Machine Learning Engineer Nanodegree

Capstone Proposal

David Ascencios (davidascencios20@gmail.com)

April 5th, 2017

Proposal: Mechanisms of Action (MoA) Prediction

Domain Background

This project is inspired by a (current) competition in the data science platform Kaggle [1]. As stated in the competition website, this data challenge was presented by The Connectivity Map, which is a project within the Broad Institute of MIT and Harvard [2], the Laboratory for Innovation Science at Harvard (LISH) [3], and the NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS) [4]. This project has generated a library containing over 1 million of gene expression profiles, using a relatively inexpensive and rapid high-throughput gene expression profiling technology. This competition corresponds in a Health/Biology field, as the data that will be used corresponds to cell and gene information, and the main problem to solve is one oriented to drug discovery.

I decided to choose this as my final project because of my deep interest in Bioinformatics / Health since my undergrad days. Having some years of experience in Machine Learning, I am always looking for applications in Medicine or Biology, and the projects or papers I found on the topic are astonishing.

Problem Statement

With the continuous improvements in the technology applied to the field of Pharmaceutics and Biology, it is possible to understand a disease, and its response in the human body in a much more precise way than years ago. Such improvements include gene profiling [5] (which is basically the measurement of multiple genes at once) and cell viability [6] (the measure of the proportion of live and healthy cells). Having this information at hand, the experts can have a great overview of the situation of the human body when a pathogen is attacking

its system. In that way, *"scientists seek to identify a protein target associated with a disease and develop a molecule that can modulate that protein target"* [1]. To simplify and classify the whole biological activity occurring to the body, scientists label each group of these measurements as "mechanism-of-action". Taking these into account, a dataset of all these biological measurements is provided, with a high number of labels (or MoAs) associated with each scenario (or row). With all these information, is possible to create a model that can have the capacity to, given a group of measurements for a particular disease, it could be possible to assign one or more MoAs to it, giving it a guide to the scientists and new ideas to create an efficient drug.

This implies that we are treating a multilabel classification problem, and we will apply supervised learning as we have the true labels of a training set, as described in the next section.

Datasets and Inputs

The datasets are provided by the main Connectivity Map project on Kaggle competition website.

Description of the files:

- train_features.csv (875 features)- Features for the training set.
  Features g- signify gene expression data, and c- signify cell viability data (these features are numeric). Feature "cp_type" indicates samples treated with a compound (cp_vehicle) or with a control perturbation (ctrl_vehicle); control perturbations have no MoAs; "cp_time" and "cp_dose" indicate treatment duration (24, 48, 72 hours) and dose (high or low). These three last features can be considered as categories.
- train_targets_scored.csv (206 targets)- The binary MoA targets that are scored (or not), for every label.
- train_targets_nonscored.csv - Additional (optional) binary MoA responses for the training data. These are not predicted nor scored.
- test_features.csv - Features for the test data (that will be evaluated in the Kaggle leaderboard).

Solution Statement

This is a supervised learning problem, more specifically a multilabel classification task. The goal of this project is to develop an algorithm that labels each case (with specific biological information about gene expression and cell data) a as one or more MoA classes. For that, different models will be trained and will be evaluated in a KFolds manner (within the same training data), and also we will

take into account the score that the public leaderboard will give us when we generate prediction with the test features (this is an open competition, and for that reason we cannot see the private leaderboard with the 100% of the test dataset yet).

Benchmark Model

For this problem, the benchmark model will be a Random Forest model trained almost directly with the features at hand, and no further hyperparameter tuning to it. This will give poor results, which will be fixed with the use of more sophisticated models and techniques, a deeper feature engineering step and some fine tuning of hyperparameters.

Evaluation Metrics

The metric that I will use for the results that we will be getting by experimenting with our dataset will be the log-loss function. Also, we will evaluate two results:

- The Cross-Validated log-loss mean score with the training data, using 5-folds.
- The log-loss obtained in the public leaderboard, with the test data.

Because the public leaderboard has only the score for 25% of the total test data, and in order to not overfit our models, we will give this result a weight of 25%, and the CV score will have a weight of the 75%, so our final evaluation metric for each experiment will be 0.75*score_training_cv + 0.25*score_test_public.

Project Design

The steps of these projects will be the following:

1. The first step will be the Exploratory Data Analysis. As we have a lot of features, it will be important to check the distribution of these, check if there is a high correlation between any of them, validate for outlier values. Also, an analysis to the targets will be made, looking for any relations between the 206 classes.
2. After an initial EDA, some feature engineering will take place, creating features that could represent relations between features, encoding categorical variables, and applying techniques such as PCA or t-SNE for feature extraction. Outliers and null values will be treated properly.
3. After we have this set of features, several models will be fitted: XGBoost, LightGBM, CatBoost and a DNN. Several hyperparameters will be tested to achieve better results.

4. We will analyze and compare the scores obtained, and, if necessary, we will repeat the sequence of these steps to find new insights of the data, create new features and test new models looking for better results, Also, we will stack the results of the models obtained to see if it gives us a better result in the public leaderboard.

After the experimentation is finished, we will create a blog containing the steps of the experimentation made on these datasets and it will be shared publicly.

References

[1] Kaggle Competition: Mechanisms of Action (MoA) Prediction https://www.kaggle.com/c/lish-moa

[2] Broad Institute of MIT and Harvard https://www.broadinstitute.org/

[3] Laboratory for Innovation Science at Harvard (LISH) https://lish.harvard.edu/

[4] NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS) http://lincsproject.org/

[5] Gene expression profiling https://en.wikipedia.org/wiki/Gene_expression_profiling

[6] Synopsis of Cell Proliferation, Metabolic Status, and Cell Death https://www.cellsignal.com/contents/_/synopsis-of-cell-proliferation-metabolic-status-and-cell-death/cell-viability-and-survival#:~:text=Cell%20viability%20is%20a%20measure,as%20during%20a%20drug%20screen
.