朴素贝叶斯分类器

基本要求

a)采用分层采样的方式将数据集划分为训练集和测试集。

b)给定编写一个朴素贝叶斯分类器，对测试集进行预测，计算分类准确率。

- 分层抽样

  分层抽样采用的方式是，先生成类的全部索引，再打乱顺序，再按照比例选取顺序次数个的索引出来当做测试集。

  例如，a类有5 个， 根据比例抽40%作为测试集，那么就先生成顺序0,1,2,3,4，打乱后，5,3,4,1,2，然后选取前5*40%=2个作为测试集，就是第5,3两个索引对应的样本作为测试集。

- 贝叶斯分类器，先算出每个类的均值和方差，再根据以下公式：

  $c_{MAP} = argmax_{c_k \in C} \prod_{d=1}^{D} P(x_d|c_k)\, P(c_k)$ ，算出对于每个样本每个类的预测值，选择最大的那个预测值所属的类，作为贝叶斯分类器的结果。

- 准确率结果如下：

```
Erro number is: 10
The erro rate is: 0.2777777777777778
```

中级要求

➢中级要求：使用测试集评估模型，得到混淆矩阵，精度，召回率，F值。

根据以下的混淆矩阵的原理：

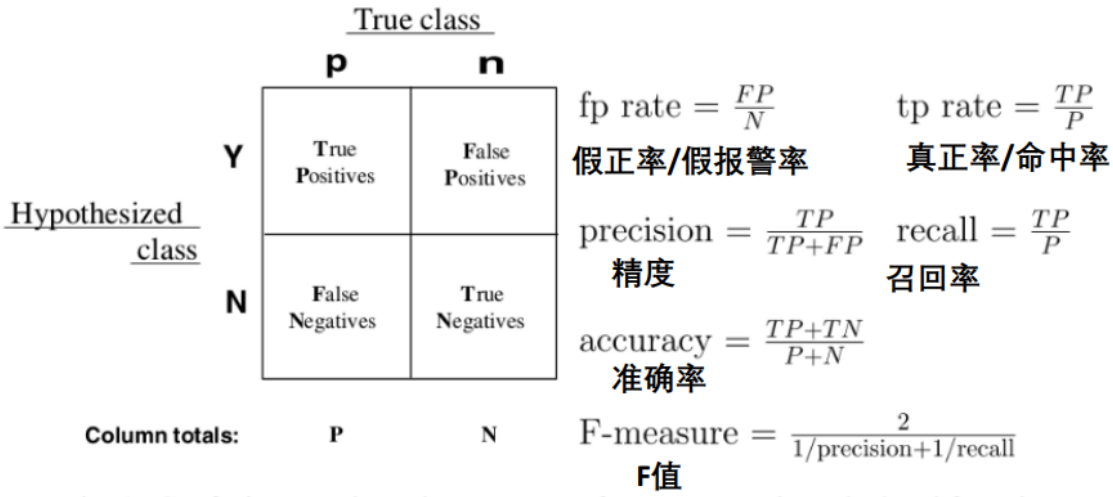


Fig. 1. Confusion matrix and common performance metrics calculated from it.

可以通过统计，TP,FP,FN,TN的值，来获得混淆矩阵和精度，召回率，F值。

```
confision matrix of class1 is
8 2
4 22
confision matrix of class1 is
9 4
5 18
confision matrix of class1 is
9 4
1 22
```

```
precision of class1 is 0.8
precision of class2 is 0.6923076923076923
precision of class3 is 0.6923076923076923
recall of class1 is 0.6666666666666666
recall of class2 is 0.6428571428571429
recall of class3 is 0.9
f-measure of class1 is 0.8
f-measure of class2 is 0.6923076923076923
f-measure of class3 is 0.6923076923076923
```

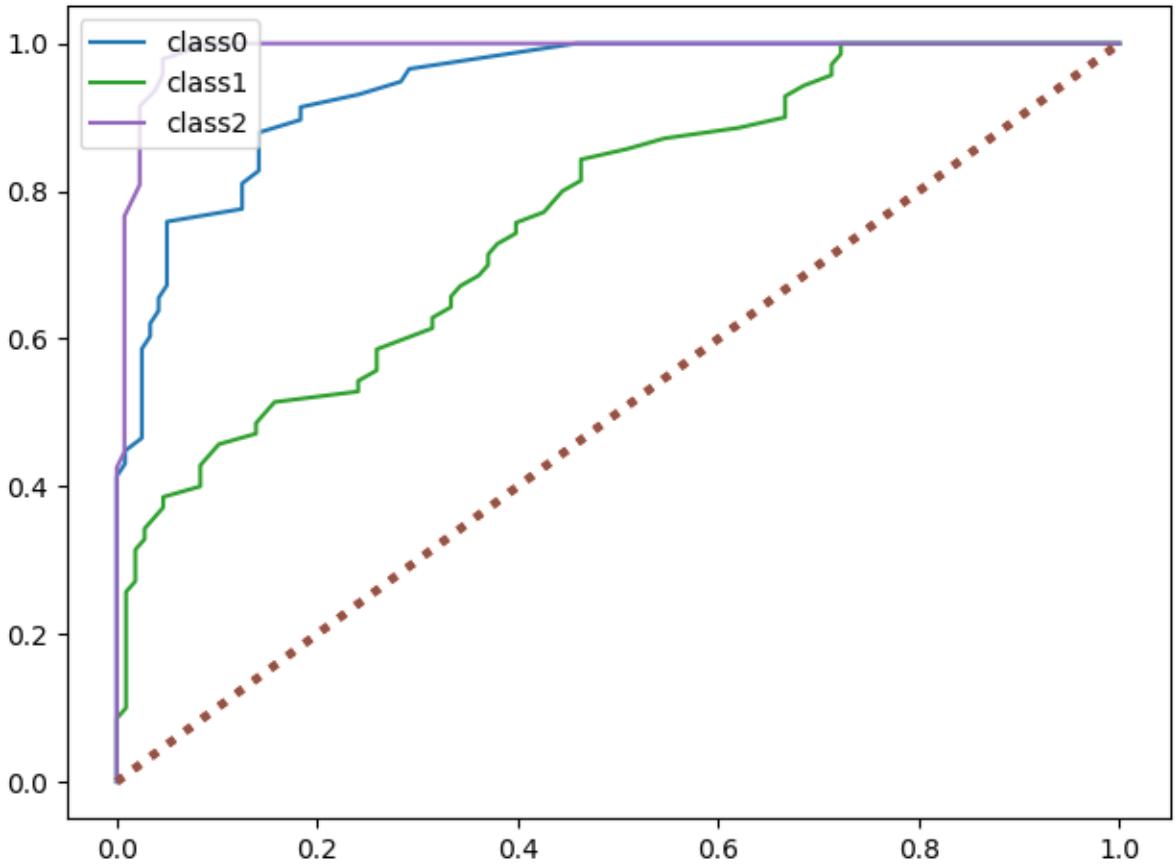
高级要求

在中级要求的基础上画出三类数据的ROC曲线，并求出AUC值。

ROC曲线

和中级中同样的方式，但是这一次将每一个预测值作为阈值进行混淆矩阵的计算，具体如下：

- 先给每个positive的预测值数组里添加一个0和无限大值--为了曲线能覆盖x轴的0-1
- 给positive数组从小到大排序
- 将每一个positive中的值分别当做阈值，计算每一个阈值下的混淆矩阵
- 根据每个混淆矩阵获得它们各自的false positive rate和true positive rate
- 绘制ROC



AUC值

最后根据rank方式的计算方法，计算auc值：

$$AUC = \frac{\sum_{ie\ positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N}$$

最后结果为：

```
The auc of class 0 is 0.8018775395824577
The auc of class 1 is 0.06966731898238747
The auc of class 2 is 0.59484375
```

（具体代码在py文件中）