Name: Deenu Khan

Email address: dr.deenukhan001@gmail.com

Contact number: 8800128247

Date: 02th Jan 2021

Case Study: Personalized Medicine: Redefining Cancer Treatment

Overview

This is one of the interesting real-world problems in the domain of medical science, first let's understand the problem here, 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 (it's simply alteration in a gene or genetic sequence) 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**.

Generally, we follow the below steps to classify the genetic mutations and these steps are majorly inspired by this link.

1. A molecular Pathologist (Just think of it as a domain expert) first selects the genetic mutation to analyze.
2. Now pathologists collect all the research and evidence related to that particular genetic mutation.
3. Now the most important and time-consuming part is, to study or analyze all those pieces of evidence and classify the genetic mutation.

Our objective here is to replace step 3 by the state of the art machine learning techniques, here we have given **9 classes of genetic mutations.**

In order to deal with the above problem **MSKCC (Memorial Sloan Kettering Cancer Center)** has provided the expert-annotated knowledge base where world-class researchers and oncologists have manually annotated thousands of mutations and we need to use this data to write an algorithm that can predict the class of the mutation.

Dataset Analysis

We are given data in two different files, one is for variations and one is for text which is nothing but evidence on which basis we are gonna predict the class of mutation.

training_variants: this file contains **3321 records** and each record is having ID, Gene, Variation, and Class values.

1. **ID:** It’s nothing but an ID number of Gene.
2. **Gene:** the gene where this genetic mutation is located
3. **Variation:** amino acid change for these mutations
4. **Class:** 1-9 the class this genetic mutation has been classified on

training_text: this file contains **3321 records** and each record is having ID, text as features.

1. **ID:** It’s nothing but an ID number of Gene. and this will be used to map the text feature in the training_variants file.
2. **text:** This text feature is having clinical evidence in the form of text to prove the class of mutation.

testing_variants and testing_text files are for testing purposes on Kaggle.

Performance Metric

Here, we will be using **Multiclass LogLoss** as a performance metric, we need to **predict the probability** for each class of 9 given class and submit the **CSV file for the evaluation**.

First Cut Approach

1. First I will be doing some EDA and trying to find Null values n all.
2. Now as we have two data files one for variants and one for text evidence, I will be **joining these two files based on the ID column.**
3. Now, when I will be having my joined dataset, it is now the turn of doing some **data cleaning** on textual data.
4. After doing all the cleaning, I will be using various techniques like **BoW, W2V,** etc. to convert these texts into vectors.
5. Once we have data to feed into the ML algorithm, then we will try **various ML algorithms** and try to find the best fit.