

A decorative graphic on the left side of the slide consists of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

Final Presentation: MoA Prediction

Bo Peng, Kyeong Joo Jung, Jinhee Lee,
Zhenhao Lu

Introduction

Drug discovery

Scientists seek to identify a protein target associated with a disease and develop a molecule that can modulate that protein target. As a shorthand to describe the biological activity of a given molecule, scientists assign a label referred to as mechanism-of-action or MoA for short.

Our data from the Kaggle: MoA prediction competition

<https://www.kaggle.com/c/lish-moa>





Project Aims

Aim: analyze the cellular responses with algorithms that search for similarity to known patterns in genomic database with known MoAs.

Predicting multiple targets of the MoA responses of different samples based on the gene expression, cell viability, treatment duration and treatment dose data.

- Example of MoA targets

NFKB_inhibitor, Serotonin_receptor_antagonist, DNA_inhibitor,
Dopamine_receptor_antagonist, Cyclooxygenase_inhibitor, Proteasome_inhibitor,
Calcium_channel_blocker, ...

Data Statistics

Train_data : 23,814 samples

Test_data: 3,982 samples

Sig_ids: Uniquely identify each sample

Features: [23,814 X 875]-training [3,982 X 875]-testing

	A	B	C	D	E	F	G
1	sig_id	cp_type	cp_time	cp_dose	g-0	g-1	g-2
2	id_000644t	trt_cp	24 D1		1.062	0.5577	-0.2479
3	id_000779t	trt_cp	72 D1		0.0743	0.4087	0.2991
4	id_000a62t	trt_cp	48 D1		0.628	0.5817	1.554
5	id_0015fd3	trt_cp	48 D1		-0.5138	-0.2491	-0.2656
6	id_001626t	trt_cp	72 D2		-0.3254	-0.4009	0.97
7	id_001762t	trt_cp	24 D1		-0.6111	0.2941	-0.9901

cp_type: samples treated with a compound(trt_cp) or a control perturbation(ctl_vehicle)

cp_time: treatment duration (24, 48 and 72 hours)

cp_dose: treatment dose (high or low)

g- : 772 gene expressions (g-0 to g-771)

c- : 100 cell viability (c-0 to c-99)

Targets: [23,814 X 206]

Probability of having MoA labels (totally 206)

	A	B	C	D	E	F
1	sig_id	5-alpha_re	11-beta-hs	acat_inhibi	acetylcholi	acetylcholi
2	id_000644t	0	0	0	0	0
3	id_000779t	0	0	0	0	0
4	id_000a62t	0	0	0	0	0
5	id_0015fd3	0	0	0	0	0
6	id_001626t	0	0	0	0	0
7	id_001762t	0	0	0	0	0

Arithmetic
data



Evaluation

Every sig_id(sample) will be predicting the probability that the sample had a positive response for each MoA target.

The Log Loss Function (From the competition)

N(sig_id) rows and M(MoA targets) columns

$\hat{y}_{i,m}$ is the predicted probability of a positive MoA response for a sig_id

$y_{i,m}$ is the ground truth, 1 for a positive response, 0 otherwise

Score is the average validation loss of all experiments (7 folds)

$$LogLoss = -\frac{1}{M} \sum_{m=1}^M \frac{1}{N} \sum_{i=1}^N (y_{i,m} \log(\hat{y}_{i,m}) + (1 - y_{i,m}) \log(1 - \hat{y}_{i,m}))$$

K-Fold Cross-Validation

Split the dataset into k groups: folds

For each fold:

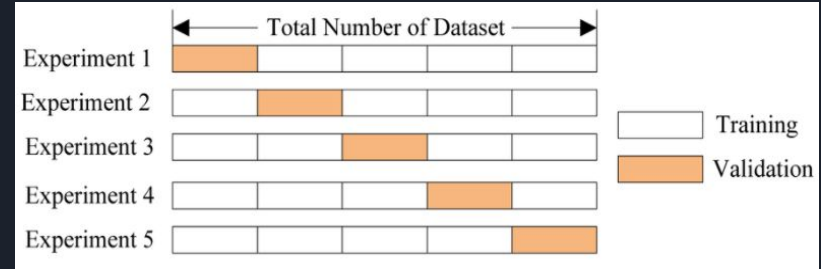
Take this fold as validation set and the remaining k-1 folds as training set

Fit a model on the training set and evaluate it on the validation set

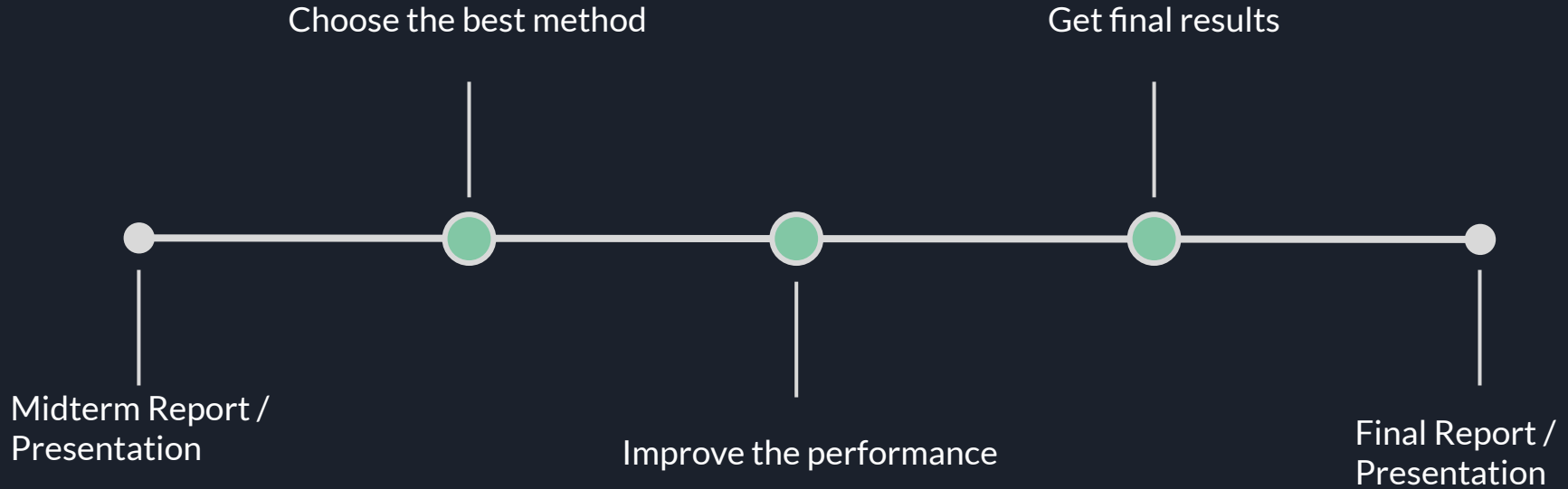
We set $k=7$ for this dataset (7-fold Cross-validation)

Reason: There are 23,814 training samples: It is divided by 7.

To reduce time for running



Plans after the Midterm Presentation





Methods

Find some baselines for the MoA prediction.

Our method:

Multi-layer neural network with permutation importance

Compared method:

1. Kernel logistic regression
2. TabNet

Compared method 1: Kernel Logistic Regression

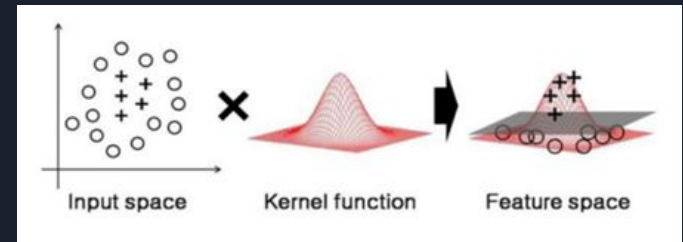
This method uses the Kernel Ridge Regression with RBF kernel and Platt scaling for the calibration (Logistic Regression)

Hyperparameter Setting (C: 1 and gamma: 100)

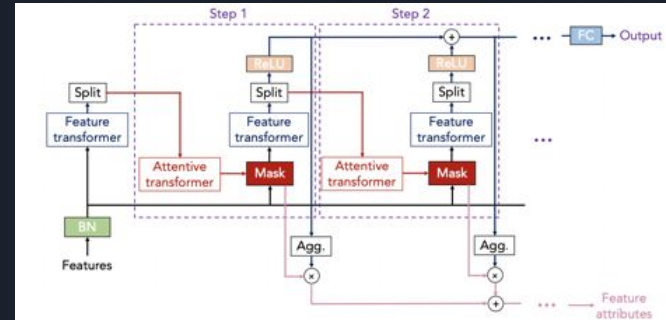
Use Nystroem Approximation (To reduce the computational costs)

Score: 0.0189

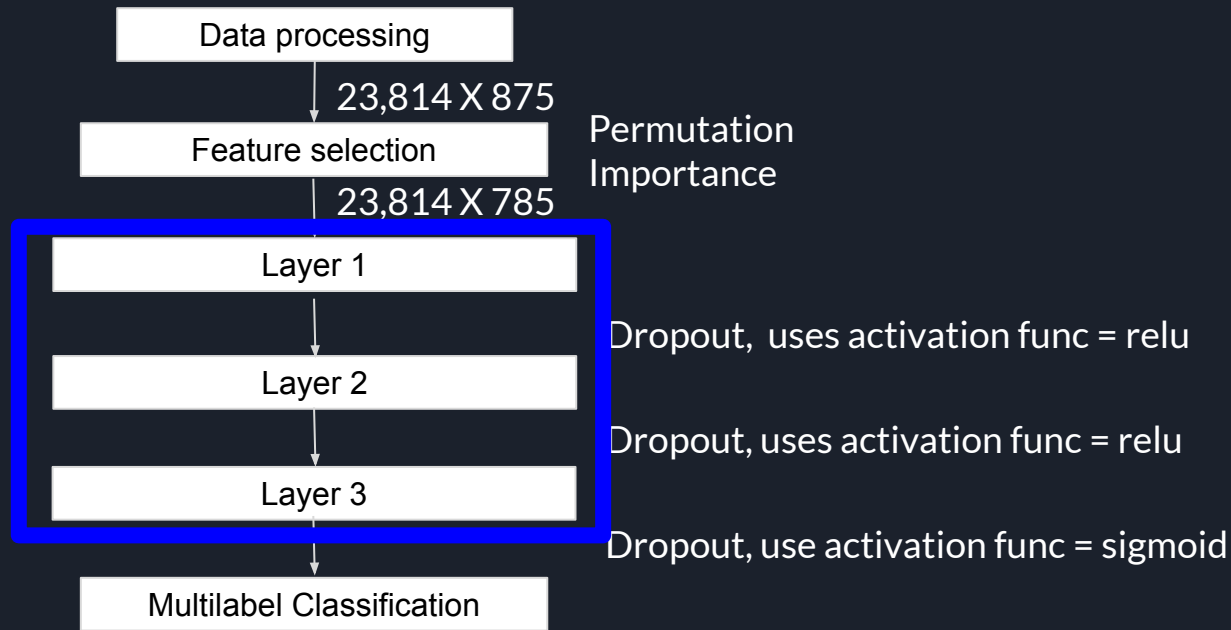
$$K_{RBF}(x, x') = \exp(-\gamma ||x - x'||^2) \quad \gamma = \frac{1}{2\sigma^2}$$



Tree-like deep method



Multilayer Neural Network with feature selection



Multi-Layer Neural Network

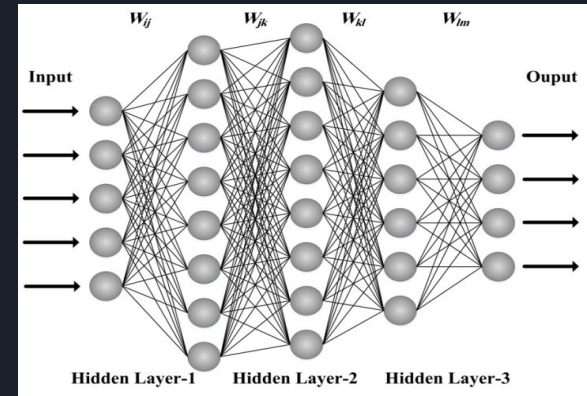
It's implemented using tensorflow

First hidden layer with 2,048 neurons

Second hidden layers with 1,024 neurons

Use Adam as the optimizer of the neural network

Set up a learning rate scheduler with ReduceLROnPlateau.



Multilayer Neural Network with feature selection

Permutation importance: Finds important features
In this method, important features are collected & used

sig_id	cp_type	cp_time	cp_dose	g-0	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9
id_000644	trt_cp	24	D1	1.062	0.5577	-0.2479	-0.6208	-0.1944	-1.012	-1.022	-0.0326	0.5548	-0.0921
id_000779	trt_cp	72	D1	0.0743	0.4087	0.2991	0.0604	1.019	0.5207	0.2341	0.3372	-0.4047	0.8507
id_000a62	trt_cp	48	D1	0.628	0.5817	1.554	-0.0764	-0.0323	1.239	0.1715	0.2155	0.0065	1.23
id_0015fd	trt_cp	48	D1	-0.5138	-0.2491	-0.2656	0.5288	4.062	-0.8095	-1.959	0.1792	-0.1321	-1.06
id_001626	trt_cp	72	D2	-0.3254	-0.4009	0.97	0.6919	1.418	-0.8244	-0.28	-0.1498	-0.8789	0.863

Features

:875



Samples

:23814



Multilayer Neural Network with feature selection

Permutation importance: Finds important features
In this method, important features are gathered & used

Shuffles data of each column and compare the result

Decrease: important
Increase: not important

sig_id	cp_type	cp_time	cp_dose	g-0	g-1	g-2	g-3	g-4	g-5	g-6	g-7	g-8	g-9
id_000644	trt_cp	24	D1	1.062	0.5577	-0.2479	-0.6208	-0.1944	-1.012	-1.022	-0.0326	0.5548	-0.0921
id_000779	trt_cp	72	D1	0.0743	0.4087	0.2991	0.0604	1.019	0.5207	0.2341	0.3372	-0.4047	0.8507
id_000a62	trt_cp	48	D1	0.628	0.5817	1.554	-0.0764	-0.0323	1.239	0.1715	0.2155	0.0065	1.23
id_0015fd	trt_cp	48	D1	-0.5138	-0.2491	-0.2656	0.5288	4.062	-0.8095	-1.959	0.1792	-0.1321	-1.06
id_001626	trt_cp	72	D2	-0.3254	-0.4009	0.97	0.6919	1.418	-0.8244	-0.28	-0.1498	-0.8789	0.863

Features

:885



Samples
: 23814



Scores: 0.0161



Scores of each method

Method	Log Loss
Kernel Logistic Regression	0.0189
TabNet	0.1553
Multi-Layer Neural Network with Permutation Importance	0.0161

Our Model





Results

Ablation Study:

- Multi-Layer Neural Network with Permutation Importance: 0.0161
- Multi-Layer Neural Network: 0.0163

Parameter Study:

- 1-Layer Neural Network: 0.0170
- 2-Layer Neural Network: 0.0164
- 4-Layer Neural Network: 0.0169
- 5-Layer Neural Network: 0.0169
- 6-Layer Neural Network: 0.0177



Improvement

Use tanh activation: 0.0165

More hidden units in each layer: 0.0163

Less hidden units in each layer: 0.0168

Using PCA for dimension reduction: 0.0166

Using PCA and variance threshold: 0.0166



Additional...

Label Smoothing

- regularization technique for the overconfident problem
 - decrease the gap among logits (i.e., output before the softmax)
 - make the learning process more smooth
- formula
 - $y_{ls} = (1-a) * y + a/k$



Conclusion&Issue

Conclusions:

- Simple model could work better than fancy models

- Hyperparameters are important

Issues:

- Hard to improve the performance

- Lack of computational resource



Q&A

Thank you!