CMSC 33750 HW3

Lynn Zheng

October 29, 2020

This writeup contains the high-level description of my approach and results. Specific design and implementation choices are described using Markdown in the code notebook files.

Note: There are 200+ drugs and 1000+ cell lines, so it's not space efficient to reproduce the lines for every table in this writeup. The tables presented here select at random 5 drugs or cell lines. The complete tables are stored in CSV format and can be directly read into Pandas DataFrames.

Code Notebooks and Table CSV Files

- data_cleaning.ipynb - for cleaning and quantizing the raw CSV files
- hw3_part1.ipynb - regression and classification on the raw data
- hw3_part1_improvements.ipynb - improve upon part 1 results by first performing PCA on the raw data
- hw3_part2_racs.ipynb - use RACS Cancer Type as a categorical variable
- hw3_part3_spearman.ipynb - comparing rank ordering between GDSC experimental results and model predictions using the three models above
- results_analysis.ipynb - more model performance comparison and analysis

- drug_df_(part1,pca,mse).csv - fitted IC50 values and predicted sensitivity responses for the three models
- result_df_(part1,pca,mse).csv - MSE, accuracy, AUC per drug for the three models
- (part1,pca,mse)_mse.csv - top 10 and bottom 10 drugs ranked in terms of MSE
- (part1,pca,mse)_top10_per_cell_line.csv - top 10 sensitive drugs per cell line
- spearman_df.csv - Spearman correlations for drug ranking per cell line

Method

For Part 1, I first experimented with using the raw data. Then I reduced the dimension of the raw data using PCA, keeping a set of components that explain 80% of the variance. This also sped up training significantly. In Part 2, I one-hot-encoded the `Cancer Type` categorical variable from RACS data and concatenated that with the raw data reduced using PCA. Using this approach, for each drug, the train matrix usually have dimensions `(num_cell_lines, num_features) = (400, 150 to 300)`.

I used Linear Regression and K-Nearest Neighbors Classification algorithm.

To perform classification, for each drug, I quantized the log IC50 values into 3 bins with `pandas.cut`.

1 Regression and Classification

1.1 Raw Data

Table 1: Top 10 and bottom 10 in terms of MSE using raw data

drug_id_top10	mse_top10	drug_id_bottom10	mse_bottom10
1262	0.269010	1248	7.180334
266	0.291851	135	6.315710
150	0.301788	190	5.065923
1264	0.321767	268	4.938611
91	0.381167	346	4.889768
341	0.402388	51	4.356716
1502	0.425869	302	4.309938
205	0.437814	344	4.270240
193	0.491845	3	4.193851
202	0.493775	299	4.142698

1.2 Improvements: PCA, 80% Variance Explained

Table 2: Top 10 and bottom 10 in terms of MSE using raw data reduced with PCA

drug_id_top10	mse_top10	drug_id_bottom10	mse_bottom10
1262	0.222263	135	6.373095
266	0.236688	1248	6.334078
150	0.256308	268	5.119414
1264	0.277344	346	4.779838
91	0.335878	190	4.731712
341	0.346542	3	4.485934
205	0.374317	51	4.451908
1502	0.384483	299	4.172587
312	0.439273	344	3.965724
1018	0.440379	302	3.921599

2 RACS Cancer Type

Table 3: Top 10 and bottom 10 in terms of MSE using raw data after PCA and RACS cancer type data

drug_id_top10	mse_top10	drug_id_bottom10	mse_bottom10
1262	0.223508	135	6.376384
266	0.238326	1248	6.363803
150	0.257682	268	5.102002
1264	0.285128	3	4.879453
91	0.332634	346	4.736399
341	0.349962	190	4.735117
205	0.378678	51	4.572035
1502	0.394210	299	4.157039
312	0.438978	344	3.978235
1018	0.450332	56	3.930705

3 Rank Drugs Per Cell Line

If the cell line has fewer than 10 sensitive drugs, the few sensitive drugs are reported in order of best to worst.

3.1 Tables for Top 10 Drugs Per Cell Line

Table 4: Top 10 drugs for 5 randomly selected cell lines, using raw data

cell_line_id	top_10_drug_ids
687802	268,194,200,273,272,1243,24
1290810	1031,200,1166,190,1243,1239,245,115
924102	201,180,1007,1004,1003,135,140,1031,268,19
1299064	201,180,1007,1004,140,194,135,1248,200,34
684062	1007,283,1003,268,1031,1004,346,200,194,18

Table 5: Top 10 drugs for 5 randomly selected cell lines, using raw data after PCA

cell_line_id	top_10_drug_ids
906800	283,1248,180,201,11,1494,226,1003,1004,14
724869	1372,268,200,299,1091,1243,119
1240159	201,1261,1007,180,1494,194,200,182,157,137
949177	201,104,1007,1248,1494,1004,11,140,180,26
687561	1007,268,1248,1494,1031,1057,194,1026,153,152

Table 6: Top 10 drugs for 5 randomly selected cell lines, using raw data after PCA and RACS cancer type dataA

cell_line_id	top_10_drug_ids
905990	104,268,201,140,180,1031,11,194,83,8
908464	268,200,273,272,274,276,45,291,24
753572	201,140,268,1494,208,180,135,194,133,152
908443	180,140,283,208,346,268,135,194,200,30
909748	1372,282,22

3.2 Spearman Correlation

I used Spearman's correlation to compare the drug ranking predicted by the model with the GDSC experimental results. We generally get a higher Spearman correlation as we go from using the raw data, PCA, and integrating RACS data.

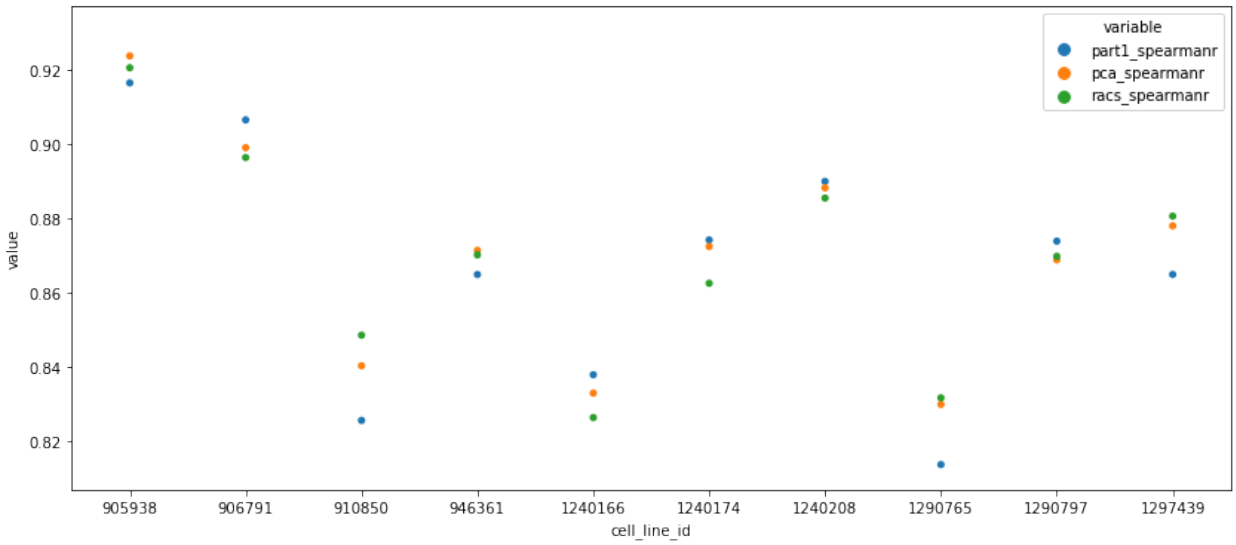

Figure 1: Spearman correlations for the three models for randomly selected cell lines

Table 7: Spearman correlations for 5 randoml selected cell lines for all three models

cell_line_id	part1_spearmanr	pca_spearmanr	racs_spearmanr
724863	0.733454	0.753834	0.739104
1331040	0.819355	0.851435	0.839068
1297449	0.854064	0.854597	0.859085
907073	0.891929	0.897533	0.897490
1479993	0.859944	0.874960	0.865560