓜农拉来一车西瓜，我们用训练好的模型对这些西瓜进行判别。

错误率衡量有多少比例的瓜被判别错误，但是若我们关心的是“挑出的西瓜中有多少比例是好瓜”，或者“所有好瓜中有多少比例被挑选了出来”。

这两类问题分别对应查准率和查全率，错误率是反映不了这两类问题。

例如在信息检索中，我们经常会关心“检索出的信息中有多少比例是用户感兴趣的”，“用户感兴趣的信息中有多少被检索出来的”；在视频监控中，我们关注的是“人脸识别的罪犯中有多少比例是真的罪犯”，“所有罪犯中有多少比例被识别出来”。

错误率计算较笼统，查准率和查全率是更为适用于此类需求的性能度量。查准率关注的问题是筛选的样本中是正样本的比例，查全率关注的问题是筛选的样本中有多少比例的正样本被筛选出来。

混淆矩阵是计算查准率和查全率或其他模型性能评估方法的基础。

混淆矩阵定义：

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

TP：真正例（true positive），即真实结果和预测结果都是正例。

FP：假正例（false positive），即真实结果是反例、预测结果是正例。

TN：真正例（true negative），即真实结果和预测结果都是反例。

FN：假反例（false negative），即真实结果是正例、预测结果是反例。

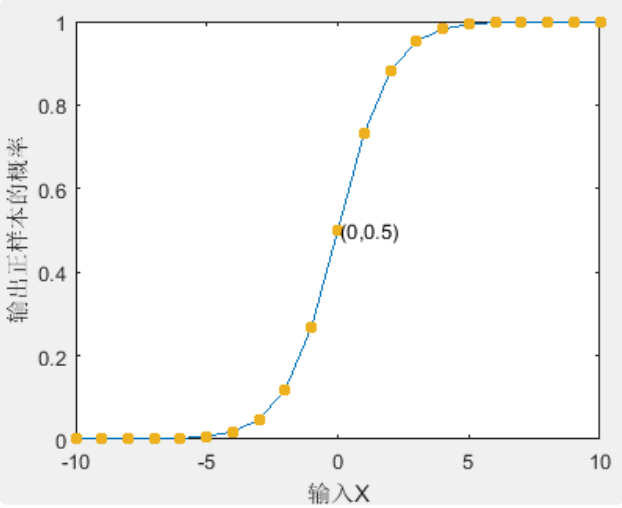
查准率P（Precision）与查全率R（Recall）分别定义为：

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

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

查准率与查全率是一对矛盾的度量，我们可以用极端的方式去理解这一矛盾。若所有测试样本的分类结果都是正样本，那么模型的查全率为1，查准率就很低；若几乎所有测试样本的分类结果都是负样本，那么模型的查准率很高，查全率就很低。

很多情形下，学习模型对测试数据输出的结果是具体的数值，如逻辑斯谛克生成模型P(Y=1|X)，表示输入变量X输出为正样本的概率，曲线图如下：

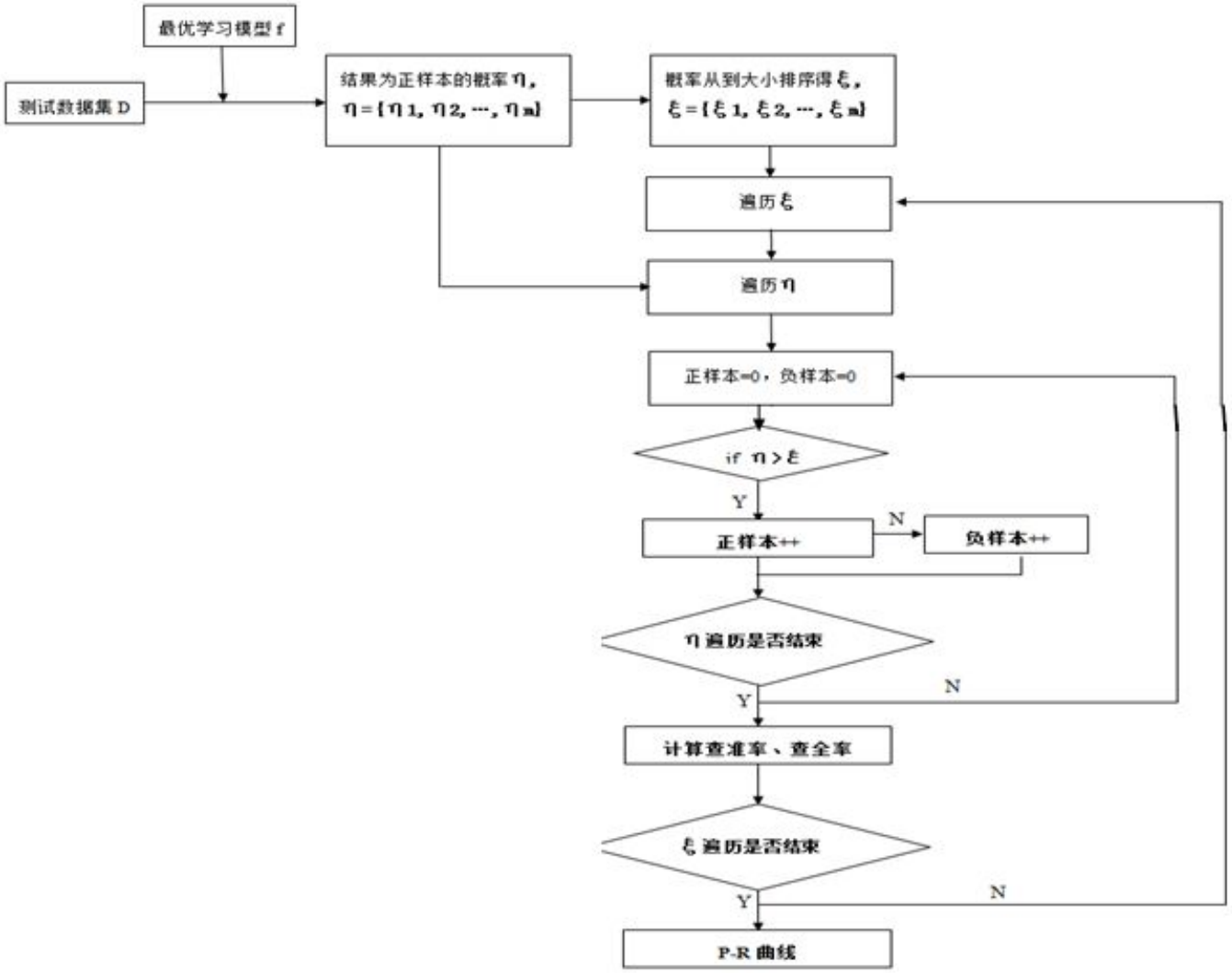


因此，学习模型P(Y=1|X)对测试数据集输出一系列为正样本的概率，根据概率由大到小排列，然后依次设置阈值，若大于该阈值，则为正样本；反之则为负样本。每次阈值

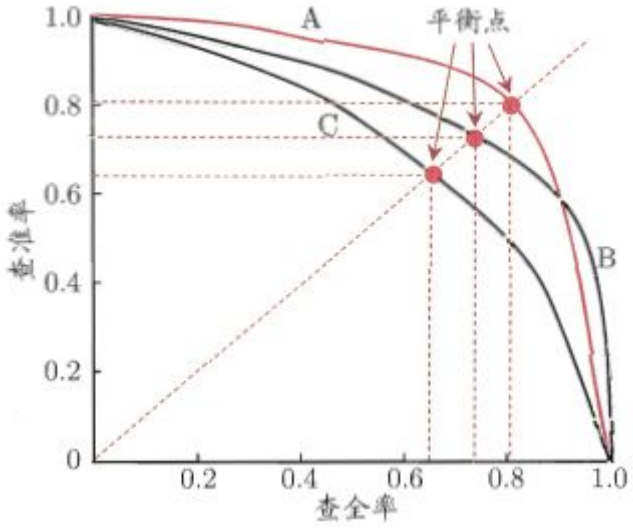
的设置都有对应的查准率和查全率，因此以查全率为横坐标，查准率为纵坐标，就可以得到查准率-查全率曲线，检测“P-R”曲线。

假设测试数据集D样本量为m，$D = \{(X1,Y1),(X2,Y2),\cdots,(Xm,Ym)\}$

P-R曲线流程图如下：



P-R曲线图如下：



根据P-R曲线来评估模型的性能：

（1）若一个学习模型的P-R曲线完全包住另一个学习模型的P-R曲线，则前者的性能优于后者。即查全率相同的情况下，查准率越高模型的泛化性能越好，如模型A优于模型B。

（2）若两个学习模型的P-R曲线互相交叉，则可通过“平衡点”（Break-Event Point，简称BEP）来评价模型的优劣，BEP是“查准率=查全率”的数值。由上图可知，模型A的平衡点大于模型B的平衡点，即模型A优于B。

（3） 由于BEP过于简化，更常用的是F1度量：

$$F1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{样例总数+TP-TN}$$

F1越大，性能越好。

（4） F1度量认为查全率和查准率的重要性程度一样，若考虑到查全率和查准率的重要性程度不一样，如推荐给用户的信息尽可能是用户感兴趣的，那么查准率更重要；抓捕逃犯时更希望尽可能少漏掉逃犯，此时查全率更重要（概念有点模糊的可以参考查准率和查全率公式）。

为了描述查准率和查全率的相对重要程度，则用F1度量的一般形式：Fβ。

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

其中，β > 0度量了查全率对查准率的相对重要性，β=1时退化为标准的F1；β > 1时查全率更重要；β < 1时查准率更重要。

2.3 ROC曲线与AUC

P-R曲线是从查准率和查全率的角度去衡量学习模型的泛化性能，ROC曲线则是从更一般的情况下去衡量学习模型的泛化性能，若没有任何先验条件的限制情况下，推荐用ROC曲线去衡量模型的泛化性能。

绘制ROC曲线的思想与P-R曲线一致，对学习模型估计测试样本为正样本的概率从大到小排序，然后根据概率大小依次设置阈值，认为大于阈值的测试样本为正样本，小于阈值的测试样本为负样本，每一次阈值设置都计算“真正例率”（True Positive Rate，简称TPR）和“假正例率”（False Postive Rate，简称FPR）。

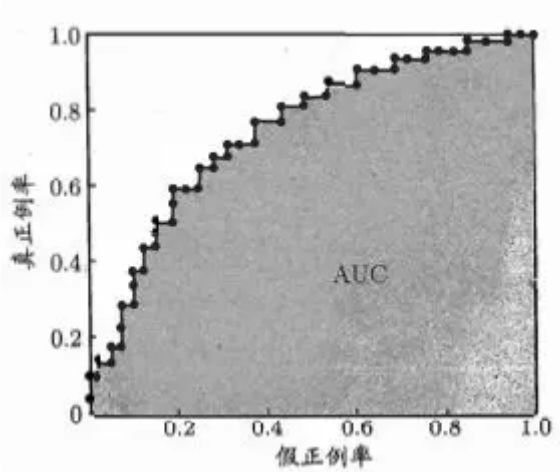
TPR和FPR的定义如下：

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

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

TP，FP，TN，FN的定义可参考上节的混淆矩阵。

ROC曲线横坐标为假正例率，纵坐标为真正例率，曲线图如下：



本文对ROC曲线的首末两点进行解释：

测试数据集包含N例正样本和M例负样本，若阈值设置的最大，则学习模型对所有测试样本都预测为负样本，混淆矩阵如下：

真实情况	预测结果	
	正例	反例
正例	0	N
反例	0	M

TPR = TP/(TP + FN) = 0/(0+N) = 0;

FPR = FP/(TN + FP) = 0/(0+M) = 0;

因此，当阈值设置最大时，TPR与FPR均为0。

若阈值小于所有模型估计测试样本为正样本的数值时，则测试样本均为正样本，混淆矩阵如下：

真实情况	预测结果	
	正例	反例
正例	N	0
反例	M	0

TPR = TP/(TP+FN) = N/(N+0) = 1;

FPR = FP/(TN+FP) = M/(M+0) = 1;

因此，当阈值设置最小时，TPR与FPR均为1。

AUC（Area Under Curve）为ROC曲线的面积，面积可以通过梯度面积法求解。

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1})$$

AUC的计算表达式理解起来有点别扭，假设正负样本数均为m例，大家回想下ROC曲线的算法思想，假正例率对应的是真实负样本中分类结果为正样本的比例，真正例率对

应的是真实正样本中分类为正样本的比例。

假正例率和真正例率的增长性具有互斥性，每次都只能增加一个，且每次增加的梯度为1/m，横坐标和纵坐标共增加了m次。

理解了这个原理，相信对ROC曲线的绘制和AUC面积的计算应该有更深的认识了吧。

AUC是衡量模型泛化能力的一个重要指标，若AUC大，则分类模型优；反之，则分类模型差。想象一下，若假正例率不变的情况下，真正例率越大，对应的AUC也越大，则模型的泛化能力强，这与实际情况相符。

参考资料：

《机器学习》周志华 著

未完，待续。。。