02-620 Machine Learning Project
Proposal

Siddharth Reed
Department of Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213
slreed@andrew.cmu.edu

1 Introduction

1.1 Problem & Motivation

It goes without saying that cancer is an incredibly complex, variable disease, affecting the expression of many many genes. The catalysts of malignancy can also be quite variable, resulting from certain behavioural, environmental or genetic factors. As a result of this coming up with effective, personalized treatment strategies are also quite difficult since one needs to understand both the genes driving the disease and how to target them. There has been significant progress on treating patients to kill cancer cells and have them enter remission but with the driver mutations still laying the patients own cells it can re-emerge at any time. Having a better understanding of the genetic landscape and architecture of various cancers is critical to effective, timely and personalized treatments.

There is often too much information to be parsed by a physician to understand both if a patient is developing cancer and how to start treatment. Despite this scientists are still able to capture much of this information with genomic and transcriptomic data. And with the application of machine learning methods we can help learn and parse the important pieces to help direct doctors in treating people. Much of the noise can be parsed to help point doctors in the right direction and help decide on necessary follow-up tests and early treatments strategies, narrowing down much of what needs to be considered.

1.2 Data

I intend to use the copy number variation and transcriptomic data from TCGA to train classifiers to predict the type of cancer present in a sample. These data are available directly from TCGA or through the R package `recount` which does much of the pre-processing work to make sure that the read counts from samples are processed such that they are comparable. The `recount` package also has done much work to harmonize clinical annotation in the data for easier training and provides various kinds of clinical, treatment and survival data for samples in TCGA. More information is available here

1.3 Methods

A man, John, who goes to a job interview woefully unprepared and sees another man, Smith, waiting outside the interviewer's office. Smith shows John the lucky coin in his pocket, does his interview and finishes, confiding in John that he is sure he will get the job. John believes that the man with a coin in his pocket will get the job and does his interview. During the interview John in offered the job over Smith and accepts, upon leaving he notices that he in fact has a coin in his pocket. John is right that the man with a coin in his pocket will get the job but for completely incorrect reasons.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In our cancer problem a model we don't want to use is "black box solver" specifically for the cases where it can end up like John. He may get it right but why he get's it right is just as important if not more for these kinds of problems. It is able to accurately predict a pateint's disease given the appropriate sequencing data but provides no interpretable information about how the features (sequencing results) relate to the diagnosis. Also if the model is wrong the doctors using it will have no where to start "debugging" that diagnosis, figuring our what the patient's issue really is. Since I am working with biological data for diagnosis I want to use methods that are both accurate and informative.

Thankfully many classifiers are much more transparent than about what they are learning from the data and are much better for humans to interpret and dissect. For this goal I chose to use the following supervised learning methods

- Random Forest
- Support Vector Machine
- Naive Bayes

Random forests provide decision paths for each tree and you can look at how informative certain splits or conditions in the data are. They also provide the relative importance or features (genes) which can be hugely beneficial for follow experiments and treatments (especially has gene therapies become usable). Similarly SVMs provide support vectors, points (patients) that are nearest to the decision boundaries and can help physicians come up with decision rules for diagnoses. They can also provide confidence levels for their classifications, as a function of the distance of a newly classified patient to the decision boundary. Naive Bayes is generative and builds a distribution of the underlying data which is used for classification. The naive assumption can also be relaxed for known clusters of genes to potentially improve performance and the method scales well to having many (in the case $\approx 20,000$) features.