Homework 3

Noah Simon

Jan 26, 2017

Please turn this in by Thursday, 1/26/17 by 5:00 pm.

Please submit on Canvas, in a compiled R-markdown file (to pdf or html).

All code in this assignment should be cleanly written and well commented, with appropriate use of functions/arguments. Imagine you need to give this code to someone else and they need to understand it (which you may need to do!)

High Throughput Screens and Predictive Models

The existence of a large number of necrotic cells in a tumor can be indicative of a successfully mounted immune defense. One might be interested in understanding biomolecular pathways regulated/dysregulated in a tumor that make it more/less susceptible to the immune system. By identifying genes with expression (in the tumor microenvironment) related to quantity of necrotic tumor tissue, we might hope to a) build a better picture of the biology of immune regulation/dysregulation in cancer and/or b) find potential targets for therapy. In this problem we would like investigate the relationship between gene-expression values in the tumor and the existence and extent of necrotic tissue.

To evaluate this we will again work with the **NOAH** data. The data can be found on the course website in the following files:

- **clinical_data.csv** contains the clinical/phenotypic information.
- **expression_data_probeID.csv** contains the expression information (by probeset).
- **annotation.csv** contains the genename identifiers corresponding to each probeset.

To get more information on the probesets, feel free to look into the affymetrix human 133 plus 2.0 array annotation (on the affymetrix site).

In particular, the variable **necrotic_cells.pct** (the percentage of necrotic tissue in a tumor found by pathology) may be useful.

Two potentially useful frameworks for addressing this question are the prediction-based, and screening-based ideas we discussed in the last two classes. Please be sure to properly evaluate over-optimism and/our account for multiplicity in your analyses. For the screening-based framework one should use a measure of association between continuous variables (as opposed to the difference of means that we used with a binary variable, in class)