

## 数据描述

该数据集有 train 和 test 分为两个文件，表示为两个  $N \times (d+1)$  的矩阵， $N$  为样本数， $d$  为特征数，矩阵最后一列是样本的类别（1 或 -1，即这是一个二类问题）。数据集有 10 个特征，其中 train 有 8285 个样本（因此矩阵有 8285 行，11 列），用于训练；test 有 2072 个样本。数据存为 .data 文件，可直接在 matlab 中读入，或用文本编辑器打开查看。

## 项目要求

- 请各自独立完成，鼓励讨论、交流，**严禁抄袭。一经发现，按 0 分处理。**
- 提交报告一份，该报告需至少包含以下内容：
  1. 个人信息（姓名、学号）
  2. 数据预处理方法（若未采用则可不写）
  3. 算法介绍
    - 采用的分类模型及其训练方法<sup>a</sup>
    - 分类器输出后期处理方法（若未采用可不写）
  4. 实验
    - 实验采用的分类器性能评价标准<sup>b</sup>
    - 整体实验方法及步骤<sup>c</sup>
    - 分类器训练算法的参数调整步骤
    - 实验结果（整理成标准的图表）

a 若使用了软件包，请简要介绍之；若完全为自己编程实现也请注明，并将源程序随报告一同提交。自己编程实现算法将获得加分。

b 使用多种评判标准将获得加分。

c 请注意 validation 和 testing 两个概念的区别，并比较 validation 和 testing 的结果是否一致。

## 注意事项及截止时间

- 所有报告均需通过电子邮件提交（jingze88@hotmail.com）。正文需为 pdf 文件，源程序请一律打包为 zip 文件。**邮件标题请采用“学号姓名模式识别作业”的格式。例如：SA11011xxx 王二模式识别作业。未按此标题提交而造成漏收邮件，责任自负。**
- 所有报告须在**期末考试开始之前**提交（以提交邮件时间为准），迟交的一律按 18 分处理，成绩统计时仍未提交的按 0 分处理。