Machine Learning Engineer Nanodegree

Capstone Proposal

Ahmet Hamza Emra

July 28st, 2017

Proposal

Domain Background

Natural language processing (NLP) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora. Challenges in natural language processing frequently involve natural language understanding, natural language generation (frequently from formal, machine-readable logical forms), connecting language and machine perception, dialog systems, or some combination thereof.

Problem Statement

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. 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. MSKCC (Memorial Sloan Kettering Cancer Center) is making available an expert-annotated knowledge base where world-class researchers and oncologists have manually annotated thousands of mutations.

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

Datasets and Inputs

Data set is originally coming from kaggle.com. It was a competition.

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 aminoacid change for this mutations), Class (1-9 the class this genetic mutation has been classified on). Gene and Variation might be important and we need to consider them in our study. Hard to tell before the EDA.

training_text - a double pipe (||) delimited file that contains the clinical evidence (text) used to classify genetic mutations which is written by the experts of this area. 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

Our goal is classifying the class of genetic mutation, so we will use supervised learning approach to solve this problem. We believe text from experts has very valuable information so we will try to find the words that are key to classify the class. Ones we transfer text data to some kind of number or vector we will try to combine them with gene and variation variable then fit them to Classification algorithm.

Benchmark Model

We will be comparing our predictions with the actual values that are provide in the training data. The test will have considered successful if the separated test data's prediction has %80 accuracy.

Evaluation Metrics

We are going to be using the accuracy score to measure the performance of our model. Since it is classification problem, accuracy is suitably.

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

which is basically Percentage of how may predictions are correct.

Project Design

I will be following this steps in order to solve this problem.

1. EDA for Gene and Variation to check if there is new and useful feature
2. Text processing
   - Cleaning the text files with stemmer and stop words
   - Count vectorization to convert texts into vectors.
   - Tfidf transformer in order to find term frequency or inverse term frequency. we will be checking both of them to find optimal one.

1 Classification, we are going to check coupe algorithm (Naive Bayes, Logistic Regression and some ensemble models) to find the best score.

Concussion with performance metrics.

Reference:

https://en.wikipedia.org/wiki/Natural_language_processing
https://www.kaggle.com/c/msk-redefining-cancer-treatment
http://scikit-learn.org/stable/python
http://www.nltk.org