# 生物信息学概论 第二次作业

1. UCSC(University of California at Santa Cruz) 的 Cancer Genome Browser (https://genome-cancer.ucsc.edu/) 提供了大量的肿瘤数据，尤其是包含了不同肿瘤全基因组层次的多角度测序数据，含有已经经过原始数据预处理的 TCGA[1] (The Cancer Genome Atlas) 数据并提供免费下载。

   本次作业，我们从其中下载了乳腺癌细胞 500 多病人的基因芯片数据，已经经过预处理，在 GeneMatrix.txt 文件中记录若干病人的相关基因的表达数据。

   作业内容为：

   **利用 R 软件，进行该数据的聚类分析，注意 R 语言有 R markdown 格式，可以将代码及代码结果，以及文档整合在一个文件中，建议最后提交作业交 Rmd 文件。**

   - 利用层次聚类，对该组数据样本按照基因表达水平进行聚类，看聚类效果如何。即是否能够按照基因表达水平，将病人进行分类。距离可以选择 average。注，R 中有相应的聚类函数，请利用并尽可能输出图示 (如 heatmap)[2]，表明你的结果。
   - 实现 PCA，并利用你实现的 PCA 对该组数据的基因表达进行降维处理。请选择你认为合适的主成分数目，给出原因，再次对病人依据你给的特征进行聚类，并与 1 比较。

   文件说明：

   i. GeneMatrix.txt： 基因表达值文件，含有行名和列名，一行为一个基因，一列为一个病人

   ii. clinical_data: 记录了病人的若干信息，每一行为一个病人，病人的编号和 GeneMatrix.txt 中的相同。GeneMatrix 中病人只涵盖了这里的一部分，注意，在病人的若干描述中，有一项为 ER_Status_nature2012，可以根据这个对病人进行分类，你可以按照这个分类标准，对你的聚类进行一定的评估，看结果是否符合预期。

2. 给出第二组阅读的文献：**Principal components analysis corrects for stratification in genome-wide association studies** 中 Fig.1 的例子如何计算。(可以纸板也可以电子文档，纸板作业可上课时提交)。
3. 给出 Principal components analysis 的方法推导，注意从最大化方差及最小化信息损失两个角度，参考 Bishop 的 Pattern Recognition and Machine Learning[3] 第 12 章第一节。(作业提交方式同上)。
4. 附加题(如果完成，可以额外加分)。

   自学 EM 算法，推导混合高斯模型(GMM)求解的 EM 算法，并回答 K-means 与 GMM 模型的联系。

   可以参考文献：A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.

   K-means 与 GMM 的关系，可以参考 Bishop 的 Pattern Recognition and Machine Learning 第 9 章内容。

---

[1] TCGA (http://cancergenome.nih.gov/) 是 NIH 资助的癌症基因信息数据库，目前很多计算研究都在利用该数据。通常这里的原始数据不能直接利用，需要进行数据的预处理。

[2] R 自带的 heatmap，或者 ggplot 的 heatmap 作图。

[3] http://pan.baidu.com/s/1qWEdI2K 提取密码 jwtx;注意不要外传。