分类是一种重要的数据挖掘算法。分类的目的是构造一个分类函数或分类模型（即分类器），通过分类器将数据对象映射到某一个给定的类别中。分类器的主要评价指标有准确率(Precision)、召回率(Recall)、Fb-score、ROC、AOC等。在研究中也有采用Accuracy（正确率）来评价分类器的。但准确率和正确率这两个概念经常有人混了

1. ROC是反映敏感性和特异性连续变量的综合指标,是用构图法揭示敏感性和特异性的相互关系，它通过将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、（1-特异性）为横坐标绘制成曲线
   1. TPR=TP/(TP+FN)
   2. FPR=FP/(FP+TN)

具体方法为：将probability>=threshold分为Positive,反之为Negative。将threshold从1逐渐改变为0，对于每一个threshold 计算TPR， FPR。最后以TPR为y轴，FPR为y轴画出ROC曲线。

1. AUC是ROC曲线下面积(Area Under roc Curve)的简称，顾名思义，AUC的值就是处于ROC curve下方的那部分面积的大小。通常，AUC的值介于0.5到1.0之间，AUC越大，诊断准确性越高。在ROC曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。
2. 查准率和查全率

对于2分类算法而言，分类结果只有4种，分别为：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 实际 | |
| P | N |
| 预测 | P | True Positive(TP) | False Positive(FP) |
| N | False Negative(FN) | True Negative(TN) |

查准率（准确率）Precision(TPR)=TP/(TP+FN)

查全率（召回率）Recall=TP/(TP+FP)

1. 其他衡量指标：
   1. 训练集大小与模型准确度的关系
   2. 计算速度
   3. 可解释性
2. 现存问题

模型是3分类问题，虽然在计算查准率和查全率时可以用混淆矩阵求出，但是计算ROC图时，对于3分类问题，需要用至少两个阈值来区分每一个类别的Positive和Negtive。

再加上TPR, FPR 是一个四维数据，无法直接绘图。

1. 2分类问题示例：

现有分类结果如下，真实标签为Label，模型给出每一个样本像Label 0 的概率 Probability

Label Probability

0 0.900000

0 0.800000

1 0.700000

0 0.600000

0 0.550000

0 0.540000

1 0.530000

1 0.520000

0 0.510000

1 0.505000

0 0.400000

1 0.390000

0 0.380000

1 0.370000

1 0.360000

1 0.350000

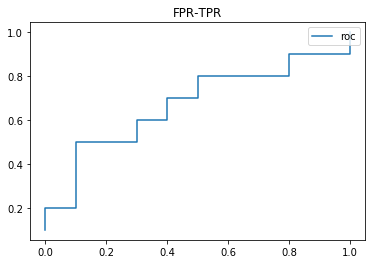
0 0.340000

1 0.330000

1 0.300000

0 0.100000

1. roc 曲线



1. AUC=0.670000
2. P=Label 0，N=Label 1

Threshold=0.5

Tag(x)=if probability>=Threshold then 0 else 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 实际 | |
| 0 | 1 |
| 预测 | 0 | 6 | 4 |
| 1 | 4 | 6 |

查准率（准确率）Precision(TPR)=TP/(TP+FN)=0.6

查全率（召回率）Recall=TP/(TP+FP)=0.6