清华大学电子工程系

媒体与认知 课堂 2

2021-2022 学年春季学期

作业 4

程书鹏 2022 年 4 月 28 日

理论部分

1 单选题（15 分）

1.1 **D**

1.2 **C**

1.3 **B**

1.4 **A**

1.5 **D**

2 计算题（15 分）

2.1 **假设邮件粗略分为垃圾邮件和正常邮件，且存在一种垃圾邮件的检测方法，其中垃圾邮件被正确检测的概率为 a，正常邮件被误判为垃圾邮件的概率为 b。针对某一邮箱，所有邮件中垃圾邮件占的比例为 c，如果某封邮件被判定为垃圾邮件，根据贝叶斯定理，这封邮件是垃圾邮件的概率是多少？(提示：全概率公式 $P(Y) = \sum_{i=1}^{N} P(Y|X_i)P(X_i)$)**

解：令 A_1：是垃圾邮件. A_2：不是垃圾邮件 B：判定为垃圾邮件.

$\therefore P(A_1) = c$, $P(A_2) = 1-c$, $P(B|A_1) = a$, $P(B|A_2) = b$.

$$\therefore P(A_1|B) = \frac{P(A_1)\cdot P(B|A_1)}{P(B)} = \frac{P(A_1)\cdot P(B|A_1)}{P(A_1)\cdot P(B|A_1) + P(A_2)\cdot P(B|A_2)}$$

$$= \frac{ac}{ac + b(1-c)}$$

图 1: 2.1 解答

2.2 给定样本集合, 其均值为 $\mu = [1, 2]^T$, 样本协方差矩阵为 C, 且已知 $CU = U\lambda$。

其中 $U = \begin{bmatrix} 0.5 & -0.4 \\ 0.5 & 0.4 \end{bmatrix}$, $\lambda = \begin{bmatrix} 10.7 & 0 \\ 0 & 0.4 \end{bmatrix}$。

试用主成分分析 PCA 将样本 $x = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ 变换至一维。

(提示: 样本数据应减去均值；特征向量应归一化)

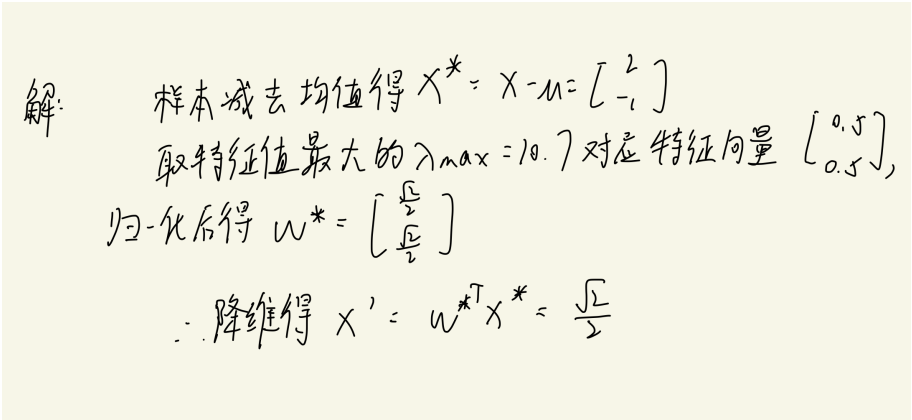

图 2: 2.2 解答

2.3 设有两类正态分布的样本集，第一类均值为 $\mu_1 = [1, 0]^T$，第二类均值为 $\mu_2 = [0, -1]^T$。两类样本集的协方差矩阵和出现的先验概率都相等：$\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 1.2 \end{bmatrix}$，$p(\omega_1) = p(\omega_2)$。试计算分类界面，并对特征向量 $x = [0.2, 0.5]^T$ 分类。

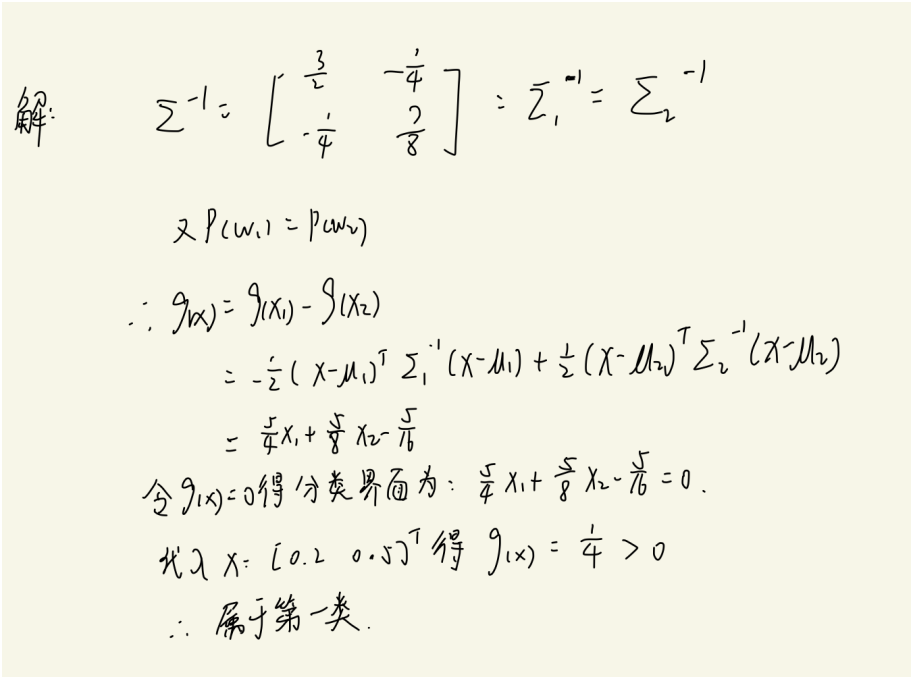

图 3: 2.3 解答

编程部分

3 编程作业报告（代码见附件）

3.1 hinge loss 测试结果

```
PS E:\Desktop\媒体与认知\第四次作业\hw4> python check.py
Linear successully tested!
Hinge successfully tested!
SVM_HINGE successfully tested!
```

图 4: 测试成功截图

3.2 训练、验证、可视化

3.2.1 使用 hinge loss 模拟 SVM

使用 hinge loss 模拟 SVM 的正确率为 92.8%

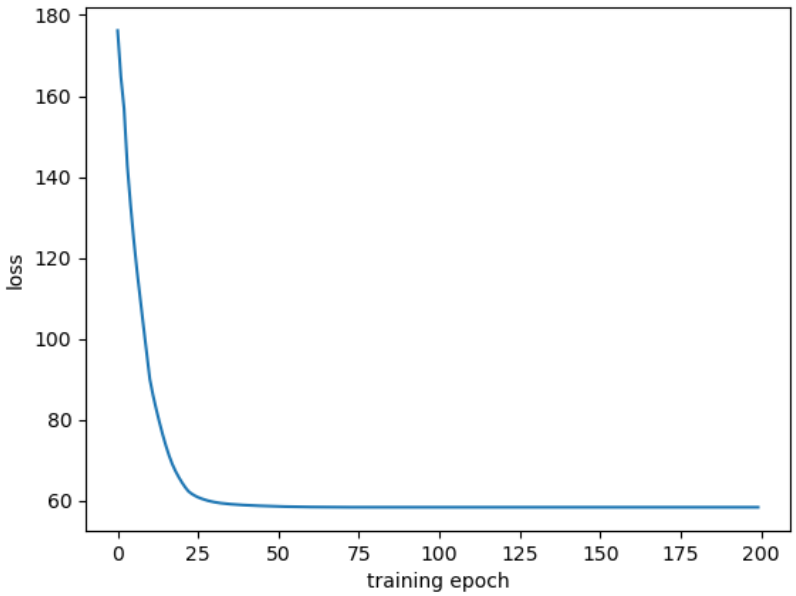

图 5: 使用 hinge loss 模拟 SVM 的 loss 曲线

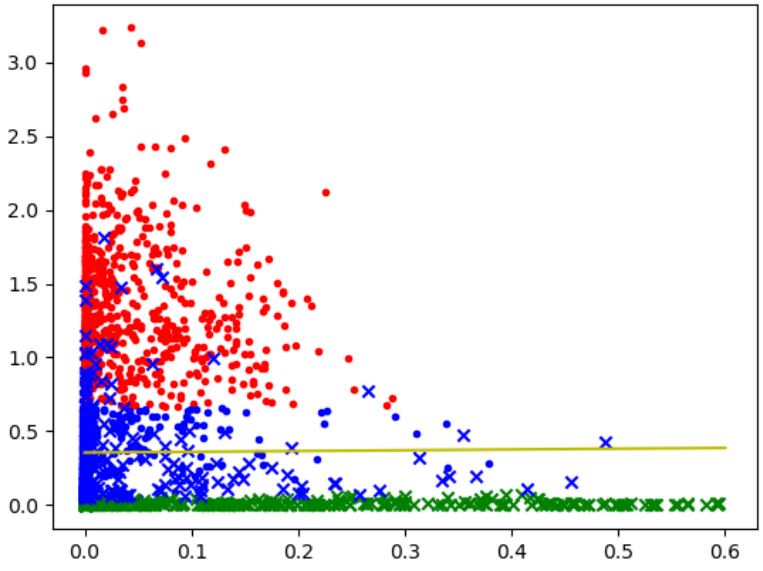

图 6: 使用 hinge loss 模拟 SVM 在训练集上的特征点分布图

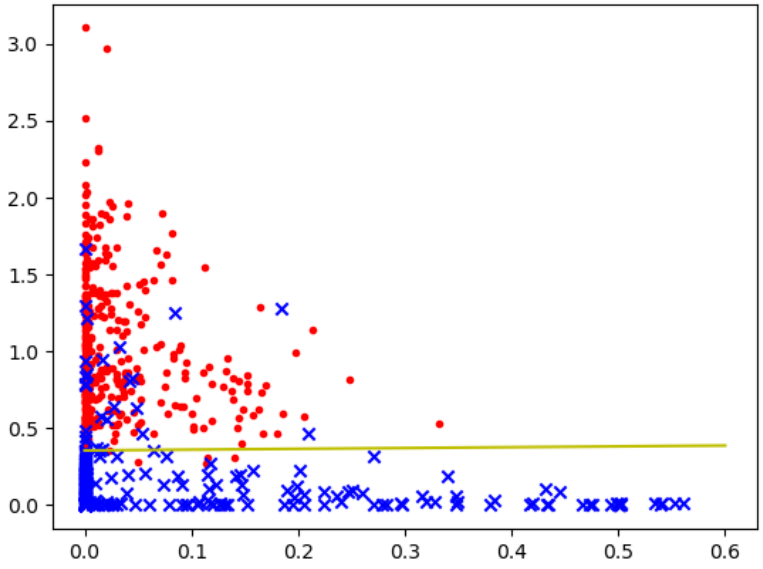

图 7: 使用 hinge loss 模拟 SVM 在验证集上的特征点分布图

3.2.2 使用 libsvm 库

使用 libsvm 库的正确率为 92.75%

```
PS E:\Desktop\媒体与认知\第四次作业\hw4> python classify_hw.py --mode baseline
*
optimization finished, #iter = 380
nu = 0.266243
obj = -58.385376, rho = 1.178906
nSV = 641, nBSV = 638
Total nSV = 641
Accuracy = 92.75% (742/800) (classification)
```

图 8: 使用 libsvm 库的正确率

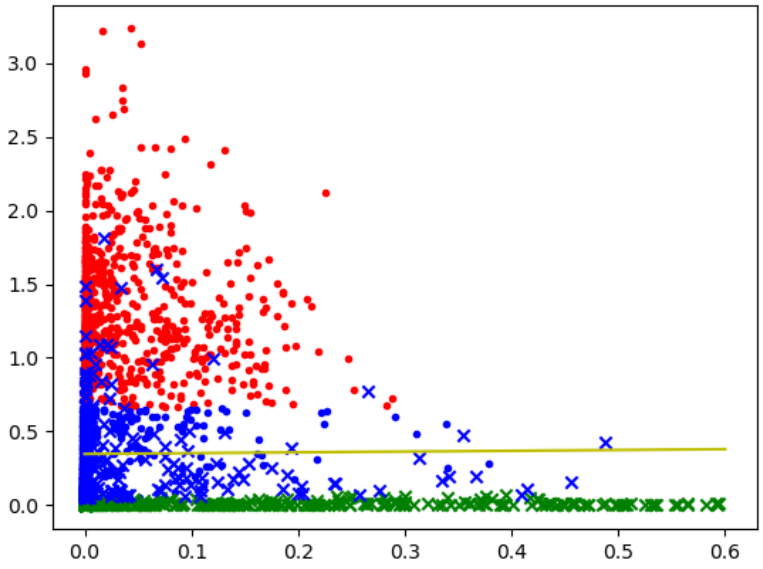

图 9: 使用 libsvm 库在训练集上的特征点分布图

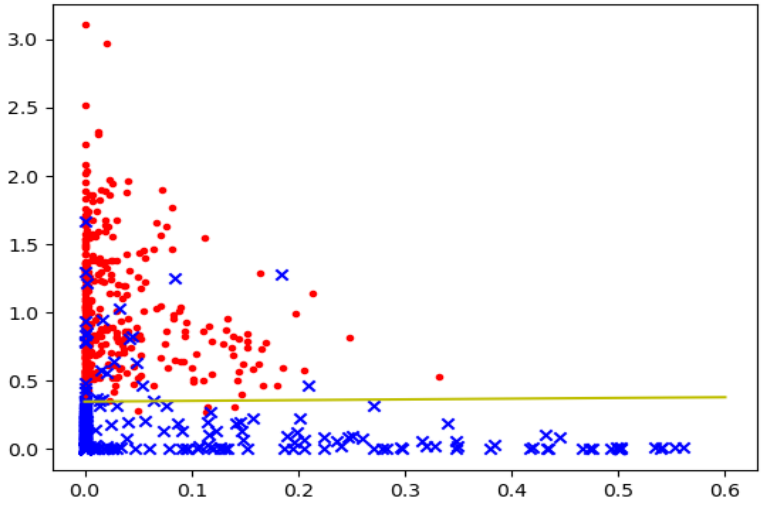

图 10: 使用 libsvm 库在验证集上的特征点分布图

综上，可以发现使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式在准确率、特征点分布图上基本上没有差别。

3.3 不同的正则化系数 C 对分类结果的影响

3.3.1 C=0.0001

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 50.0% 和 54.875%

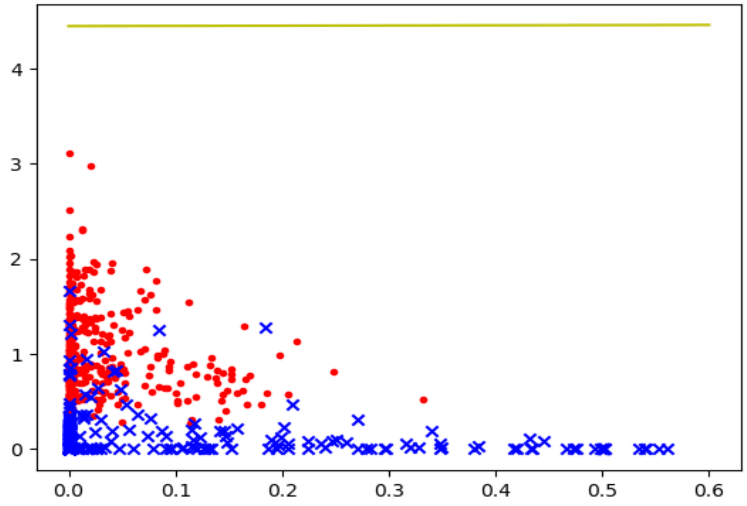

图 11: C=0.0001,hinge loss 在验证集上特征点分布图

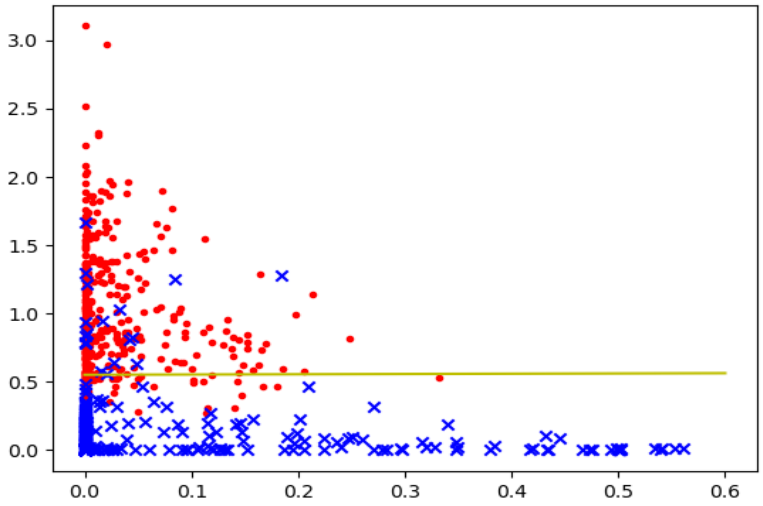

图 12: C=0.0001,libsvm 库在验证集上特征点分布图

3.3.2 C=0.001

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 68.6% 和 67.625%

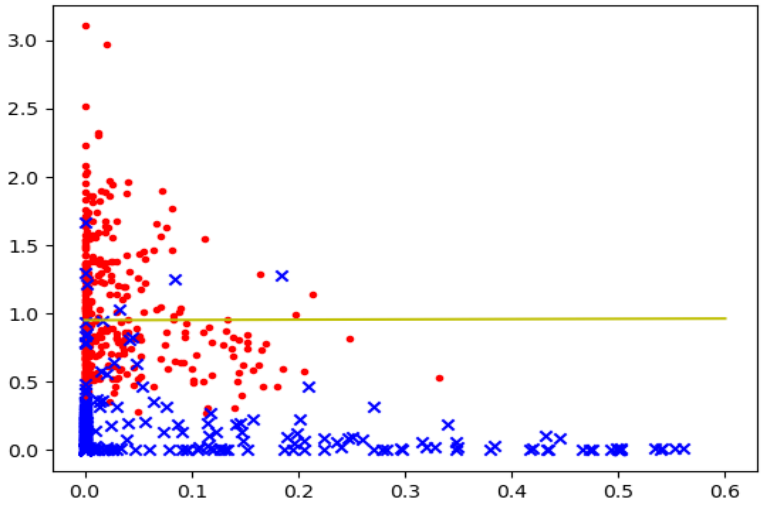

图 13: C=0.001,hinge loss 在验证集上特征点分布图

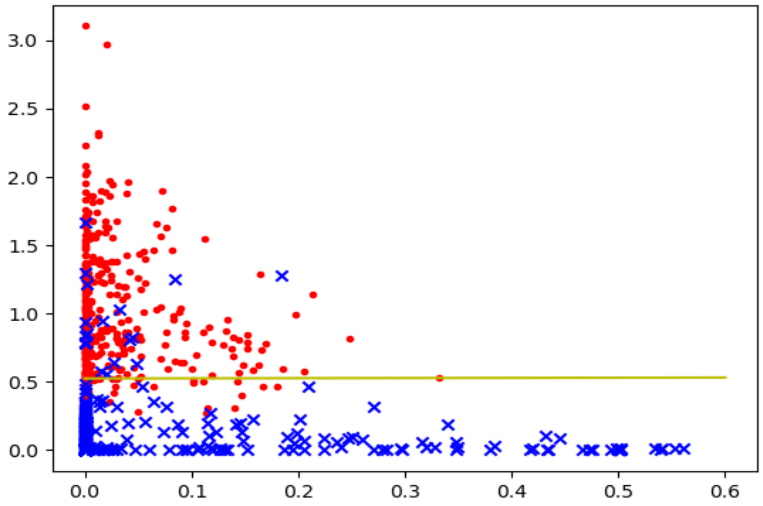

图 14: C=0.001,libsvm 库在验证集上特征点分布图

3.3.3 C=0.01

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 91.4% 和 91.375%

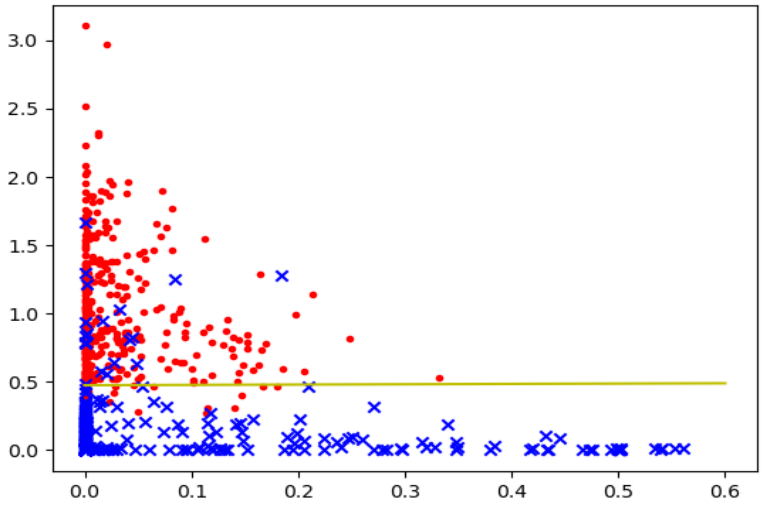

图 15: C=0.001,hinge loss 在验证集上特征点分布图

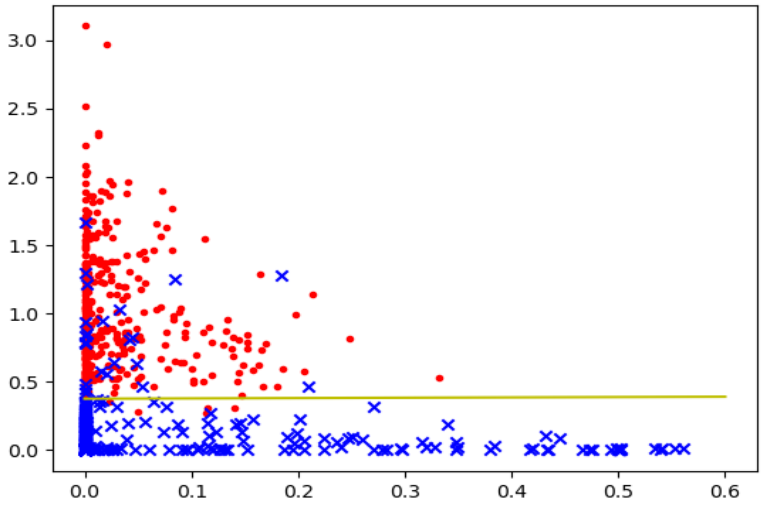

图 16: C=0.01,libsvm 库在验证集上特征点分布图

3.3.4 C=0.1

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 92.8% 和 92.75%

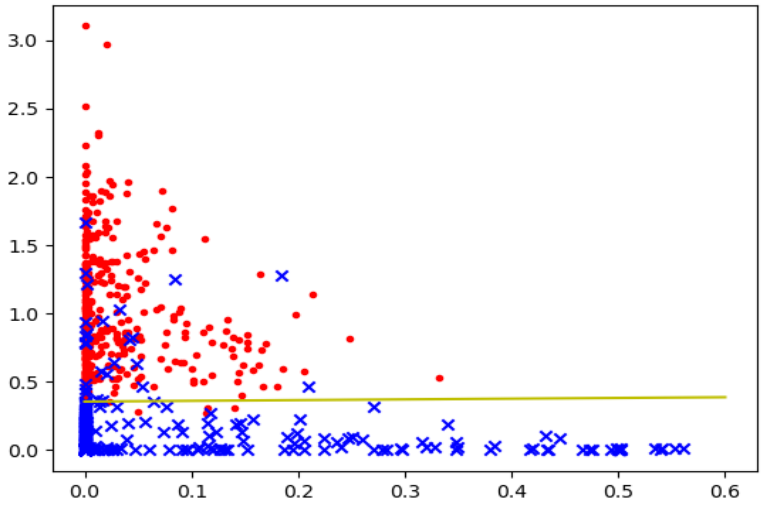

图 17: C=0.1,hinge loss 在验证集上特征点分布图

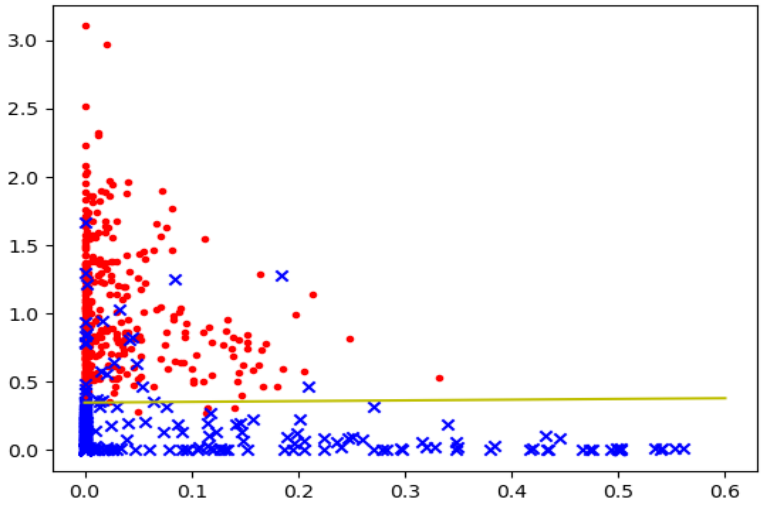

图 18: C=0.1,libsvm 库在验证集上特征点分布图

3.3.5 C=1

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 92.4% 和 92.375%

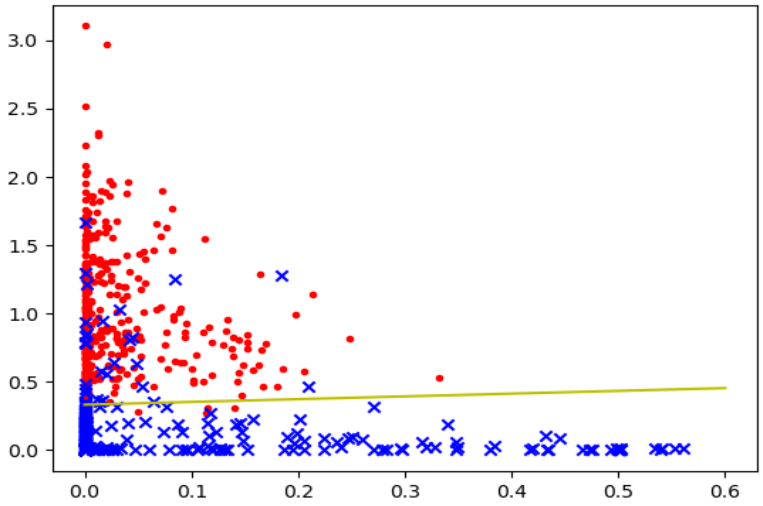

图 19: C=1,hinge loss 在验证集上特征点分布图

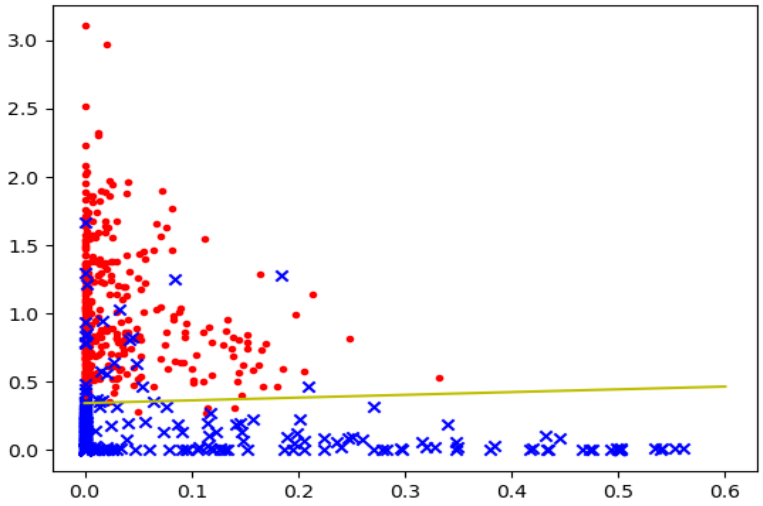

图 20: C=1,libsvm 库在验证集上特征点分布图

3.3.6 C=10

使用 hinge loss 模拟 SVM 和使用 libsvm 库两种模式的准确率分别为 92.4% 和 92.375%

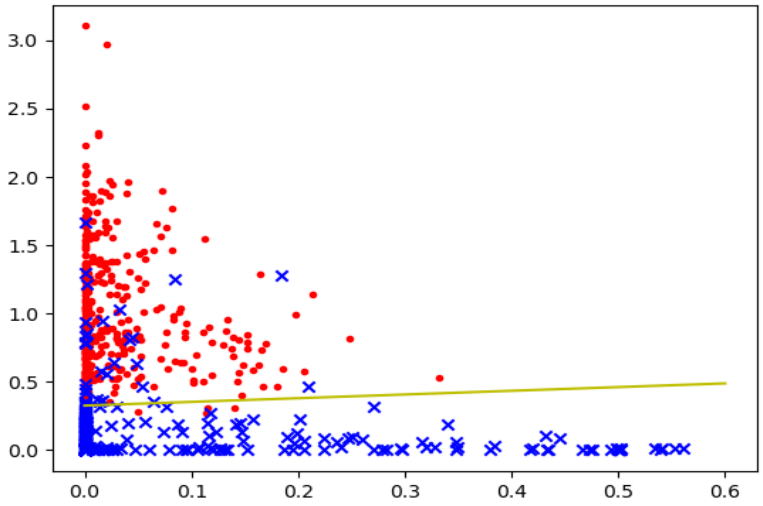

图 21: C=10,hinge loss 在验证集上特征点分布图

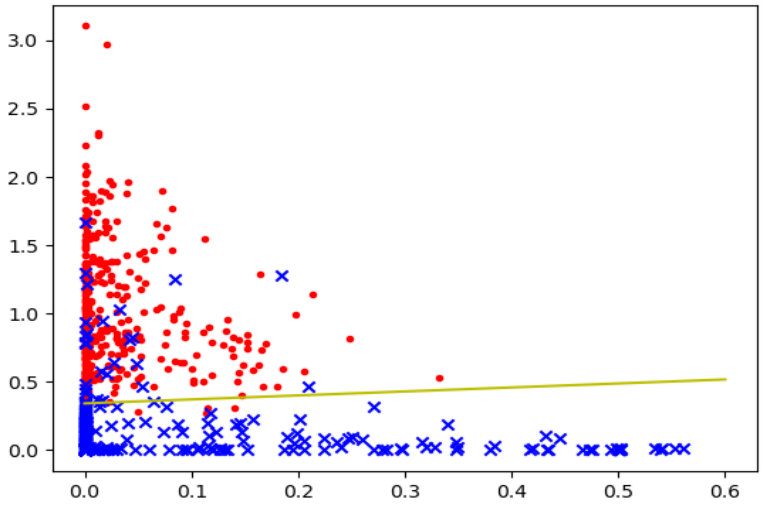

图 22: C=10,libsvm 库在验证集上特征点分布图

通过比较不同 C 值下两种模式的分类结果，可得出如下结论：1. 在 C 值很小时，分类极度不精确，这是因为模型能够容忍的错误分类太多，出现欠拟合，随着 C 增大，分类准确率总体上越来越高，但是在 C=0.1 左右达到最大，C 继续增大则不会使准确率更高，反而使其略有降低，这可能是因为模型对分类要求太严格从而出现过拟合。2. 在 C 值较小时，libsvm 库的分类效果比 hinge loss 好，在 C 值较大时，linge loss 的分类效果比 libsvm 库好。

4 本次作业遇到的问题及解决方法

本次作业进行得较为顺利，助教在习题课上的详细讲解和辅导给了我巨大的帮助。在前面的选择题和简答题中，由于有概率论等课程的数学基础，我完成得也比较顺利。最后，再次感谢助教的细致讲解和耐心辅导！