Capstone Project 1 Proposal:
Predicting clinical genetic variants that will have conflicting classifications by clinicians

Gurdeep Sullan
11/25/2019

Background:

Genetic testing is an important part of disease diagnosis and continues to grow in capacity (in terms of number of patients and detected diseases) as gene sequencing technologies develop. Patients can undergo genetic testing for numerous reasons, among which are preemptive screening for inherited diseases and diagnosis confirmation. In the first case, genetic testing is done when the patient has a family history of a disease and would like to know the likelihood of getting it in the future. In the latter case, a doctor sends a DNA sample to test for the presence of a disease-related mutation after the patient reports symptoms of disease. Thus, genetic testing provides results that have serious implications for guiding treatment and healthcare planning.

Once the decision to perform genetic testing is made, patients submit a sample of their DNA for testing against a large panel of known mutations, also known as variants. After their sample is run and variants are found, an in-house clinical geneticist makes the determination of where on a five-step scale the variant lies:

Benign	**Likely Benign**	**VUS**	**Likely Pathogenic**	**Pathogenic**

Where benign is an indication of no disease, VUS stands for a variant of uncertain significance, and pathogenic means disease is present. Many genetic variants, or mutations, are benign. Thus it is important to note here that ‘mutation’ is not synonymous with disease.

This classification is based upon a number of factors such as published literature and studies on that particular variant. There are multiple in silico predictors of deleteriousness, or harmfulness, based upon amino acid changes of a variant. However, these predictors are limited to protein-expressing genes and, by themselves, are not reliable for predicting disease. Some labs do use in silico predictors in conjunction with other resources when making their classification while others do not. Consequently, the same mutation can be classified as pathogenic by one lab and benign by another.

Problem:

A large problem in genetic testing is that some mutations are consistently classified, while other mutations receive conflicting classifications when tested at different laboratories. This project aims to address the following question: Can we predict whether or not a mutation will have conflicting classifications by two or more labs based upon the features of the mutation?

Client:

The first group of clients for this project are medical professionals who make the decisions on ordering genetic tests as well as communicate the results of a genetic test to a patient. The second client is the patient themselves, who make the final decision to test their DNA for mutations. The resultant model from this project can help both patients and caregivers decide whether or not to send out DNA samples to different labs when testing for a particular mutant. If they know the mutant of interest is likely to have conflicting classifications, it would be worthwhile to get additional labs' opinions before making major treatment or healthcare decisions.

Finally, a third group of clients is the community of genetic testing professionals. This model can reveal trends in mutants that are likely to be conflictingly classified, and may highlight patterns in classification processes that can be standardized.

Data:

I will be using a Kaggle dataset (https://www.kaggle.com/kevinarvai/clinvar-conflicting), originally taken from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). ClinVar is an online repository of genetic mutations and their features, updated in real time by submitters who submit their data to the database. Submitters to ClinVar include academic institutions, genetic testing laboratories, and hospitals.

The Kaggle dataset I am using includes data up until April 7th, 2018. Each of the 65,188 data entries has 46 features. All of the data points are genetic variants that have two or more classifications on the five-step scale outlined above. The 'CLASS' column contains a binary value for each data entry, either 0 = consistent classifications or 1 = conflicting classifications.

Approach:

First, I will upload the data in a pandas dataframe and complete exploratory data analysis in Python. I will use pandas to understand the structure of the dataframe and the data types in it. To visualize the data and identify interesting trends, I will use the matplotlib and seaborn packages.

Second, I will perform data cleaning. This includes identifying and replacing null and NaN values. I will convert columns to either categorical or numerical, with the exception being the column for disease indication. For this column, I will use NLP and tokenization to extract distinguishing keywords for each disease. Additionally, I will convert the necessary columns that are really bins, such as 'CHROM' (chromosome number), to categorical columns.

Thirdly, I will build a classification model to get a first-pass working model to improve upon. The model will take the necessary features as input and output either 0 or 1 for the 'CLASS' feature. I will be creating a hyperparameter table to track the hyperparameters of my model as I optimize it. I will perform MinMax scaling of features to be able to provide a normalized perspective on weights in the final model.

Deliverables:

A key deliverable will be the hyperparameter table, which will track the building of the predictive model. For clients in the genetic testing professional space, normalized weights of features along with descriptive statistics of the dataset will be presented separately. The final project with all of its code and Jupyter Notebooks will be shared on Github. All relevant reports and data will be included in the Github repository.