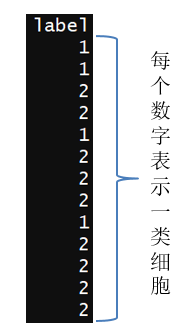
**《模式识别基础》期末大作业**

细胞是动植物的结构和功能的基本单位。人体大约由几十万亿个细胞组成，根据细胞的形态、功能等可以把细胞分为众多不同的类型，如生殖细胞、神经细胞等等。研究发现，即使是同一类的细胞，在形态、基因表达等方面仍存在着差异。为了探究细胞与细胞之间的差异及其原因，单细胞测序技术应运而生。关于单细胞研究技术的常识可以参见：https://en.wikipedia.org/wiki/Single\_cell\_sequencing。本次作业将使用单细胞基因表达数据对细胞分类问题进行探究。

单细胞基因表达数据如图1(a)所示，每一行为一个细胞，每一列为一个基因的表达值。我们可以把每一个细胞看做一个样本，把它对应的基因表达量作为特征向量。图1(b)则表示了每一个细胞的标签。



1. (b)

图1 单细胞基因表达数据和细胞标签

基于以上介绍的单细胞基因表达数据，请同学们研究以下问题。

1. 在文件夹“数据1”中，我们提供了两类细胞的单细胞RNA-seq数据。请对所给的测试数据（2ctrainX.txt和2ctrainY.txt）进行分析，设计分类器，并针对测试数据（2ctestX.txt）给出类别标签预测。
2. 在文件夹“数据2”中，我们提供了十类细胞的单细胞RNA-seq数据（10ctrainX.txt、10ctrainY.txt和10ctestX.txt）。请在该数据集下重复（1）中的问题，给出测试样本的类别标签预测。
3. 请对（2）中的十类细胞数据使用PCA或tSNE等方法进行降维和可视化，并对数据进行聚类分析（建议在降维后的空间进行聚类），比较聚类结果和真实标签之间的关系。

**评分**：

总分**100**分。报告占**60**分，请详细写出每一问的处理方法、原因和结果，并给出结论，我们会根据同学们报告的质量进行打分。

预测结果40分。其中问题（1）的测试集正确率占**15**分，（2）的测试集正确率占**25**分，评分标准如下。

得分=

其中最优准确率和最劣准确率分别是所有同学中的最高准率和最低准确率。

**说明**：

1. 数据已经区分了文件夹，文件名中2c指2 classes，10c指10 classes。
2. 对于给出的测试数据2ctestX.txt和10ctestX.txt，请使用你得到的最优分类器进行分类，然后将结果保存为文本文件，细胞标签顺序应与测试数据细胞顺序一致。标签请使用我们在题目中给出的细胞标签，以方便我们进行比对，以免出现比对出错给大家带来不必要的麻烦。提交的文件中请不要有文件头，即第一行就是第一个细胞的标签。
3. 如果聚类时电脑内存不够，可以随机抽取部分样本进行聚类分析。
4. 此次作业的截止时间是7月5日，所以时间充裕，希望大家多去尝试一些方法，开阔自己的眼界。