CMSC 33750:
Machine Learning in Cancer

Homework 1

October 1, 2020
Version 1.0

This homework and the next homework uses RNA-seq data from the 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.

The goal of this homework is to download and explore some RNA-seq data from the GDC. In this homework (Homework 1), you should become familiar with the GDC, some of its data, and become familiar how to access the data. In the next homework (Homework 2), you will get some hands-on experience clustering RNA-seq data.

Please download some RNA-seq data from the TCGA head and neck cancer project (TCGA-HNSC), explore the data, and create a histogram of the counts (from at least one) of the samples.

Please write up your assignment, showing how you downloaded the data and the code to plot the histogram. In the next assignment, you will cluster the data.

Please work individually.

Homework 1 must be turned in by midnight, **October 8, 2020**.