Gene Expression EDA

Matt Oehler

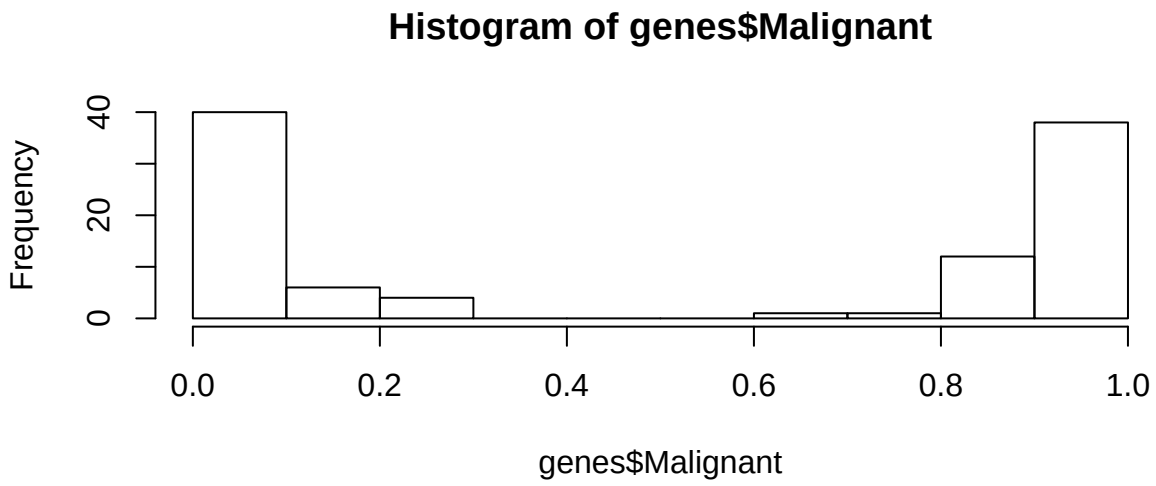
January 29, 2018

1. Motivation

Cancer is bad. The more that we can learn and understand about cancer, the more likely we will be able to find ways to fight it and/or prevent it. Through gene expression profiling we can hopefully more accurately understand types of tumors. Using the gene profiling data we hope to be able to identify which genes are associated with malignant tumors.

2. Data Exploration

The data consists of measures of 5150 genes of 102 cancer patients. Since there are so many covariates, creating a scatter-plot matrix of the whole data set is impractical. The covariates are all quantitative Below is a histogram of the malignance metric for all of the patients to help us understand the shape/spread of that variable. We can see that it isn't normally distributed, which might cause problems in the analysis if it isn't transformed or otherwise accounted for.



3. Statistical Method

I think that logistic regression would be a good model choice for this problem. We are interested in determining whether or not a tumor is malignant. Using logistic regression we should be able to determine the probability of a tumor being malignant given the observations of the various covariates. This will help us to address two research questions of interest: What makes tumors malignant, and given a tumor with certain features, what is the probability it is malignant.

4. Things I don't know

I don't really know how to visualize data that has more columns than observations (or really any dataset that has too many columns to make a pairs plot practical). I feel like that is going to cause some problems in the analysis because of matrix singularity.