CMSC 33750:
Machine Learning in Cancer

Lab Assignment 2

February 11, 2019
Version 1.0

This lab uses RNA-seq data from the NCI Genomic Data Commons (gdc.cancer.gov). You can download the data using the TCGAbiolinks R package, using the GenomicDataCommons R package, or directly using a curl command, as shown in Lecture 4.

The goal of this lab is to cluster some RNA-seq data from the GDC. In this lab you should become familiar with the GDC, its data, and get some hands-on experience clustering RNA-seq data, but you are not expected to do a comprehensive analysis of RNA-seq data, though, of course, the more thorough the analysis the better.

Please download the RNA-seq data from the TCGA Breast cancer project (TCGA-BRCA) and the TCGA Ovarian cancer project (TCGA-OV) and cluster both datasets separately. Please also visualize the clusters and include the visualizations of the clusters in your report. If you use R, you may want to consider using the limma package for RNA-seq clustering, but you can use whatever software you would like, including principal components analysis.

Please carefully write up your assignment, explaining:

1. How you prepared the data for clustering.
2. What algorithms that you used for clustering.
3. How you visualized the clustering.
4. What you noticed about the differences, if any, between clustering the breast cancer RNA-seq data and the ovarian cancer RNA-seq data.
5. How you convinced yourself that the clustering was valid.
6. What suggestions that you have for improving the clustering.

Please work individually.

Please prepare:

1. A written report addressing questions above.
2. Please annotate the code that you wrote to download the data, prepare the data, cluster the data, and visualize the clusters and include it in the report. Please be sure that the code is sufficiently annotated that a third party can understand it.

The lab must be turned in by midnight, **February 25, 2019**.