Machine Learning Engineer Nanodegree
Capstone Proposal

Bharat Mehan
26th July 2020

Personalized Medicine: Redefining Cancer Treatment
(Kaggle Competition)

Domain Background

A lot has been said during the past several years about how precision medicine and, more concretely, how genetic testing is going to disrupt the way diseases like cancer are treated.

But this is only partially happening due to the huge amount of manual work still required. Memorial Sloan Kettering Cancer Center(MSKCC) launched this competition, accepted by the NIPS 2017 Competition Track, because we they help to take personalized medicine to its full potential.

Problem Statement

Once sequenced, a cancer tumor can have thousands of genetic mutations. But the challenge is distinguishing the mutations that contribute to tumor growth (drivers) from the neutral mutations (passengers).

Currently this interpretation of genetic mutations is being done manually. This is a very time-consuming task where a clinical pathologist has to manually review and classify every single genetic mutation based on evidence from text-based clinical literature.

Hence, we need to develop a Machine Learning algorithm that, using this knowledge base as a baseline, automatically classifies genetic variations.

Datasets and Inputs

The datasets are provided by MSKCC on Kaggle competition website. They are free to download.

In order to get the dataset please create a login account to Kaggle and go to this problem statement page(given above) and download 2 datasets ***training_variants.zip*** and ***training_text.zip.***

In this competition we will develop algorithms to classify genetic mutations based on clinical evidence (text). And, there are nine different classes a genetic mutation can be classified on.

This is not a trivial task since interpreting clinical evidence is very challenging even for human specialists. Therefore, modelling the clinical evidence (text) will be critical for the success of our approach.

Both data sets are provided via two different files. One (training variants) provides the information about the genetic mutations, whereas the other (training text) provides the clinical evidence (text) that our human experts used to classify the genetic mutations. Both are linked via the ID field.

Therefore the genetic mutation (row) with ID=15 in the file **training_variants**, was classified using the clinical evidence (text) from the row with ID=15 in the file **training_text**

File descriptions

- **training_variants** - a comma separated file containing the description of the genetic mutations used for training. Fields are ID (the id of the row used to link the mutation to the clinical evidence), Gene (the gene where this genetic mutation is located), Variation (the amino acid change for this mutations), Class (1-9 the class this genetic mutation has been classified on)
- **training_text** - a double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations. Fields are ID (the id of the row used to link the clinical evidence to the genetic mutation), Text (the clinical evidence used to classify the genetic mutation).

Solution Statement

First we will analyze the data to gain clear understanding of the problem and the dataset. Following which we will be creating a training, validation and test datasets. We will also evaluate Gene, Variation and Text columns to gain a clear picture and relationships among the data.

Finalizing our data preparation and exploration we will move onto developing and training our models.

Benchmark Model

Text Classification – We will take Naïve Bayes as our benchmark model and will use other models to try and beat the performance of Naïve Bayes.

Evaluation Metrics

We will optimize and evaluate our solution against **Multi Class Log Loss** between the predicted probability and the observed target. For each ID in our test dataset, we must predict a probability for each of the different classes a genetic mutation can be classified on.

Project Design

Before training ML models, I will first take a deep dive into our data and see how it's formatted. Then I will start working on natural language processing and extract information such as character counts, sentence length, TF-IDF vector...etc. We will also do some graphical visualizations for better understanding of the data distribution.

I plan to compare 3 to 4 different models. Because this is a classification problem, a few options would be NB, Random Forest, SVM and KNN.

I expect to spend 60% of the time on data cleaning and natural language processing part and 40% of the time on training models and tweaking parameters. The final accuracy will be calculated against our test dataset

References:

1. https://www.kaggle.com/c/msk-redefining-cancer-treatment
2. NIPS 2017 Competition Track