Homework 10

A7: KAGGLE-MOA

Cheung Wai Chan
Seyedeh Elnaz Sadat Mansouri
Janar Aava

Task 1. Setting up:

https://github.com/aavajanar/KAGGLE-MOA

Task 2. Business Understanding:

Identifying your business goals:

- **Background**

In the past, scientists derived drugs from natural products or were inspired by traditional remedies. Today, drug discovery is a targeted approach based on understanding the biological mechanism of a disease. This modern approach involves identifying protein targets associated with a disease and finding a molecule with mechanism-of-actions (MoAs) to modulate those protein targets. However, researching a molecule's MoAs is very resource intensive. Decoding the mechanism of action (MoA) helps the development of new therapeutics and the evaluation of drug side effects.

- **Business goals**

Development of new computational methods for MoA would get faster drug delivery and development with more efficiency and low cost. Therefore, we will create a model to predict a molecule's MoAs based on the gene expression and cell viability patterns, which are relatively faster and cheaper to obtain. This model would significantly narrow down the potentially desired drugs to further study and develop.

- **Business success criteria**

The value of the logarithmic loss function on each drug-MoA annotation pair will be measured for this project. We can compare our results with the results of kaggle competition. Logarithmic loss function under 0.02 might be a good outcome which can help scientists in the drug discovery process.

Assessing your situation:

The dataset of 5,000+ drugs has their MoAs, gene expression and cell viability and it's publicly available on Kaggle. The drug names, gene expression and cell viability are anonymous to prevent identifying the specific drug. This makes it impossible to group genes or cell types but it's still enough information needed.

- **Inventory of resources**

We have people with data science, programming, medical and genetics knowledge. Data provided by Laboratory for Innovation Science at Harvard on Kaggle. In terms of hardware we have our personal computers and possible cloud services(like Colab) and for software we are using jupyter notebooks running Python 3.

- **Requirements, assumptions, and constraints**

This project must be completed by 14 Dec 12:00. The project will be considered completed when the poster and video presenting the results are done and submitted. The project should take roughly 90 hours divided between 3 people.

- **Risks and contingencies**

Given the project's small size there shouldn't be many risks that arise. The most likely risks will be related to lack of data science or domain knowledge, which could be solved by consulting the instructors, checking/starting discussions on Kaggle or analyzing notebooks of others on Kaggle.

- **Terminology**

- **Mechanism of Action (MoA)**
  specific biochemical interaction through which a drug substance produces its pharmacological effect. The MOA can occur on the cell membrane, within the cell, or outside the cell.
- **Cell viability**
  A measure of the proportion of live, healthy cells within a population.
- **Gene expression**
  It is a process by which the instructions in our DNA are converted into a functional product, such as a protein.

- **The costs and benefits**

The gene expression and cell viability data are collected with a new technology that measures all 100 different cell types in the same sample. This method to test new drugs is quick and cost effective compared to researching the MoA. If the data could accurately predict future potential drugs, it could improve the success rate of drug development phase 2 and 3, which cost $20 - $30 million. Overall success rate is 13.8%, hence a small improvement can make a big difference.

Defining your data-mining goals:

- **Data-mining goals**

Create models to predict a drug's mechanism of action based on gene expression and viability data and compare the results of different models to find which one is the best performing. Report those results in the form of a poster and a video.

- **Data-mining success criteria**

Based on the MoA annotations, the accuracy of solutions will be evaluated on the average value of the logarithmic loss function applied to each drug-MoA annotation pair. Reaching 0.02 log loss on the Kaggle competition would be acceptable for the scope of this project, but ideally we are aiming for 0.019 and beyond.

Task 3. Data Understanding:

Gathering data

- **Outline data requirements**

The list of necessary data to identify a MOAs based on the gene expression and cell viability patterns:

1. gene expression
2. cell viability
3. Type of sample (samples treated with a compound or with a control perturbation)
4. treatment duration (24, 48, 72 hours)
5. dose (high or low).

- **Verify data availability**

Experimental data are publicly available at the website (https://www.kaggle.com/c/lish-moa/data).

- **Define selection criteria**

Data selection was performed by the Laboratory for Innovation Science at Harvard and we will make use of all the data presented by them that we determine to help us get a better final result.

Describing data

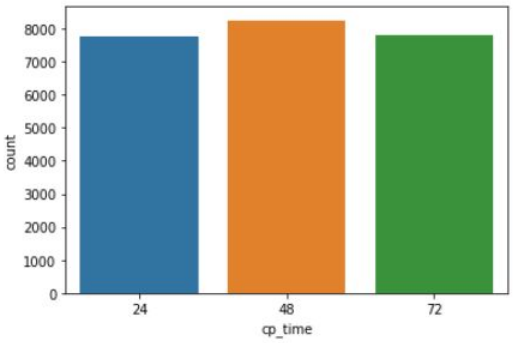
Data source is available in Kaggle website and it is a project within the Broad Institute of MIT and Harvard, the Laboratory for Innovation Science at Harvard (LISH), and the NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS), present this challenge with the goal of advancing drug development through improvements to MoA prediction algorithms. The dataset includes gene expression and cell viability data. The data is based on a new technology that measures simultaneously (within the same samples) human cells' responses to drugs in a pool of 100 different cell types (thus solving the problem of identifying ex-ante, which cell types are better suited for a given drug). In addition, the dataset contains MoA annotations for more than 5,000 drugs.

The train dataset includes 23814 values with 876 columns which are sig_id,co_type,cp_time,cp_dose, g(0:771),c(0:99). And the test dataset has 3982 values for predicting the probability of each scored MoA with 876 columns. Data types per dose, type and id are string and the other columns are numeric.

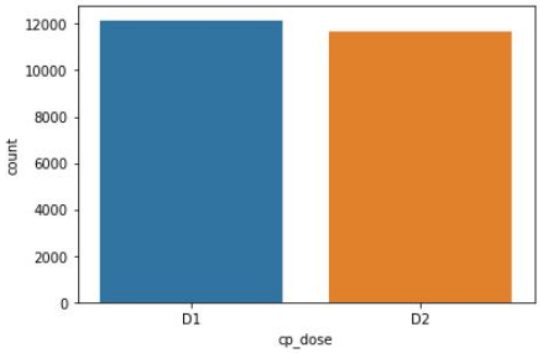
Exploring data

Features:

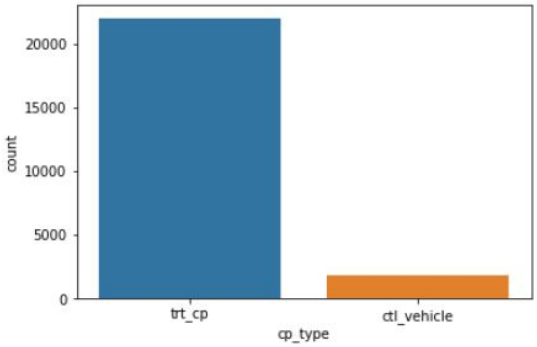
- cp_time - treatment duration
    - Categorical
    - 3 possible values (24, 48, 72 hours).
    - Mean - 48.02
    - Std - 19.40



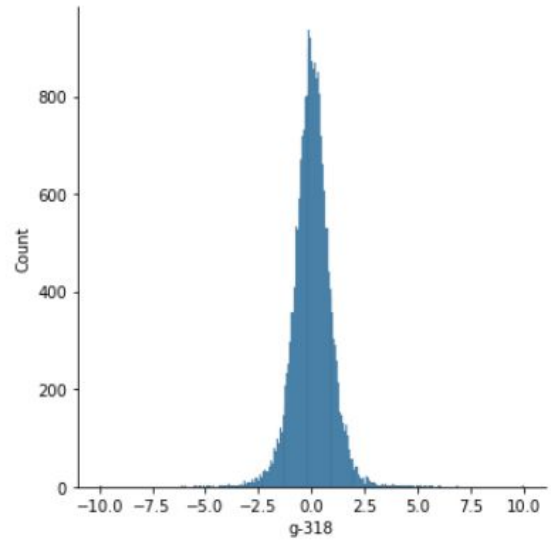
- cp_dose - treatment dose
    - Categorical
    - 2 possible values
        - D1 - high
        - D2 - low

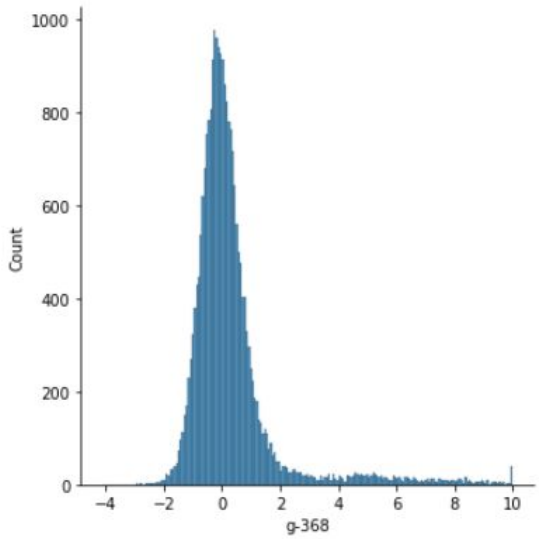


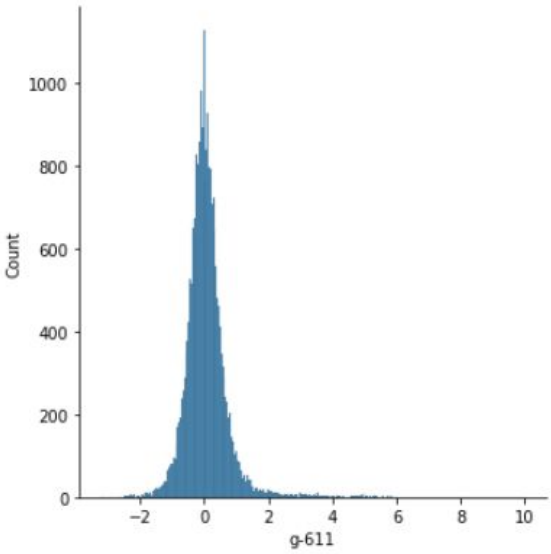
- cp_type - the type of perturbation used to treat the samples
    - Categorical
    - 2 possible values
        - trt_cp - compound perturbation
        - ctl_vehicle - control perturbation (have no MoAs)

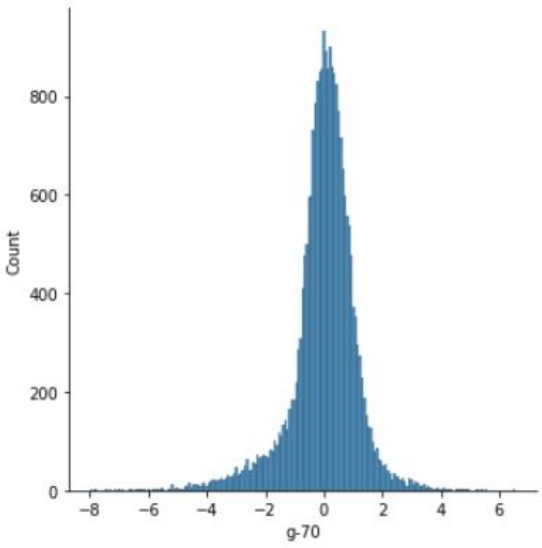


- g-[0:771] - gene expression data
    - Numerical
    - Value range: [-10, 10]
    - Example summary for g-0
        - mean - 0.2484
        - std - 1.3934
        - median - -0.008850
    - Distributions of randomly selected features

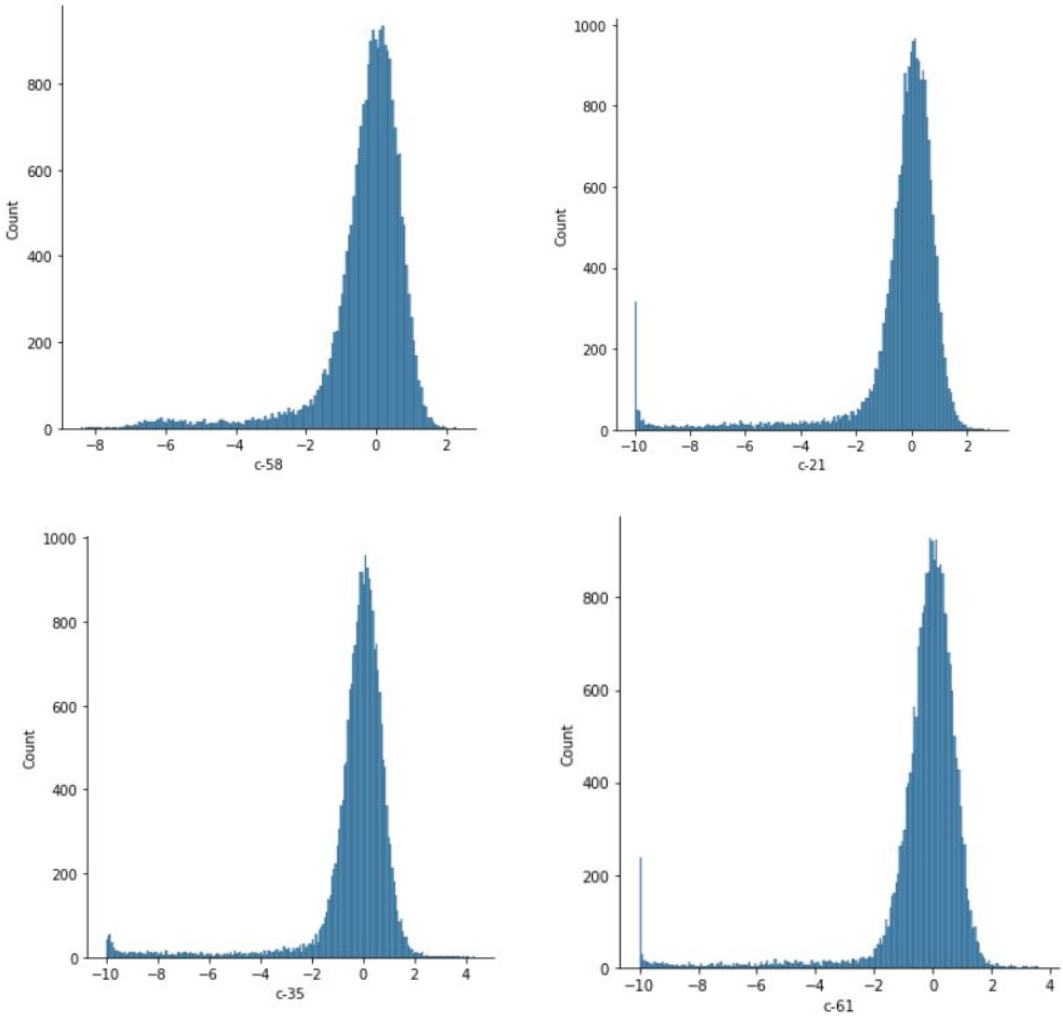








- c-[0:99] - cell viability data
    - Numerical
    - Value range: [-10:6.412]
    - Example summary for c-0:
        - mean - -0.355
        - std - 1.752565
        - median - -0.009
    - Distributions of randomly selected features



Targets:
Include 206 different binary MoA responses that will be predicted. Possible values are 0 and 1. A single case can have multiple MoA responses so this is a multi-label classification, not multiclass classification.

Verifying data quality

The data appears to be of very high quality. The data presented to us by the Laboratory for Innovation Science at Harvard has everything we need easily presented to us. The data is well organized, doesn't have any missing values and doesn't appear to have any incorrect values. All of the features appear to be useful and there seem to be enough samples for good generalization of the model. Gene expression and cell viability data seems to be clipped at -10 and 10 which explains the spikes seen in the distribution.

Task 4. Planning your project:

We are planning to use a jupyter notebook with python.

- **Data Analysis** (Cheung: 2 hours, Seyedeh: 2 hours, Janar: 4 hours)
  Trying to find patterns in the data and looking at the relationship between targets and their target genes. Check the correlation between the different gene expression features and the different cell viability features. Look at different target types and correlation between them (possibly check scored_target correlation with nonscored_target). Gain a better understanding of the relationship between train features and test features.
- **Data Preparation** (Cheung: 4 hours, Seyedeh: 5 hours, Janar: 4 hours)
  The data we are working with is of very high quality so little data preparation is necessary. The primary data preparation will involve applying techniques like one-hot-encoding, normalization and feature engineering in order to increase the performance of a model. This will have to be done/evaluated on a per model basis as different machine learning models like their input data in different formats.
- **Modelling** (Cheung: 17 hours, Seyedeh: 16 hours, Janar: 17 hours)
  Constructing different models in order to find the best performing one. Finding the best hyperparameters for the models. The exact models used are to be determined, but most likely will include at least one ensemble learning model and experimenting with a neural network.
- **Evaluation** (Cheung: 2 hours, Seyedeh: 2 hours, Janar: 2 hours)
  Compare the performance of different models and check whether the best one has good enough performance. If it doesn't try to improve the models.
- **Presentation** (Cheung: 5 hours, Seyedeh: 5 hours, Janar: 3 hours)
  Create a poster and a video presenting the project.