生物信息学概论第二次作业

1. UCSC(University of California at Santa Cruz)中的Cancer Genome Browser (https://genome-cancer.ucsc.edu/)平台提供了大量的肿瘤数据，尤其是包含了不同肿瘤的多层次组学数据并提供免费下载。本次作业，我们从其中下载了500多例病人乳腺癌的基因芯片数据（转录组）并已经过预处理，保存在GeneMatrix和clinical_data这两个文件中。

作业内容为：

利用**R**软件或其他数据分析语言，进行该数据的聚类分析。

1）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。

2）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。

数据文件说明：

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。

3. 附加题(此题非必需完成，如果完成，可额外加分)。

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。

备注：

[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;