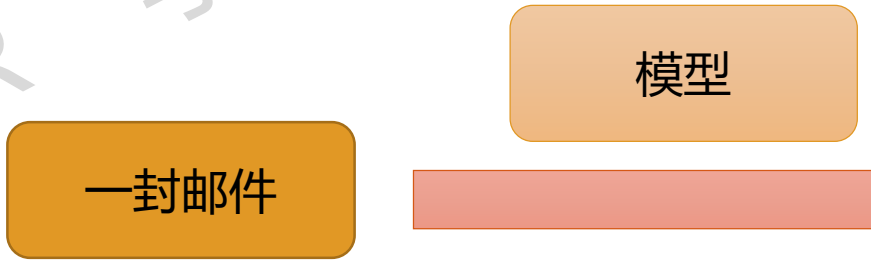
单标签分类评价指标介绍

https://github.com/sangjianshun/Master-School/blob/master/SingleClassEvaluationIndex.py

- 业务背景



是否是垃圾邮件
1
0.7

1	0.6
1	0.8
1	0.2
1	0.9

0	0.1
0	0.2
0	0.2
0	0.3
0	0.6
0	0.7

1	1
1	1
1	0
1	1

0	0
0	0
0	0
0	0
0	1
0	1

单标签分类评价指标介绍

- 混淆矩阵（Confusion Matrix）

	标签类别			
预测类别		Positive	Negative	Total
	Positive	TP	FP	预测为正
	Negative	FN	TN	预测为负
	Total	所有正例	所有负例	样本数

1	1
1	1
1	0
1	1

0	0
0	0
0	0
0	0
0	1
0	1

TP=3	FP=2
FN=1	TN=4

$$Precison = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{3}{5} = 0.6$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{3}{4} = 0.75$$

$$F_\alpha = \frac{(\alpha^2+1)P*R}{\alpha^2*P+R}$$

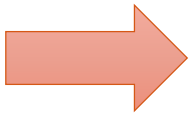
$$F1\ Score = \frac{2*P*R}{P+R} = 0.67$$

单标签分类评价指标介绍

- 精确率和召回率存在的问题

0	1

1	1
1	1
1	1
1	1
1	1
1	1



TP=6	FP=1
FN=0	TN=0

$$Precison = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{6}{7}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{6}{6}$$

单标签分类评价指标介绍

ROC曲线（AUC）

1	0.6
1	0.8
1	0.2
1	0.9

0	0.1
0	0.2
0	0.2
0	0.3
0	0.6
0	0.7

阈值0.5

1	1
1	1
1	0
1	1

0	0
0	0
0	0
0	0
0	1
0	1

TP=3	FP=2
FN=1	TN=4

$$TPR = \frac{TP}{TP + FN} = \frac{3}{4} = 0.75$$

$$FPR = \frac{FP}{FP + TN} = \frac{2}{6} = 0.33$$

阈值0.65

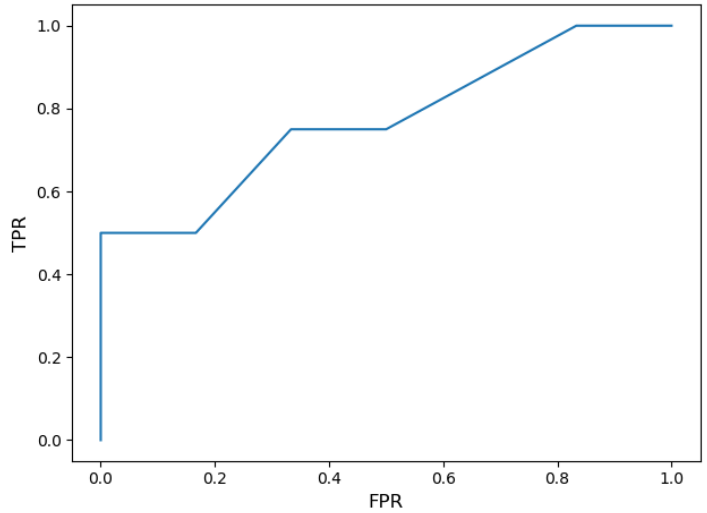
1	0
1	1
1	0
1	1

0	0
0	0
0	0
0	0
0	0
0	1

TP=2	FP=1
FN=2	TN=5

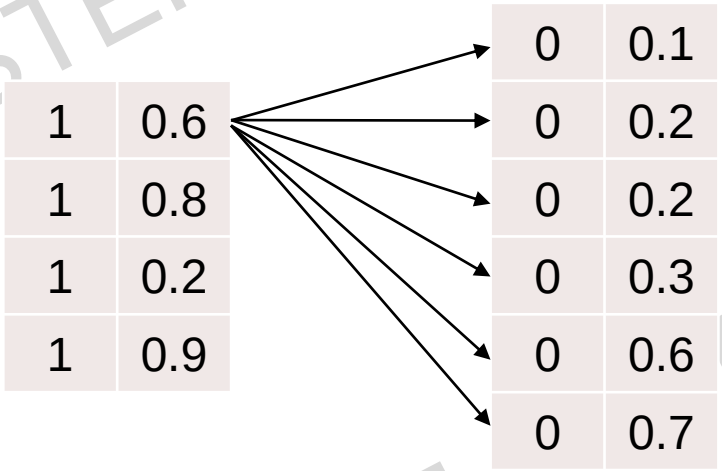
$$TPR = \frac{TP}{TP + FN} = \frac{2}{3} = 0.67$$

$$FPR = \frac{FP}{FP + TN} = \frac{1}{6} = 0.17$$



单标签分类评价指标介绍

- AUC的计算



$$\mathrm{AUC} = \frac{4.5 + 6 + 2 + 6}{4 * 6} = \frac{18.5}{24} = 0.77$$

```
for pos1 in pos:
    for neg1 in neg:
        if pos1>neg1:
            count += 1
        elif pos1==neg1:
            count+=0.5
        else:
            break
```