2.1

|  |  |  |
| --- | --- | --- |
|  | 预测正 | 预测反 |
| 实际正 | TP=200 | FN=100 |
| 实际反 | FP=200 | TN500 |

欠拟合

原因：模型较简单，特征数较少(精度、召回率低)

2.2

|  |  |  |
| --- | --- | --- |
| a | 成立 | 1）当阈值大于所有样本的预测值时，所有的样本都会被归为负类，这时正例数为0，因此FPR和TPR都为0，对应于ROC曲线的原点;当阈值小于所有样本的预测值时，所有的样本都会被归为正类，此时负例数为0，因此TN=FN=0，进而FPR和TPR都为1，对应于ROC曲线的(1,1)点。 |
| b | 成立 | 在给定样本分布的情况下，如果某个分类器A的ROC曲线高于另一个分类器B的ROC曲线，说明在相同假正例率的条件下，A的真正例率大于B，因此分类器A要优于分类器B。 |
| c | 成立 | 对于随机猜测的情况，被归为正例的样本中有一半是猜对的有一半是猜错的，被归为负例的样本也一样。  假设总共有  M个样本归为正类，N个样本归为负类，  则TN=FN=N/2，TP=FP=M/2  因此FNR=TNR=M/(M=N)  这对应于y=x 曲线。 |
| d | 成立 | 由于ROC曲线是由TPR和FPR直接决定的，而TPR和FPR都对分布不敏感，因此ROC曲线也对样本分布不敏感。这条性质说明，ROC曲线是对分类器本身性质的表征，与样本的分布无关。 |

2.3

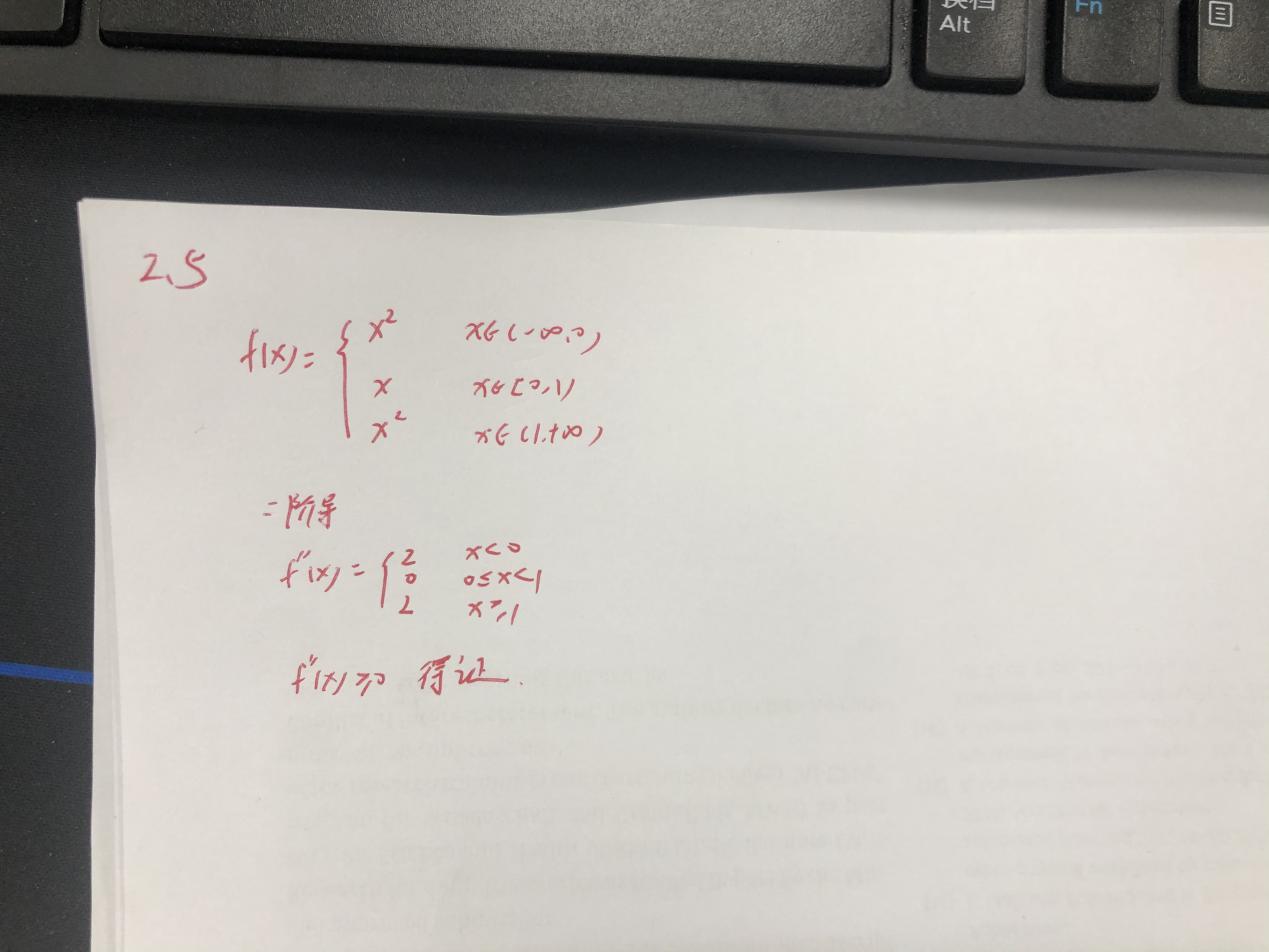
查全率: 真实正例被预测为正例的比例 真正例率: 真实正例被预测为正例的比例 显然查全率与真正例率是相等的。 查准率:预测为正例的实例中真实正例的比例 假正例率: 真实反例被预测为正例的比例 两者并没有直接的数值关系。

2.4

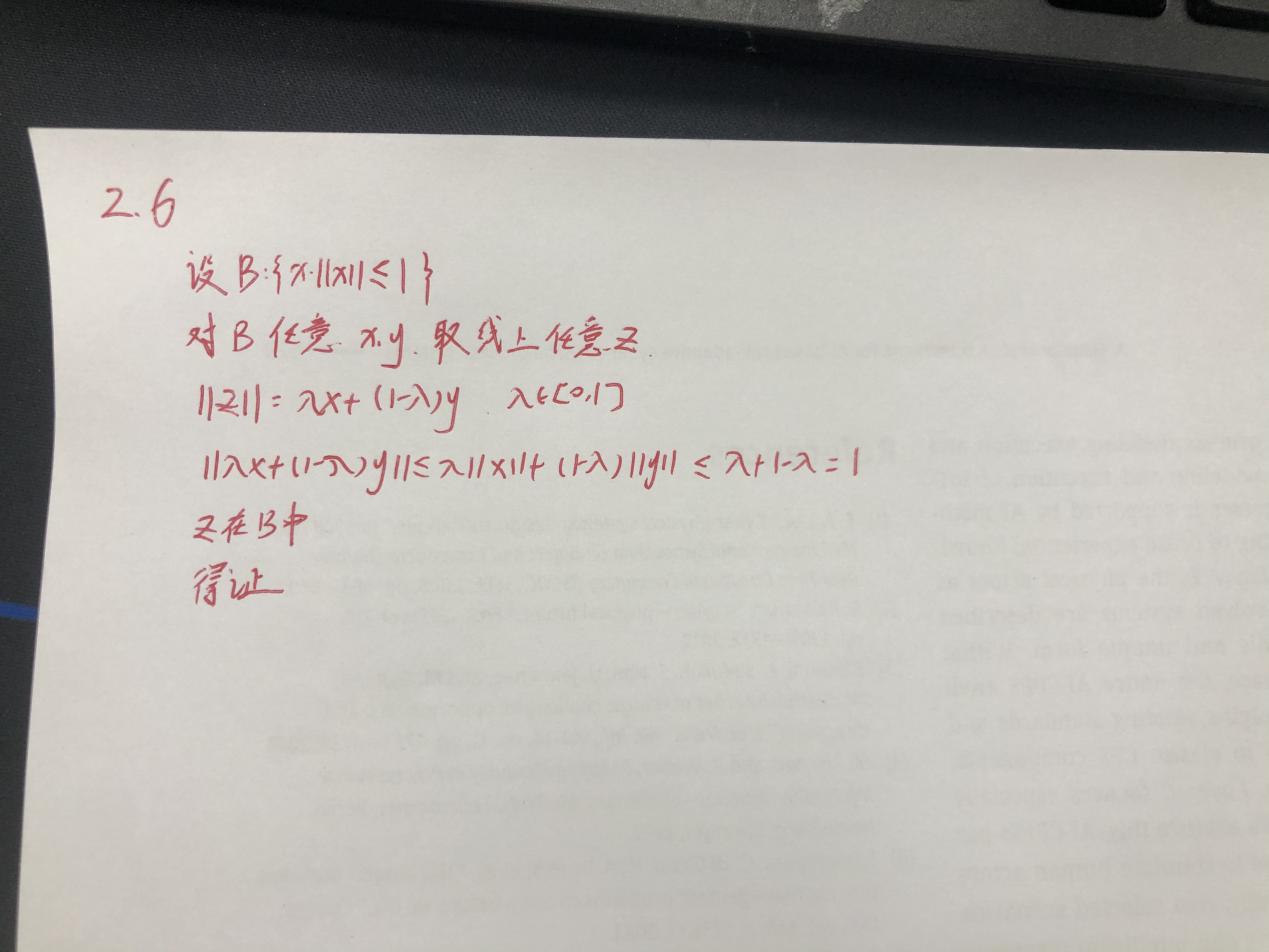
留一法：测试集1个样本，训练集99个样本，假如测试集真实情况为反例，则训练集包含50个正例和49个反例，所以测试集将被预测为正例，反之，将被预测为反例。错误率100%。

交叉验证法：按比例分层抽样的情况下，两类样本数量相同，预测结果等价于随机猜测。错误率50%。

2.5



2.6



2.7

,F1 = 0.6