Mini-Kaggle Competition: Survival Time Prediction

June 16, 2015

In this challenge, you will receive pre-processed clinical and genomic data for patients with colorectal, bladder and prostate cancer. Each data file contains, for each patient in the study, the observed time-to-replace or censorsing time, an indicator for whether the patient's relapse was observed or not (1=observed, 0=censored), and a list of 541 binary covariates that represent a mutation or copy number alteration event in the corresponding gene. Each training dataset will have a testing dataset of patients with the same type of cancer. However, the testing data will not be revealed until the end of the challenge.

Your task is to come up with a supervised learning pipeline that ranks patients in the testing dataset in the correct ascending order of survival time. Specifically, we will use Harrell's concordance index [1, 2] to evaluate the quality of a solution.

References

[1] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[2] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.