Personalized Medicine: Redefining Cancer Treatment

Machine Learning Capstone Project Proposal

1 Domain Background

The genetic testing is crucial us to understanding more about cancer, while the research has been slow due to significant amount of manual work, Therefore the Memorial Sloan Kettering Cancer Center (MSKCC) proposed this competition on Kaggle as one of the item of the NIPS 2017 Competition Track [1].

2 Problem Statement

There are many genetic mutations inside a cancer tumor, but only some mutations (drivers) contributing to the growth of the tumor. Currently, those drivers are captured manually. It is a really time-consuming task as the clinical pathologist need to classify every single genetic mutation based on evidence from the text-based clinical literature. Therefore it will be a natural language plus a classification problem.

3 Datasets and Inputs

The dataset contains 3 features as shown below.

Feature	**Data type**	**Description**
Clinical Evidence	Text	The clinical evidence used to classify the genetic mutations.
Gene	Text / Enum	The gene where the genetic mutation is located.
Variation	Text / Enum	The amino acid change for this mutations

Table 1. The description of 3 different features of the dataset.

The dataset is expert-annotated and published by MSKCC. Those features are included because those are the evidence used by the experts. There are 3321 training samples and 5668 testing samples for tuning the model.

4 Solution Statement

The problem is to find out the class of genetic mutation based on the given clinical evidence. It can be done by a supervised learning model with the help of natural language processing techniques. As all features are text, they should be first encoded into real number vectors for training a machine learning model.

5 Benchmark Model

Since the objective of this competition is to find a faster way to replace this manual process, one of the possible benchmark model would be the human. In terms of accuracy, The machine learning model may not perform better than a expert, but the model should be able to predict faster.

There are many benchmark models in the Kaggle Leaderboard. The best result so far is 0.42453 loss/error.

6 Evaluation Metrics

The performance of the model is evaluated by Multi Class Log Loss [2], the lower the best.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

Figure 1. The Multi Class Log Loss equation.

Since the prediction will be a probability distribution across class 1 to class 9 genetic mutations, the Multi Class Log Loss equation calculates the difference between the true label y and the prediction p of all N samples with M classes.

7 Project Design

7.1 Exploratory Data Analysis

Exploratory Data Analysis will be conducted for observing the importances of the three features and getting a sense to pick encoders for them.

7.2 Model Construction & Selection

The whole pipeline consist of a text to vector encoder, and a supervised learning model.

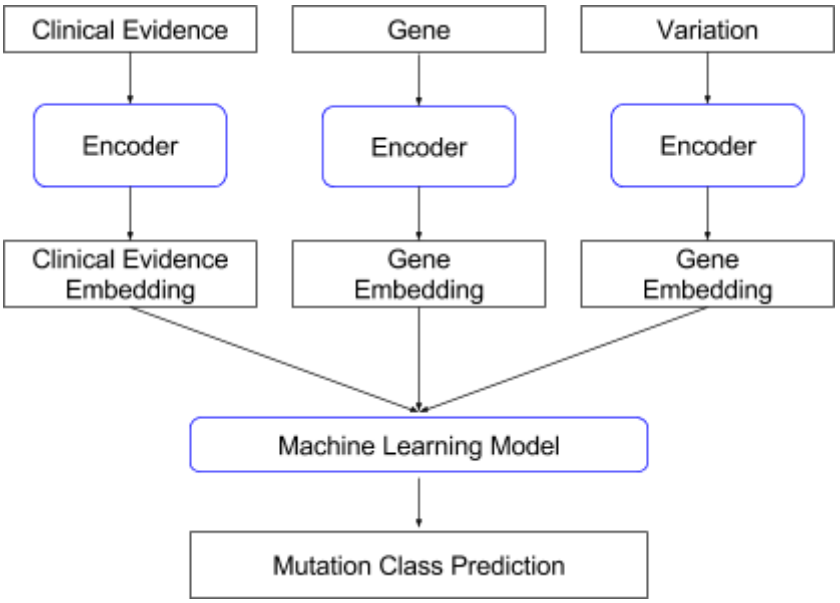

Figure 2. The proposed overall model architecture.

7.2.1 Data preprocessing

Since all 3 features are text, they have to be converted into vectors. These can be done using NLP techniques, such as word embedding [3] or terms frequency [4].

7.2.2 Machine Learning Model

The machine learning model should be a classification model. Different machine learning models will be tried including deep learning models.

7.2.3 Model Selection

K-fold cross validation [5] will be used for selecting models.

7.3 Model Evaluation and Kaggle Submission

The final model will be evaluated by both the given testing set and the Kaggle Leaderboard.

Reference

[1]"NIPS Competition Track", Nips.cc, 2017. [Online]. Available: https://nips.cc/Conferences/2017/CompetitionTrack. [Accessed: 12- Sep- 2017].

[2]"Multi Class Log Loss | Kaggle", Kaggle.com, 2017. [Online]. Available: https://www.kaggle.com/wiki/MultiClassLogLoss. [Accessed: 12- Sep- 2017].

[3]"Vector Representations of Words | TensorFlow", TensorFlow, 2017. [Online]. Available: https://www.tensorflow.org/tutorials/word2vec. [Accessed: 12- Sep- 2017].

[4]"Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining", Tfidf.com, 2017. [Online]. Available: http://www.tfidf.com/. [Accessed: 12- Sep- 2017].

[5]"3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.19.0 documentation", Scikit-learn.org, 2017. [Online]. Available: http://scikit-learn.org/stable/modules/cross_validation.html. [Accessed: 12- Sep- 2017].