CMSC 33750:
Machine Learning in Cancer

Homework 2

October 8, 2020

This homework uses TCGA head and neck cancer (TCGA-HNSC) project data from the Genomic Data Commons (gdc.cancer.gov).

The goal of this lab is to cluster the TCGA-HNSC RNA-seq data from the GDC. In this lab, you should 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 also visualize the clusters and include the visualizations of the clusters in your report. You may use R, Python, or another language.

If you use R, you may want to look at the TCGAbiolinks R package or the GenomicDataCommons R package. 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 packages for computing principal components.

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. How you convinced yourself that the clustering was valid.
5. Any biological interpretations that you can associate with the clusters (extra credit).

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.

Homework 2 must be turned in by midnight, **October 16, 2020**.