2.1

$\binom{500}{350}^2$

2.3

BEP值为$P=R$时的值
$P=\frac{TP}{TP+FP}$
$R=\frac{TP}{TP+FN}$
$P=R$
所以$FN=FP$
$F1=\frac{2TP}{2TP+FP+FN}$
因为$FP=FN$
所以$F1=\frac{TP}{TP+FP}=R$
所以A的BEP值比B高

2.6

ROC曲线下面积越大 错误率越小

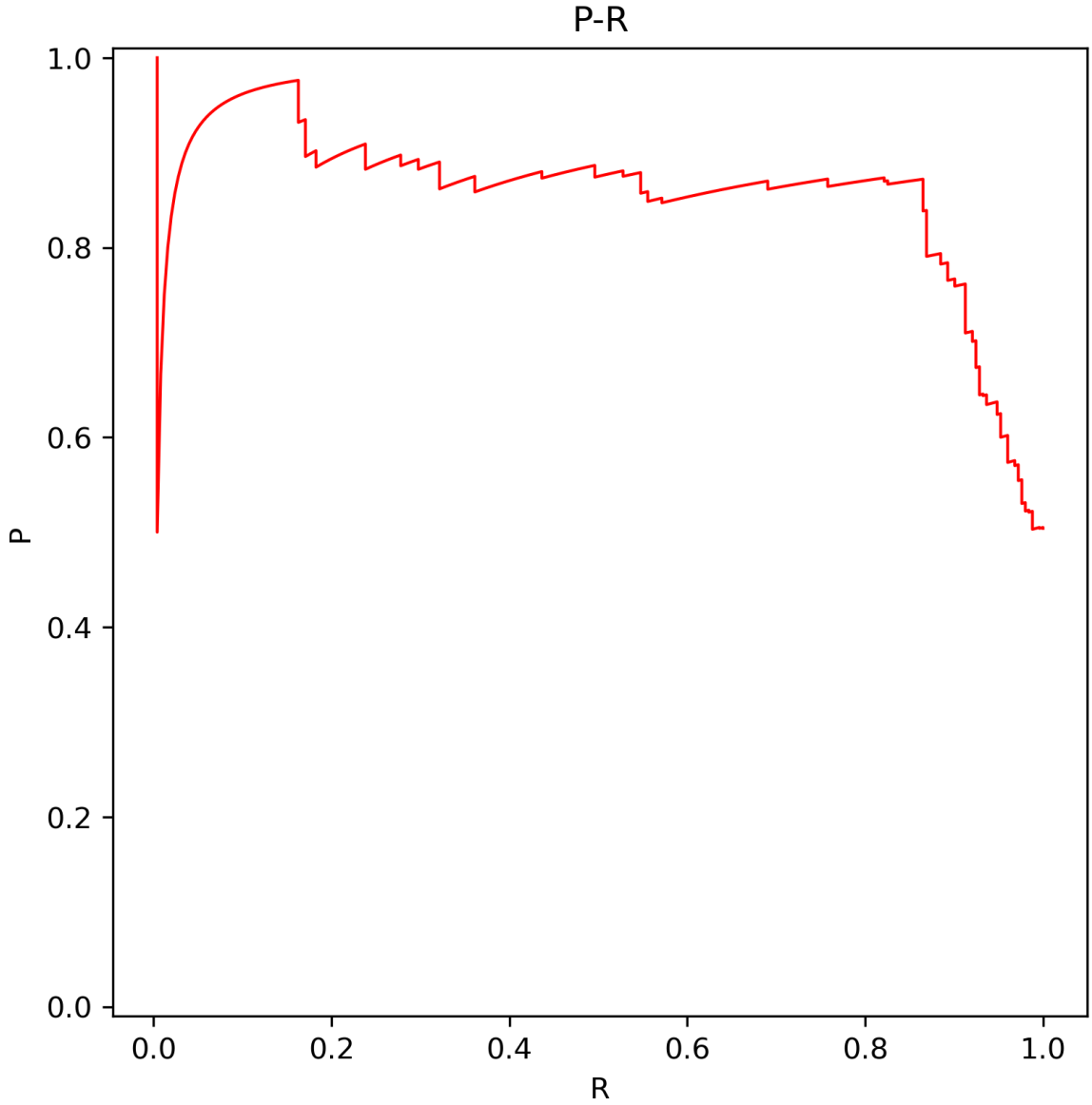
附加1

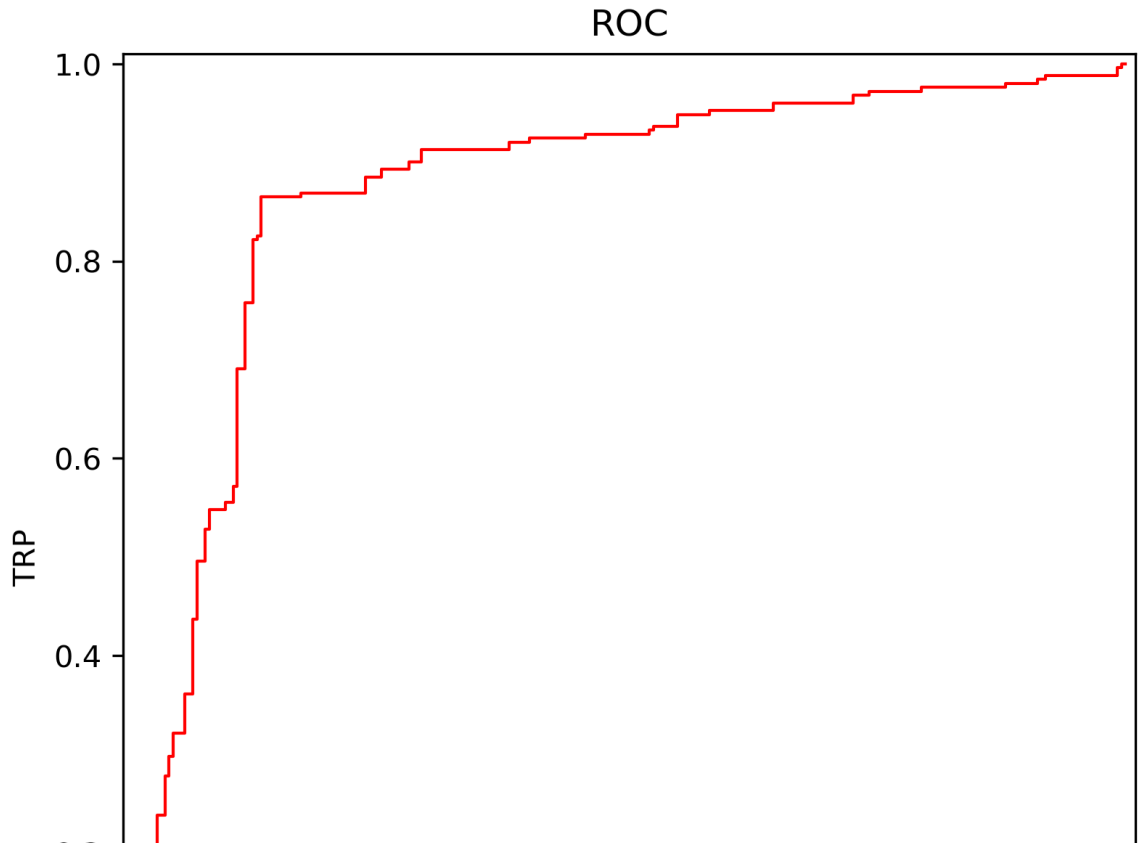
$AUC=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-)))$
考虑ROC曲线构造的过程
每次下调阈值的时候有两种情况

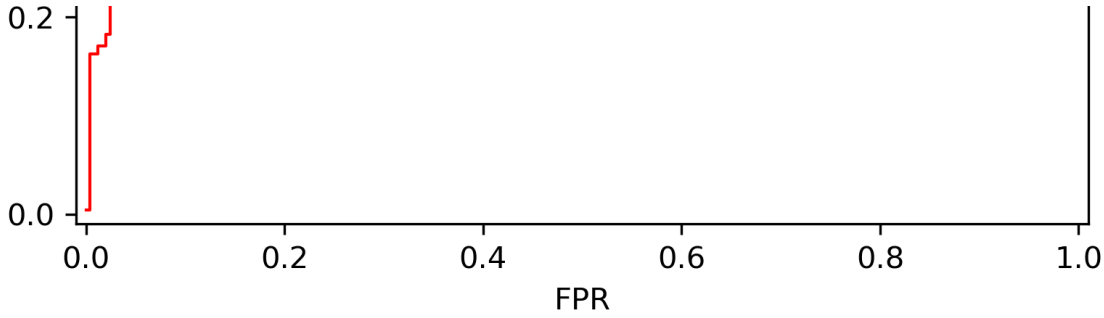
- 1. 一个正例样本被认为成正例
- 2. 一个负例样本被认为成正例
  当为情况1时, 考虑对AUC的面积贡献.
  增加的高度为$\frac{1}{m^+}$, 对应的长度为$\frac{1}{m^-}\sum_{x^-\in D^-}\mathbb{I}(f(x^-)<f(x))$
  考虑到当存在$f(x^-)=f(x)$可能因为顺序问题而产生的的面积偏差, 我们在这里取平均, 即原来统计的为一个面积为$\frac{1}{m^+}*\frac{1}{m^-}$的矩形, 现在我们对半取三角形, 则这部分对面记得贡献为$\frac{1}{m^+m^-}\frac{1}{2}\sum_{x^-\in D^-}\mathbb{I}(f(x)=f(x^-))$
  则枚举每一个正例为x, 求出对应的贡献, 即
  $S=\frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}(\mathbb{I}(f(x^+)>f(x^-))+\frac{1}{2}\mathbb{I}(f(x^+)=f(x^-)))$
  则原式成立

附加2

$AUC = 0.873720$







使用pandas库中的read_csv来读取数据集, 并用内置的排序功能对output进行降序排序.
然后枚举阈值, 分别算出 TP 和FP的数值, 放到对应的P, R list中, 和TPR FPR的数值放到对应的TPR FPR的list中, 同时利用AUC += (FPR[i] - FPR[i - 1]) * (TPR[i] + TPR[i - 1])来计算AUC
绘图用matplotlib.pyplot来绘图, 并保存.

关于重复数据 output相同的label都相同, 不需要特殊处理.

```
import numpy as np
import matplotlib.pyplot as plt
from pandas import read_csv

filename = 'F:\\杂\\大二下\\机器学习\\data.csv'
names = ['Index', 'label', 'output']
dataset = read_csv(filename, names=names)
dataset.sort_values(by="output",ascending=False,inplace=True) # inplace = True : 这一行全部变化
Size = len(dataset) - 1
Positive_Size = len(dataset[dataset.label == '1'])
False_Size = Size - Positive_Size
TP = 0
FP = 0
dataset_list = np.array(dataset[1:]).tolist()
TPR = []
FPR = []
P = []
R = []

AUC = 0.0

for i in range(Size):
    if dataset_list[i][1] == '1':
        TP += 1
    else:
        FP += 1
    R.append(TP / Positive_Size)
    P.append(TP / (TP + FP))
    TPR.append(TP / Positive_Size)
    FPR.append(FP / False_Size)
    if i > 0:
        AUC += (FPR[i] - FPR[i - 1]) * (TPR[i] + TPR[i - 1])

AUC *= 0.5

plt.figure(figsize=(6,6))
plt.plot(R,P,color="red",linewidth=1 )
plt.xlabel("R") #xlabel、ylabel：分别设置X、Y轴的标题文字。
plt.ylabel("P")
plt.title("P-R") # title：设置子图的标题。
```

```
plt.ylim(-0.01, 1.01)# xlim、ylim：分别设置X、Y轴的显示范围。
plt.savefig('P-R.png',dpi=300,bbox_inches='tight')
plt.show()

plt.figure(figsize=(6,6))
plt.plot(FPR,TPR,color="red",linewidth=1 )
plt.xlabel("FPR") #xlabel、ylabel：分别设置X、Y轴的标题文字。
plt.ylabel("TRP")
plt.title("ROC") # title：设置子图的标题。
plt.ylim(-0.01, 1.01)# xlim、ylim：分别设置X、Y轴的显示范围。
plt.xlim(-0.01, 1.01)# xlim、ylim：分别设置X、Y轴的显示范围。
plt.savefig('ROC.png',dpi=300,bbox_inches='tight')
plt.show()

print('AUC = %f' % AUC)
```