BIOS 648: Final Project

Due 5pm Tuesday May 12, 2020

For the final project you can use any method(s) to analyze a high dimensional dataset of your choice, or **one** of the two datasets described below (posted on the course website in the folder "Final Project"):

1. Dataset of your choice: prior approval from the instructor is needed for this option. If you want to work on a dataset that is of interest to you, please upload to the course d2l website (Assignments folder titled "Final Project Proposal (optional)") a one-page brief description of the data and the objective of analysis by 5pm on Thursday 3/26. In the proposal please include a brief introduction to the background, the variables involved, the main hypothesis, and the sample size.

2. ACTG 384 HIV resistance data: download the file "HIVDATA.R", save it to your working directory, and read the data into R by the command

```
source("HIVDATA.R")
```

   You will get three data frames: `dat384`, `prmut0`, and `rtmut0`. Details of the data:

   (a) `dat384`:

Variable Name	Meaning
`patid`	Subject ID
`lrna0`	Baseline log viral load
`lrna1`	Last log viral load at or before week 52
`lrna2`	Last log viral load at or before week 104
`cd40`	Baseline CD4 count
`cd41`	Last CD4 count at or before week 52
`cd42`	Last CD4 count at or before week 104
`arm`	Treatment arm

   (b) `prmut0`: baseline mutation status in the protease (PR) region of the virus.

Variable Name	Meaning
`paitd`	Subject ID
`pr6`	Mutation status at PR codon 6 (0=wildtype; 1=mutated)
`pr7`	Mutation status at PR codon 7 (0=wildtype; 1=mutated)
···	··· ···
`pr99`	Mutation status at PR codon 99 (0=wildtype; 1=mutated)

   (c) `rtmut0`: baseline mutation status in the reverse transcriptase (RT) region of the virus.

Variable Name	Meaning
`paitd`	Subject ID
`rt38`	Mutation status at RT codon 38 (0=wildtype; 1=mutated)
`rt39`	Mutation status at RT codon 39 (0=wildtype; 1=mutated)
...	... ...
`rt240`	Mutation status at RT codon 240 (0=wildtype; 1=mutated)

The main objective is to use baseline data to predict outcomes (log viral load and CD4 count) by the end of year 1 and also by the end of year 2. Please see the two papers posted on the course website for more background information about this study.

3. Karyometry data: download the file “bladder.R”, save it to your working directory, and use the following command to read the data into R:

   ```
   source("bladder.R")
   ```

   You will get one data frame called `bladder`. This data frame has 94 variables (columns), including 92 karyometric features, named `f1`, `f2` through `f92`, patient id (`id`), and recurrence indicator (`group`). There are 40 patients, each one of them with around 100 nucleus images. Each row of the data frame includes the feature values extracted from image of one nucleus, so there are altogether about 4,000 rows in the data frame. The first 20 patients (with `group == "R"`) were recurrent, and the latter 20 patients (with `group == "NR"`) had no recurrence. The main goal of this project is to use the karyometric features to predict future recurrence. Please see the paper posted on the course website for more background information about this technology. Note that the response (recurrent or not) is measured on the patient level, not on the nucleus level. You could calculate some summary measures for the 100 values of each feature in the same patient to be used as predictors in constructing your classification model, or you could work on the nucleus level and then summarize across all nuclei of the same subject to arrive at a predicted class for the patient. Performance of your model should be measured by misclassification error calculated from cross-validation.

Please write a report about your findings. The report should include the following sections:

1. **Background:** provide an introduction to the problem. What is the question(s) of interest? What are the variables of interest? In what ways could the variables be potentially associated? State the hypothesis (or hypotheses) that you will investigate.

2. **Description of Data:** briefly describe the data.

3. **Statistical Methods:** describe the statistical methods that you use to answer your research questions.

4. **Results:** provide your results in a clear, concise form. Use tables and figures to summarize your results. Please don’t include raw output from the software. Please justify your models.

5. **Conclusions:** interpret and discuss your results. Provide a summary of the results and a take-home message. Please also state any limitations.