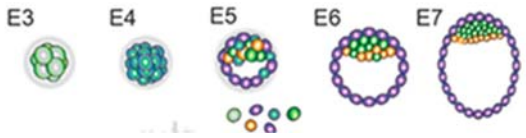
本次作业需要用到的数据文件：E3E5_data/

- 训练集：train_data_E3E5_2genes.txt
  train_data_E3E5_10genes.txt
- 测试集：test_data_E3E5_2genes.txt
  test_data_E3E5_10genes.txt

作业 2018-09-30. 贝叶斯分类器与线性分类器的实验

- 背景：从文献[Petropoulos et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell, 2016, 165(4): 1012-1026]中提取的人胚胎发育第 3 天（E3）与第 5 天（E5）的单细胞基因表达数据，根据不同的准则分别选取了 10 个基因和 2 个基因进行分类实验。



1. 用 2 个基因的数据进行贝叶斯分类实验，要求用训练样本集在假设正态分布的情况下估计概率密度函数，写出判别函数的显式表达式，分别假设 E3 和 E5 的先验概率为 1:1 和 1:5 得到两个分类器，将分类器分别应用到训练数据和测试数据上，分析错误情况。（注意：可能需要对原数据进行适当变换）
2. （选作）用 10 个基因的数据重复以上实验，考查可能遇到的问题并进行分析讨论。
3. 分别用 2 个基因的数据和 10 个基因的数据进行 Fisher 线性判别的分类实验，要求写出判别函数的显示表达式，按照自己认为合适的方式确定阈值。按两种实验设计进行测试：（1）用训练集进行训练，将所得分类器分别应用于训练集和测试集得到训练错误率和测试错误率；（2）用训练集进行 10 折交叉验证，得到交叉验证错误率，与前面两个错误率进行比较分析。
4. 分别用 2 个基因的数据和 10 个基因的数据进行感知器分类实验，考查训练过程中的训练错误率变化情况，如训练不收敛可采用步长缩减的方法强制收敛。按两种实验设计进行测试：（1）用训练集进行训练，将所得分类器分别应用于训练集和测试集得到训练错误率和测试错误率；（2）用训练集进行 10 折交叉验证，得到交叉验证错误率，与前面两个错误率进行比较分析。
5. 对于 2 个基因数据，画出两类训练样本和测试样本在二维平面上散点分布，将以上实验得到的分类边界画在图中，结合数据情况和方法原理对结果进行分析讨论。

- 交作业日期：2018 年 10 月 28 日（周日）前打包提交到网络学堂。
- 如发现抄袭、杜撰或未经说明的引用，或发现捏造结果，本次作业将记-10 到-20 分。如出现雷同报告但无法区分谁是原作者，则都按抄袭论处。

作业要求：

1、 提交内容：

a) 实验报告：PDF 文件，需适当排版，正文用 5 号字仿宋体单倍行距，不超过 3 页。正文无法包括的内容可以放在附件中，但正文必须保证在不阅读附件的情况下的完整性。作业成绩主要依据正文。

b) 附件：因篇幅限制无法放入正文但又需要介绍和讨论的细节内容可以用附件 PDF 文件提交，包括实验设计中的细节和所用程序的出处及参数设置等等，需提交实验结果的数据文件，如使用非开源程序则需提供程序源代码。附件所提供材料需保证实验内容能够完全重复。

2、 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，所提供细节必须足以完全重复实验结果。

欢迎同学间就作业相关内容进行讨论，但实验和报告必须独立完成，并在实验报告中对所参与的讨论进行说明（参加人及讨论内容）。

如发现抄袭、杜撰或未经说明的引用，或发现捏造结果，本次作业将记-10 到-20 分。如出现雷同报告但无法区分谁是原作者，则都按抄袭论处。