- 数据集：dataset3.txt　10 维特征，954 个样本（469 女、485 男）
- 数据集：dataset4.txt　10 维特征，328 个样本（78 女、250 男）

作业 2019-09-30. 用多种数据进行性别分类的实验（一）

- 用 dataset3 作为训练数据，用 dataset4 作为测试数据，采用不同的特征、训练样本数、分类方法进行比较实验，观察、分析实验结果的异同。
  - 要求采用的特征组合：a) 10 个特征都用；b) 任意选取其中两列特征（要在报告中说明是哪两个特征）。
  - 要求试验的训练样本：a) 从 dataset3 中任选 20 个训练样本（男女各 10 例）；b) dataset3 中的全部训练样本。
  - 要求试验的分类器方法：a) 最小错误率贝叶斯分类器（假设正态分布，先验概率各 50%）；b) Fisher 线性判别（FLD）；c) SVM（核函数自定）；d) 采用 BP 算法的 MLP 神经网络（网络结构自定）
- 对测试错误率用下表汇总实验结果

训练样本数	特征数	Bayes	FLD	Linear SVM	MLP
10+10	10				
	2				
469+485	10				
	2				

- 用所选出的两维特征画出样本的分布，设法在其中画出几组实验得到的分类边界（如画在一起不清楚可以画在多幅图上）。结合实验观察和对各种方法特点的理解，尝试对训练样本数、特征维数以及所选用的方法与测试结果的关系进行分析和讨论。
- 与选用了不同特征的同学进行讨论，比较使用不同特征的结果，尝试分析其中的原因和规律。
- 提示：在某些样本数和特征维数下，正常结果可能会相当不好，请注意鉴别和分析是实验错误还是正常情况，如是正常情况请尝试分析原因。

作业要求：

1、 交作业日期：2019 年 11 月 5 日（待定）前（含）打包提交。

2、 提交内容：（基本要求缺项或不符合要求要扣分）

a) 测试数据结果，按照助教要求的格式

b) 实验报告（PDF 文件，适当排版，以不超过 **5** 页为佳）

- 题目、姓名、学号、班级、日期
- 按照作业报告格式规范写作

C) 程序源代码

3、 关于编程和讨论：

鼓励自己写程序（用任何语言），也允许使用工具包，不禁止使用他人程序。在实验报告及程序报告中须明确写明程序出处和作者。但“程序运行步骤及参数”必须根据自己的实验情况自己完成。

对实验结果的分析，鼓励同学间讨论，但实验和报告必须独立完成。

如发现抄袭或未经说明的引用，或发现捏造数据（含使用往年数据），本次作业将记-15 分。如报告雷同但无法区分谁是原作者，则都按抄袭论处。

如在收集数据中或实验数据中发现捏造行为，则本次作业记-20 分。