模型评测

Allen·Huang

模型的比较

·一次训练过程中模型的比较

·多次训练的模型的比较

·不同算法的模型的比较

评测集

·评测集不混入训练集中训练

·评测集只能代表片面的数据评测

·评测集需要数量和数据分布方面的扩充

正样本与负样本

·通常我们需要判定概率为1的类型的样本叫做正样本

·通常我们需要判定概率为0的类型的样本叫做负样本

·扩展到多分类

二分类评测指标

- *Classification Accuracy*
- *Confusion Matrix*
- *ROC Curve*
- *Area under Curve*
- *F1 Score*
- *PR Curve*
- *AP Score*

Accuracy 准确率

准确率/正确率：

通常是指在所有的预测中，与正确答案相等的比例是多少

无论正样本还是负样本

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

Accuracy 准确率

缺陷：

当数据训练集和评测集的数据比例不平衡的时候，难以准确衡量模型的好坏

在训练集中，当A类型有98%的数据量，B类型有2%的数据量的时候，那么我们的模型可以轻松达到98%的准确率，如果将所有的训练结果都记做A

当评测集中60%的数据是A，40%的数据是B的时候，测试的准确率就会降到60%

Confuse Matrix混淆矩阵

n代表总体的测试样本数

第一行是实际负样本数

第二行是实际正样本数

第一列是预测负样本数量

第四列是预测正样本数量

所以对于二分类我们有4种预测后的结果

真负，假负，真正，假正

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Confusion Matrix

Confuse Matrix混淆矩阵

TP:真正

TN:真负

FN:假负

FP:假正

		预测	
		1	0
实际	1	True Positive（TP）	False Negative（FN）
	0	False Positive（FP）	True Negative(TN)

Confuse Matrix

事实上，Accuracy=(TP+TN)/n

事实上，后面的很多指标也可以用混淆矩阵中的数据来进行组合

Precision 和 Recall

Precision 查准率：

Precision=TP/(TP+FP)

就预测y=1这件事情的正确率是多少?

就预测y=1这件事情的误报率是多少? 1-Precison

Precision 和 Recall

Recall 查全率:

Recall=TP/(TP+FN)

对于所有y=1的数据有多少真的被报出来了?

对于所有y=1的数据有多少的漏报率? 1-Recall

一个小例子

- 金融诈骗2分类问题(1:诈骗,0:不诈骗)
- P-100 N-100
- Thres=0.5 当大于等于0.5的时候认为1,当小于0.5认为是0
- 预测出来TP-80 FP-20 ---报出100个样本
- Precision？ 0.8
- Recall？ 0.8
- 如果Thres=0.9，大于等于0.9的时候认为是1，小于0.9认为是0
- Precison和Recall的变化
- [0.2,0.1,0.35,0.7,0.98……..,0.23]---200个预测的样本
- 当Thres=0.9的时候报出的样本数量变多了还是变少了？

Precision 和 Recall

右图为评估+GroundTruth表

通过调整不同的阈值

score>? 则预测为p

得到一系列的数字

阈值越低，Recall越高，Precision越低

阈值越高, Recall越低,Precison越高

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35

PR曲线

Thres=0.9 P=1,R=0.1

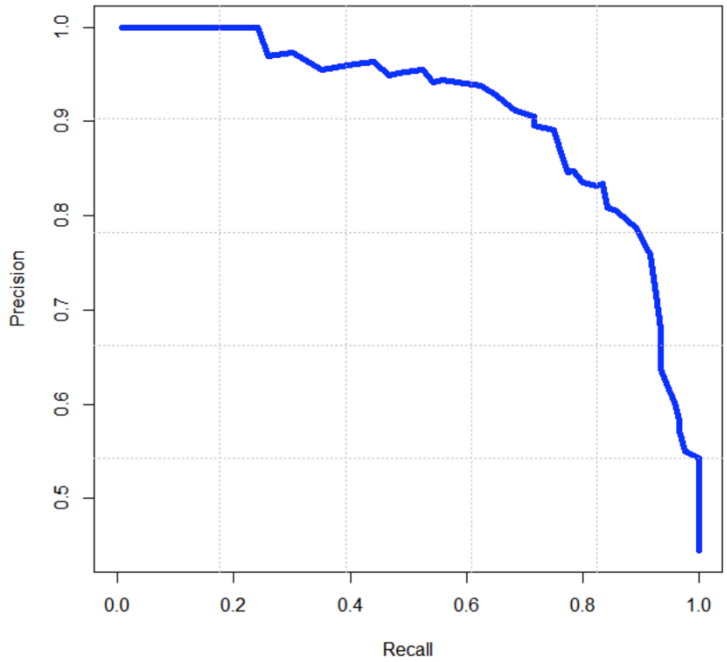
Thres=0.8 P=1,R=0.2

Thres=0.7 P=0.6,R=0.3

…

Thres=0.1 P=0.5,R=1.0

对于PR曲线越靠近右上越好



ROC曲线

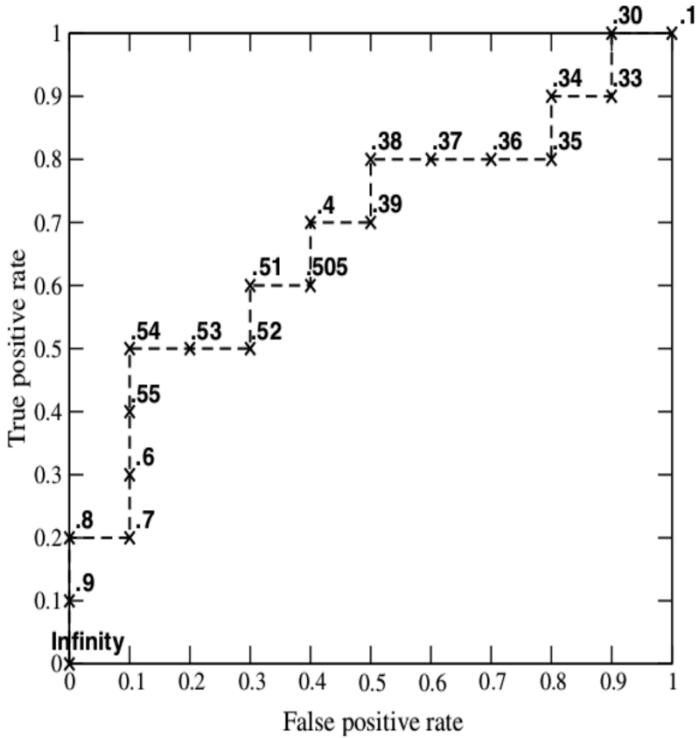
Recall=TP/(TP+FN)

TPR=Recall=TP/(TP+FN),真正例率,Recall

FPR=1-FRecall=FP/(FP+TN),

假正例率,1-FRecall,误召率

AUC=该条曲线的面积(利用微分进行计算)



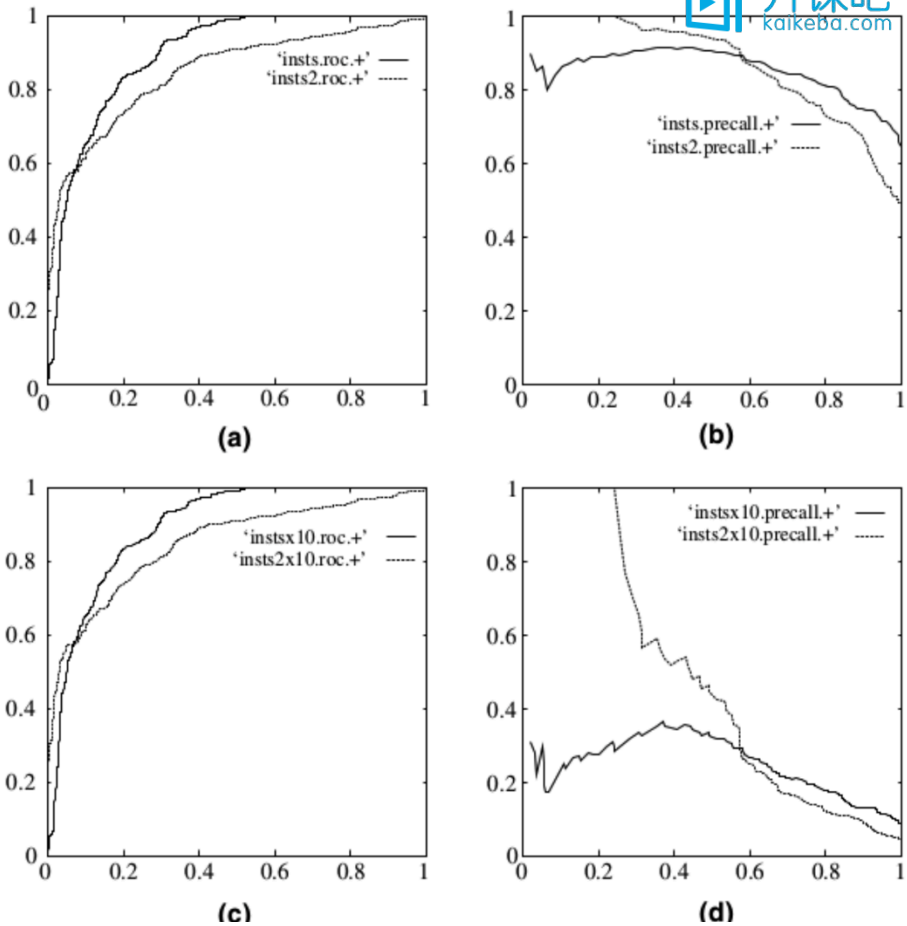
对比ROC和PR

·ROC永远是单调曲线

·正负样本失衡的时候，PR变化大而ROC可以保持不变

·在混入了更多的错误的时候也引入了更多的正确

·在Recall稳定提升的过程中，可能错误的判断如洪水般涌入，将正确率淹没



多分类评测

每一类当做是二分类来进行评测

计算Micro/Macro的PR、ROC

计算mAP

Micro

$$Micro - Precision = \frac{TruePositives1 + TruePositives2}{TruePositives1 + FalsePositives1 + TruePositives2 + FalsePositives2}$$

$$Micro - Recall = \frac{TruePositives1 + TruePositives2}{TruePositives1 + FalseNegatives1 + TruePositives2 + FalseNegatives2}$$

$$Micro - F - Score = 2.\frac{Micro - Precision\,.\,Micro - Recall}{Micro - Precision + Micro - Recall}$$

Macro

$$Macro - Precision = \frac{Pecision1 + Precision2}{2}$$

$$Macro - Recall = \frac{Recall1 + Recall2}{2}$$

$$Macro - F - Score = 2.\frac{Macro - Precision\,.\,Macro - Recall}{Macro - Precision + Macro - Recall}$$