第二次作业

Exercise 2.1

如果有个1000个样本的数据集，其中300个正例，700个反例，有个线性分类模型对这些样本进行分类，得到正例中有200个正确分类成正例，反例中有500个被分为反例，请画出此时模型的混淆矩阵？并分析此时模型是过拟合还是欠拟合？如果是欠拟合，请说明原因。

Exercise 2.2

试判断ROC曲线的下列性质是否成立，并给出原因

A. 任何ROC曲线必定经过原点和$(1,1)$
B. 位于上方的ROC曲线分类效果好于下方的ROC曲线
C. $y=x$曲线对应于随机猜测分类器在无穷多样本下的极限
D. ROC曲线对样本的分布不敏感

Exercise 2.3

试述真正例率(TPR)，真正例率(FPR)，查准率（P），查全率(R)之间的联系

Exercise 2.4

数据集包含100个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用10折交叉验证法和留一法分别对错误率进行评估所得的结果。

Exercise 2.5

证明： $f(x)=\max\{x^2,x\}$ 是凸函数。

Exercise 2.6

证明范数意义下的单位球是凸集

Exercise 2.7

根据下图数据，画图 P-R 曲线，并给出平衡点处的 F_1 值

Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1

Exercise 2.8

熟悉Python、NumPy、Pandas和Matplotlib