CS589 Machine Learning - Fall 2020

Mini-Project Information

November 6th, 2020

About:

- The mini-project is worth 10 points towards the final score.
- The amount of time and effort you put in the project should be approximately similar to what allocated to one homework, a median of 20 hours.
- You will be free to choose the techniques you'll use to solve the problem.
- You will need to hit some thresholds to get points for each part of the challenge.
- Extra points (5pts max) will be given for innovative solutions or very complete solutions (exhaustive search over multiple models, etc)
- Alternatively, extra points will be given for high scores on the Kaggle challenge website.

Custom Mini-Project:

You may replace this challenge with a 'custom; mini-project, something aligned with your own interests; but it has to be ML-related and use some of the techniques we're discussing in the class; for reasons of fairness, it also cannot be used as part of any other UMass activity for which you're given credit (other courses or independent studies). Examples of restrictions on custom mini-project:

- Something you're doing for another class project is not acceptable;
- Something you're doing as an independent study is not acceptable;

However, something that you're doing for research with a UMass prof (or with anyone else) is acceptable if you're not already getting graded for it.

Challenge

The challenge you will be working with is **Mechanisms of Action (MoA) Prediction**. This challenge aims to advance drug development by predicting mechanisms-of-action (MoA) of new drugs. Samples of human cells were treated with different drugs. The cellular responses were analyzed, in terms of their similarities to known patterns in existing genomic databases. Gene expression and cell viability data is included in the dataset. Also provided are the MoA annotations for more than 5,000 drugs.

Task, as described in the challenge: *'As is customary, the dataset has been split into testing and training subsets. Hence, your task is to use the training dataset to develop an algorithm that automatically labels each case in the test set as one or more MoA classes. Note that since drugs can have multiple MoA annotations, the task is formally a multi-label classification problem.'*

Link: https://www.kaggle.com/c/lish-moa/overview

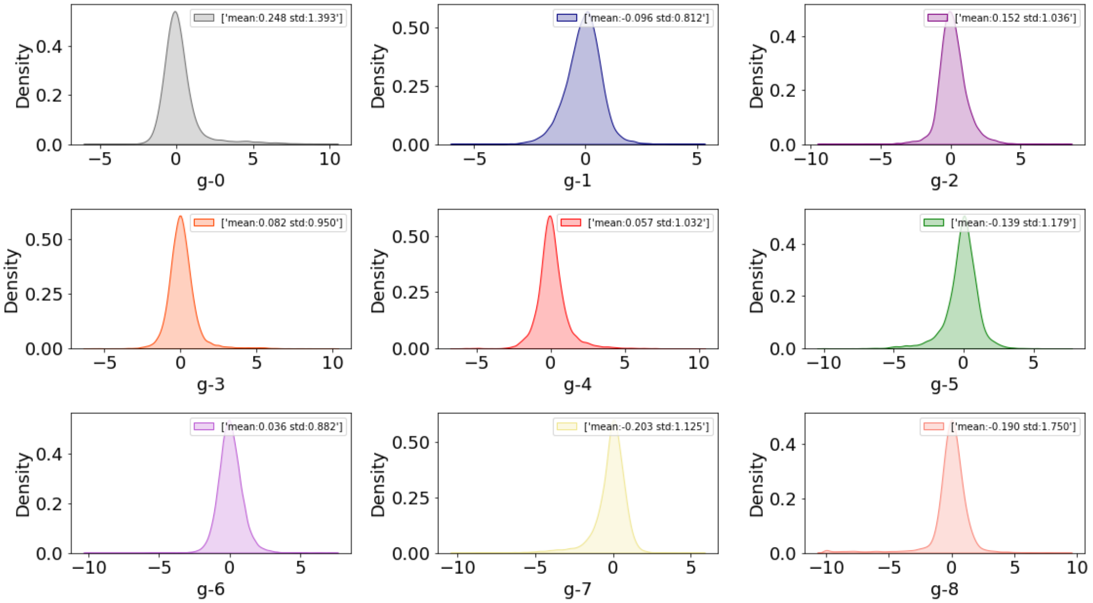
Data

This dataset contains 875 features including 3 categorical features and 872 real-valued features. There are 206 scored targets (binary classification) and 402 non-scored targets (binary, only for the training dataset) for the task. In the training dataset, there are 23,814 data points, and in the test dataset, there are 3,982 data points.

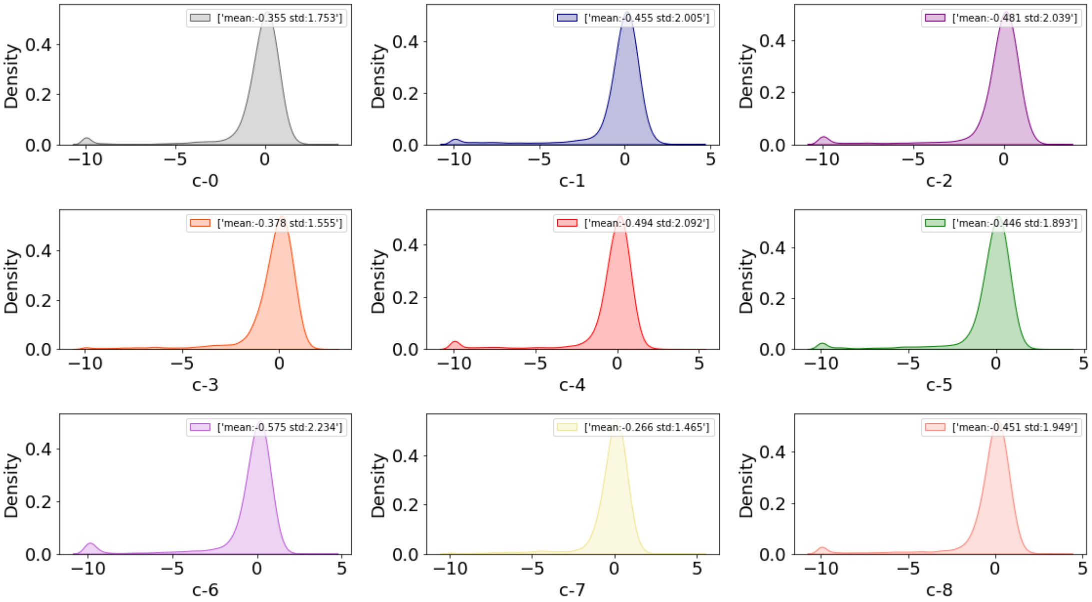
The 3 categorical features are *cp_type*, *cp_time*, *cp_dose*.

- The *cp_type* has two values: *cp_vehicle* for samples treated with a compound or *ctrl_vehicle* for treated with a control perturbation. Note: control perturbations have **no** MoAs (i.e. we expect no effect on the cells).
- The *cp_time* has three values (24, 48, 72 hours) indicating the duration of treatment.
- The *cp_dose* has two values (high or low) indicating the treatment dosage.

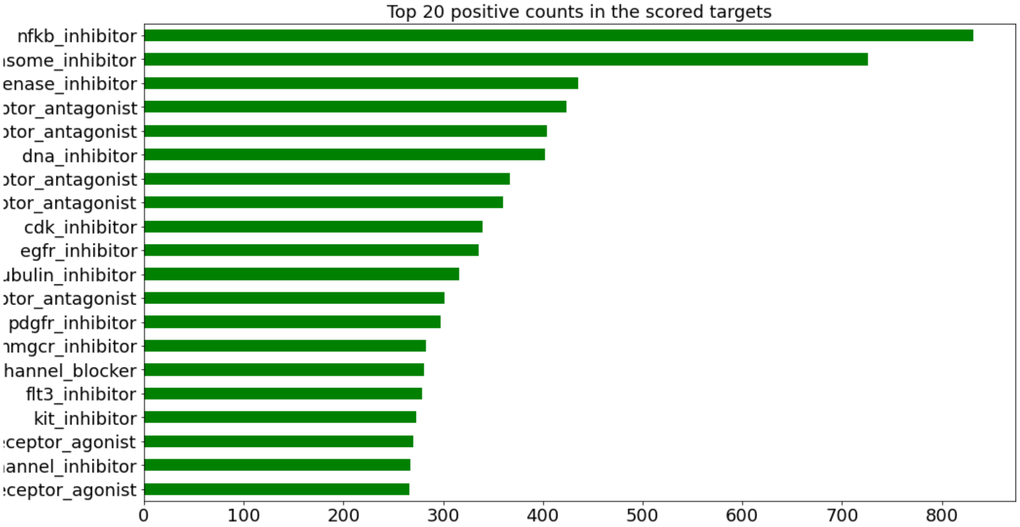
The 872 floating features include 772 gene expression data (labeled as *g-xx*) and 100 cell viability data (labeled as *c-xx*). Here the probability distributions for the first 9 **gene expression** features in the training dataset:

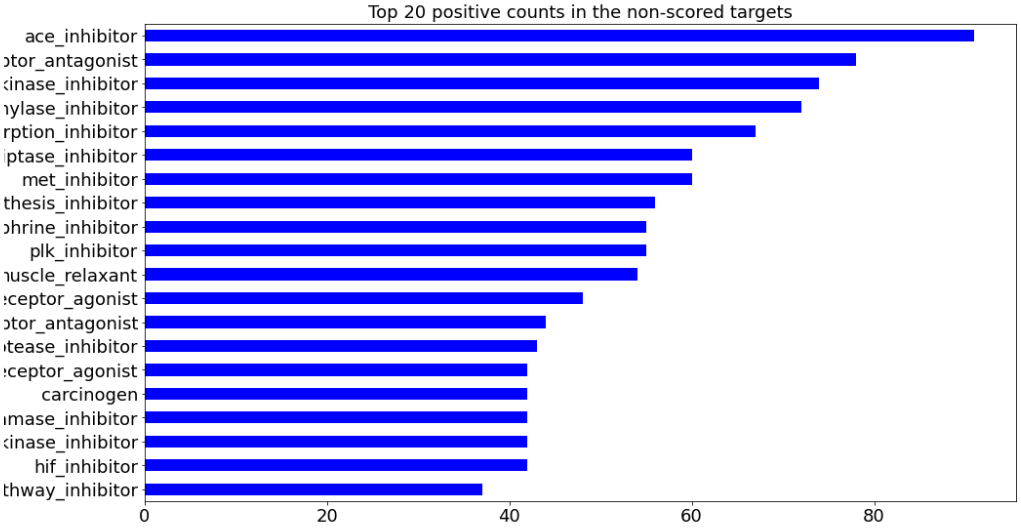


And here are the probability distributions for the first 9 **cell viability** features in the training dataset:



In the 206 binary targets, most of them are very imbalanced, and there are only a few positive samples in some targets. Here are the bar plots for the top 20 positive counts in the **scored targets** and **non-scored targets** of the training dataset:





Model

This challenge is a multi-task (multi-output) binary classification problem. Understanding and pre-processing the dataset is probably needed for a better performance in modeling the predictive task. You should explore **linear models**, **ensemble models**, **feature selections**, and **deep learning models**.

The evaluation metric used in this challenge is column-wise (206 scored targets) *average logarithmic loss.*

Because the dataset is very imbalanced, you may want to choose between multi-task learning and multiple single task learning.

Evaluation

To receive the full 10 points for the mini-project, you will need to address this challenge through some of the techniques you've learned this semester. You will write a report documenting your efforts. The report must include the following, in no particular order.

1. [2 points] Describe the scheme you will use to validate your model. You may include figures such as the ones shown the lecture on generalization and evaluation. Remember that you are not provided the test labels, as this is an active challenge. To get test scores for a challenge submission, you would need to submit your predictions via Kaggle - this is necessary to earn extra credit, but it's not a requirement for the completion of this project. Instead, to earn 'standard' credit, you'll need to select an appropriate way of testing your model, locally on your machine, using only the data you're provided in the 'training' files. You can either select a subset of it for testing or cross-validate. You'll need to report what we'll call the **local test score -** a surrogate for the test score that Kaggle would assign. The **local test score** you report should be an unbiased estimate of the Kaggle test score.

2. [1 point] Specify how you are handling the class imbalance and the sparse output labels.

3. [1 point] Describe any dimensionality reduction or feature selection you intend to do as a preprocessing step. Please remember that any feature selection needs to be performed on training data only.

4. [2 points] Try out a linear model. Describe any regularization you're using, how you are setting your hyperparameters and other details a reader would need to replicate your work. Report the local test score for this model.

5. [2 points] Try out an ensemble model. What are you using as the base classifiers? How are they combined in the ensemble? What are the pertinent values for the hyperparameters – ensemble size, for instance. Report the local test score.

6. [2 points] Try out deep learning for this task. What network architectures made the most sense here? Can you combine deep learning with some of the other models you've tried? Report the local test score for your best DL-enhanced model.

7. [2 points Extra Credit] Submit your solution to the challenge leader board and report your score. You may try several times and report your best performance.

8. [2 points Extra Credit] If you've scored in the top 1000, you get 2 extra points. You must include your Kaggle identifier and a screenshot of your placement.

9. [2 points Extra Credit] If you've scored in the top 100, you get 2 extra points (in addition to the points for the top 1000).

10. [2 points Extra Credit] If you've scored in the top 10, you get 2 extra points (in addition to the points for the top 100).

11. [2 points Extra Credit] If you've scored in the top 1, you get 30k and a (basically guaranteed) admission into our PhD program. :)

Good luck! Have fun!