混淆矩阵与ROC曲线

二分类问题的指标

对于二分类问题，其类别有两种：

- 正类
- 负类，也可以叫作反类

例如，在乳腺癌肿瘤预测这个二分类问题中，如果我们感兴趣的是类别为恶性的样本，则把恶性称为正类，良性称为负类。

两类错误

- 对于一个模型来说，总会有预测错误的时候。预测出错有两种可能性。

- 第一类错误：将一个类别为正类的样本预测为负类，就叫作假正类。
- 第二类错误：将一个类别为负类的样本预测为正类，就叫作假负类。

混淆矩阵（confusion matrix）

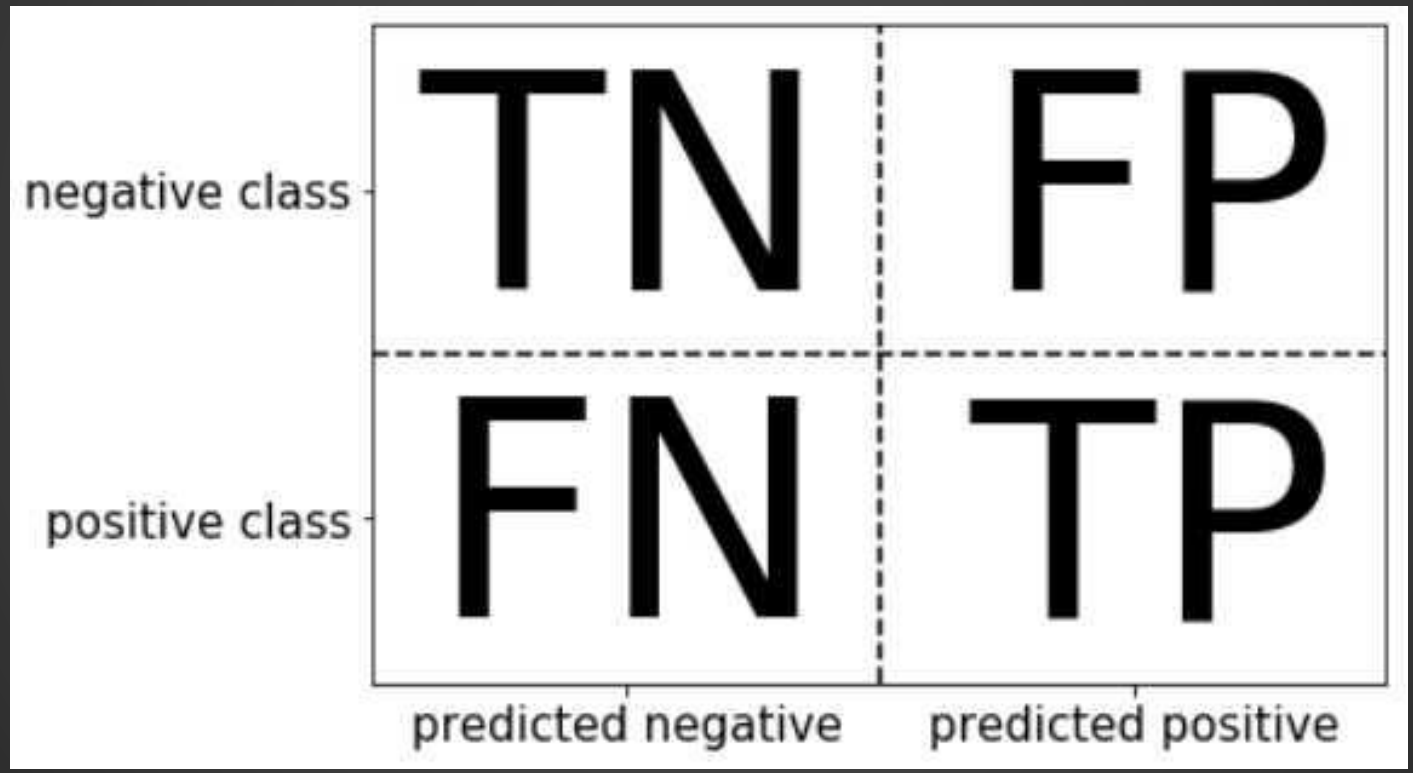
对于一个二分类问题来说，类别标签有两个类别，预测结果有两个类别。

- 样本类别：正类记为positive class，负类记为negative class。
- 预测类别：正类记为predicted positive，负类记为predicted negative。

样本类别和预测类别组合一下，共有四种情况。

1. 负类样本被预测为负类，称为真负类(true negative,TN)
2. 负类样本被预测为正类，称为假负类(flase negative,FN)
3. 正类样本被预测为负类，称为假正类(false positive,FP)
4. 正类样本被预测为正类，称为真正类(true positive,TP)

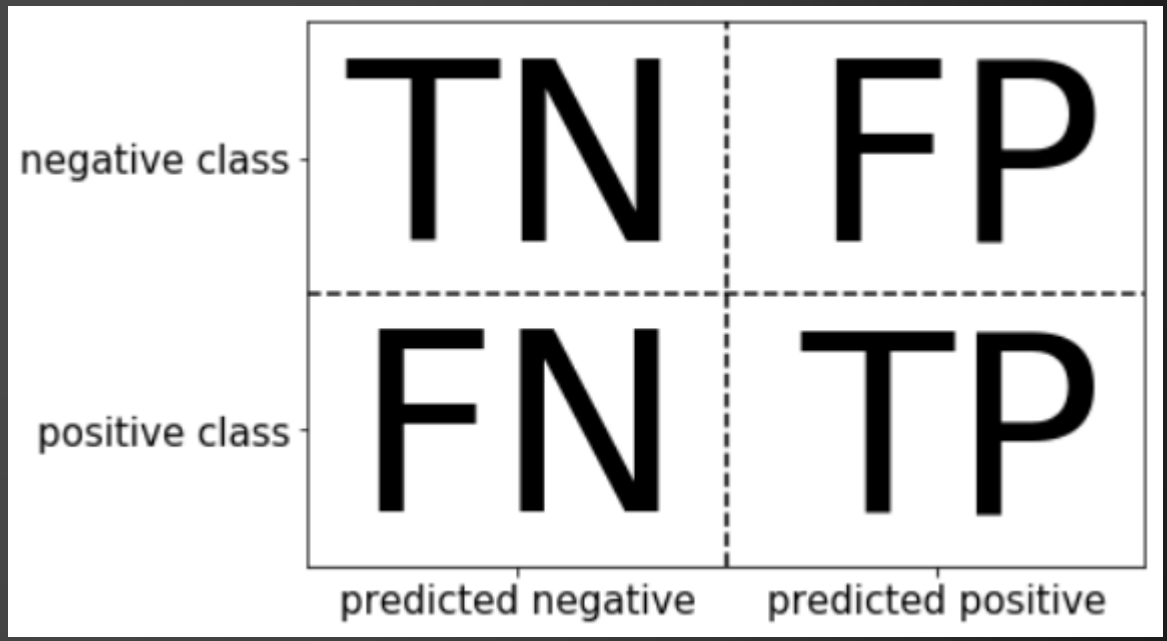
混淆矩阵（confusion matrix）



评价指标：精度

- 精度（Accuracy）：在所有正类样本中有多少被预测为正类的比例，正确分类数/样本总数。
- 精度计算公式

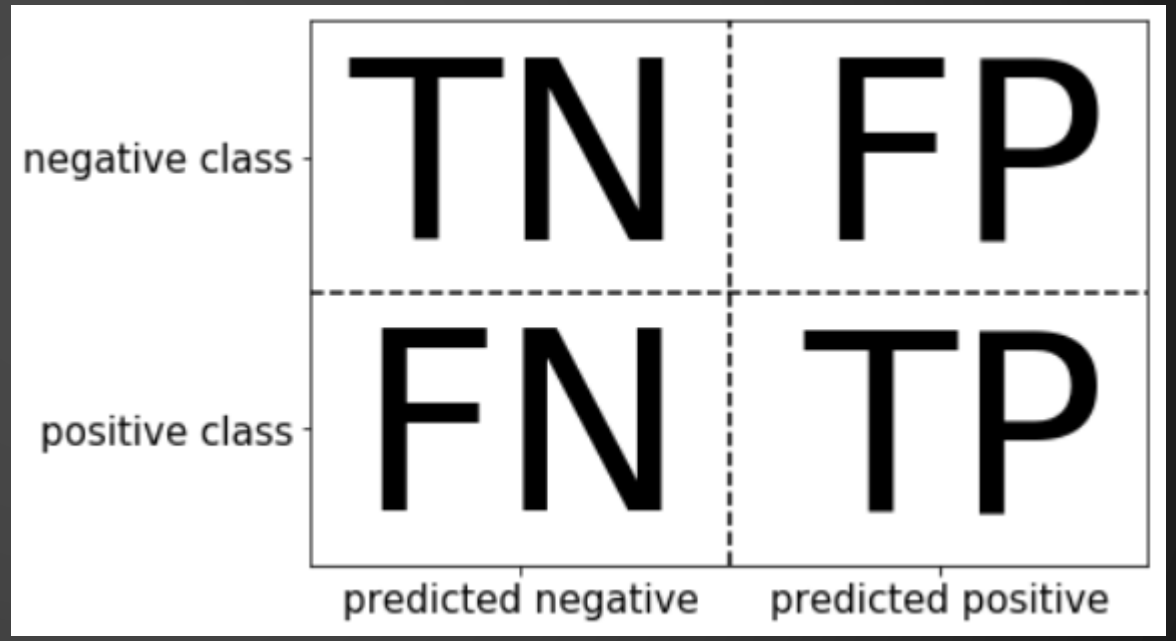
$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN}$$



评价指标：准确率

- 准确率（ Precision ）（也叫查准率）：在所有被预测为正类的样本中有多少是真正的正类。
- 查准率计算公式

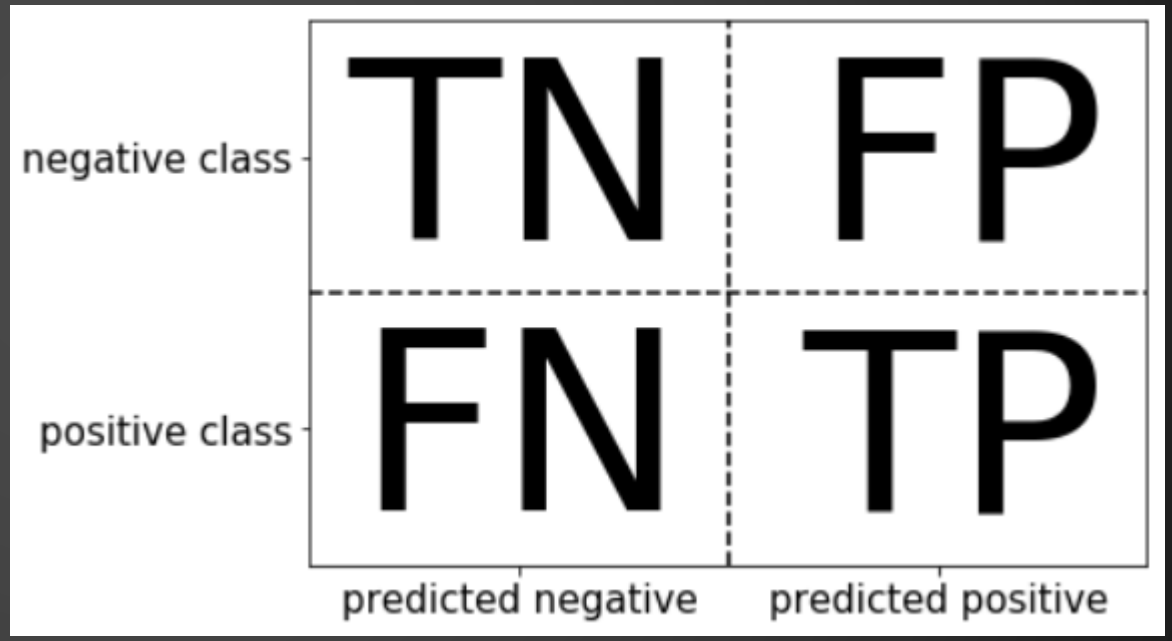
$$Precision = \frac{TP}{TP + FP}$$



评价指标：召回率

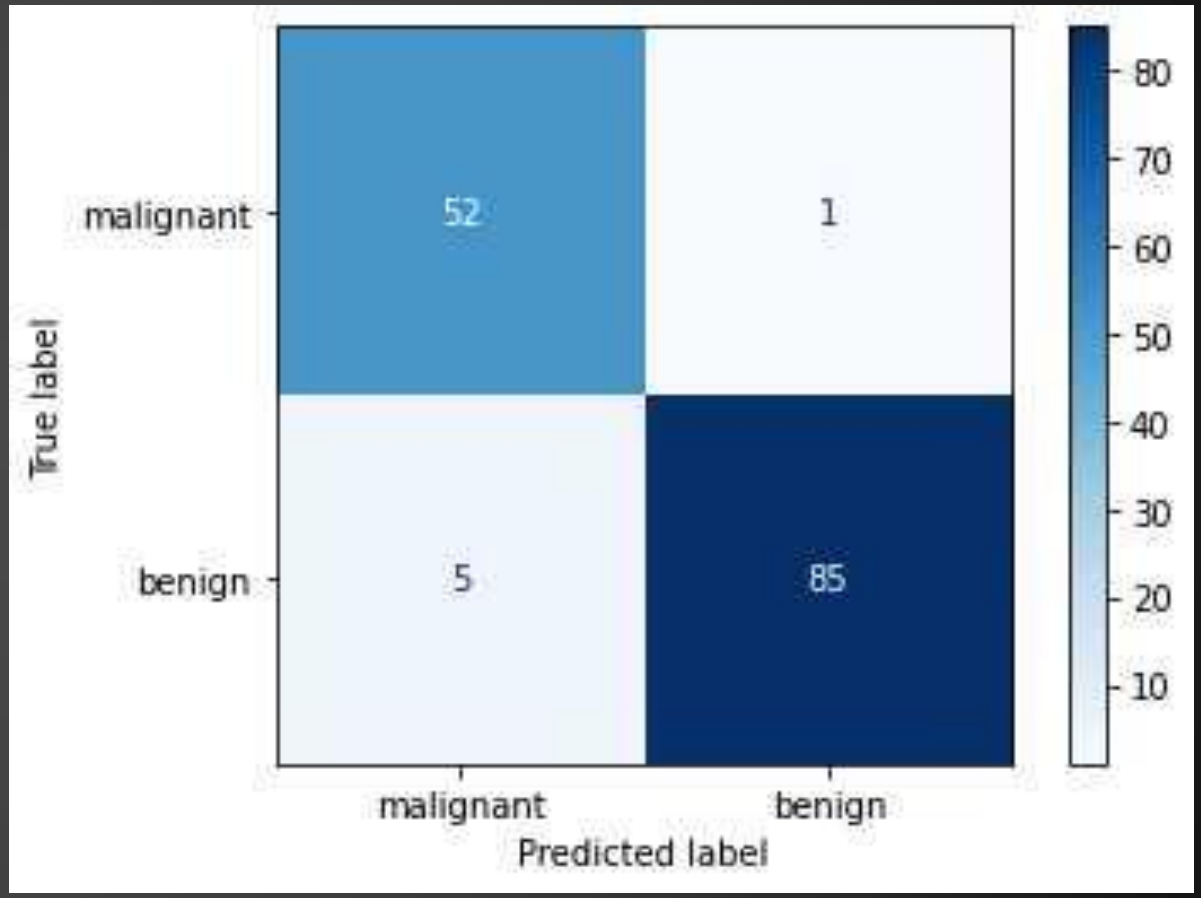
- 召回率（Recall），也叫灵敏度：所有正类样本中有多少被预测为正类，描述了分类器对正类的敏感程度。
- 召回率计算公式

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



如何得到混淆矩阵?

1. 得到混淆矩阵：confusion_matrix
2. 绘制混淆矩阵：plot_confusion_matrix



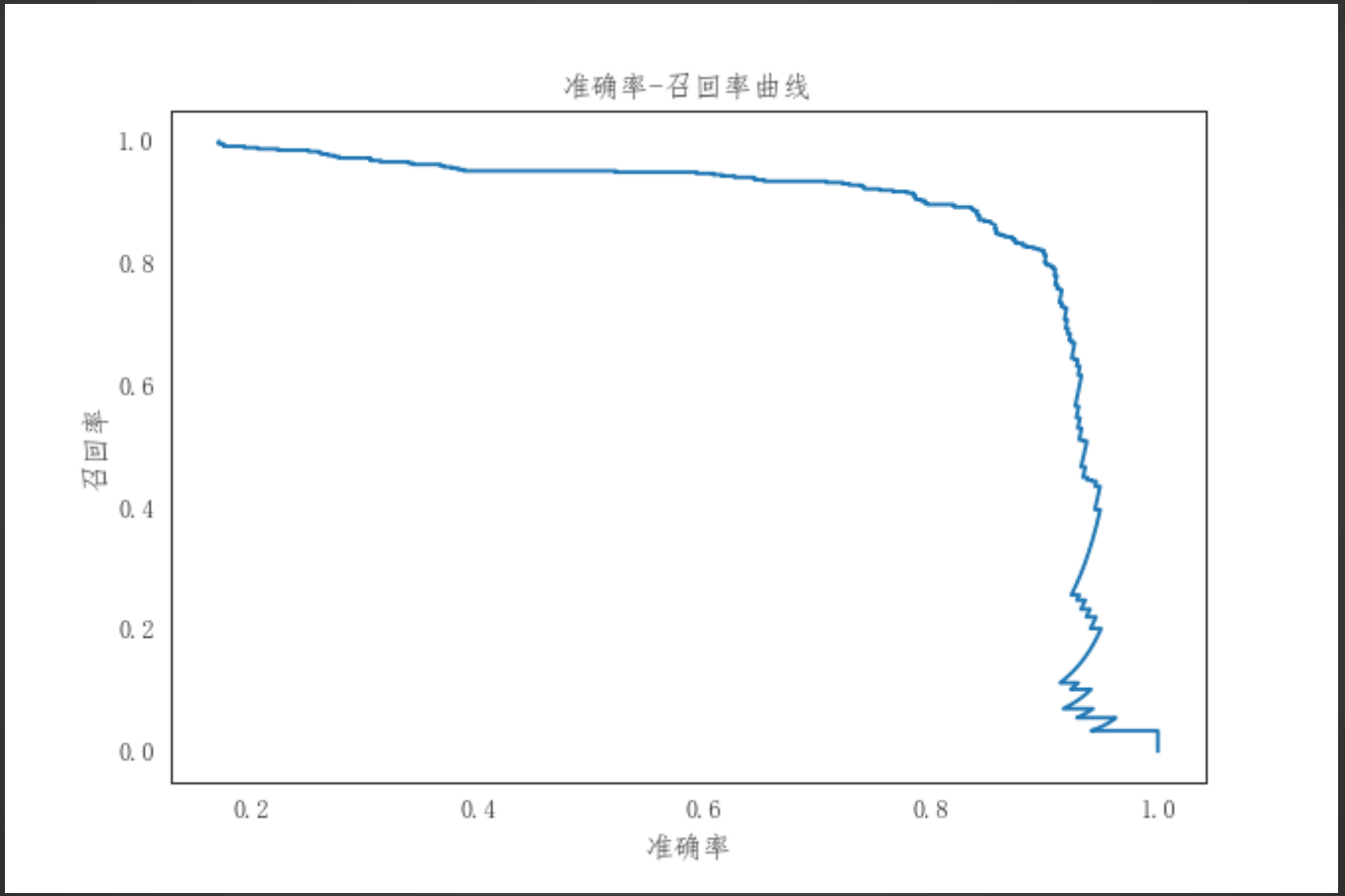
再谈准确率和召回率

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

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

- 如果预测少量确定的样本点为正类，其他均预测为负类，此时，假正类为零，即FP=0，可以得到100%的准确率。但是假负类会很多，即FN会很大，此时，召回率就会很低。
- 反过来，如果预测所有的样本均为正类，则没有假负类，即FN=0，可以得到100%的召回率，但是假正类会很多，即FP会很大，此时，准确率会很低。

准确率-召回率曲线

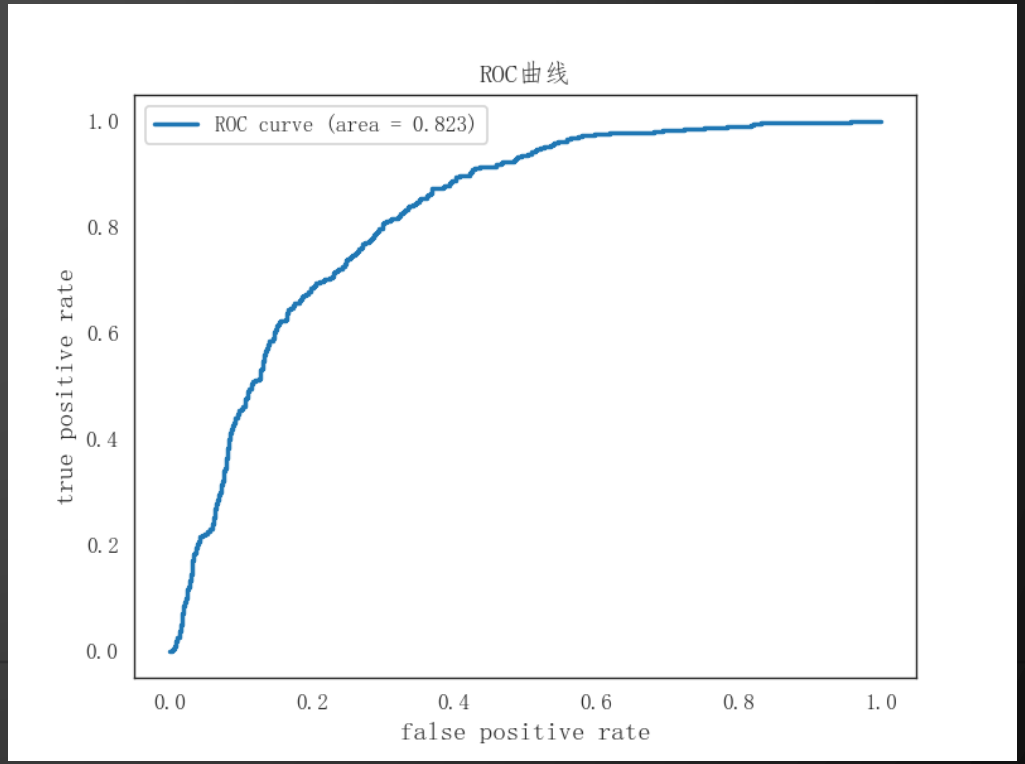


ROC曲线

- ROC是receiver operating characteristics curve的简称，中文翻译是受试者工作特性曲线，即ROC curve。
- ROC曲线考虑的是假正例率（false positive rate，FPR）和真正例率（true positive rate,TPR），真正例率其实就是之前说过的召回率，假正例率表示假正类占所有负类样本的比例，公式为：

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

- ROC曲线下的面积，即area under the curve，简称AUC



绘制ROC曲线

sklearn中的函数：

- 绘制ROC曲线：roc_curve
- 求AUC值：roc_auc_score