Math 6380: A Mathematical Introduction to Data Analysis

Project1 Report

Hao Xin   Jinxing Yu   Xun Jian

1 Introduction

We choosed the drug sensitivity data and participated in the kaggle inclass contest on Cleave Drug Sensitivity Prediction. We are the team "3654". Our best submission gets **0.11088** Mean-Square-Error. The dataset consists of 149 cell lines and their MRNA, CNV, mutation values of genes. Provided 129 cell lines' IC50 values at 24, 28, 72hrs, the task is to predict the IC50 values at 72hrs for the remaining 20 cell lines.

The key challenge we found is that the sample size of the datasets is very small (149) while the features dimension is very high (more than 35, 000). This makes the learning algorithms prone to overfitting. Another problem caused by the small size of dataset is the bias between trainning data and test data. We did K-fold cross validation on the training dataset for model selection, we found that some models which performed best on cross-validation set, however got worse submission scores.

2 Methodology

In data preprocessing, we merged MRNA, CNV, mutation values of genes to form a feature vector for each cell line and standardized features by removing the mean and scaling to unit variance. We did not use the information of gene set and the tissue type of cells.

To cope with the high-dimension of features and prevent the learning algorithms from overfitting, we performed dimension reduction or feature selection on features. We compared severial regression methods including LASSO, Ridge Regression, Support Vector Regression, Decision Tree, and Gradient Boosting. Two python packages:scikit-learn and pandas were used in the implementation.

We also manually implemented a recursive feature selection method from scratch based on Lasso regression errors on cross validation data. Different from Lasso recursive feature elimination in the reference poster which repeatly eliminates some features, our feature selection method repeatly add a feature until the Lasso regression error on validation dataset does not decrease. The advantage of our method is that it is more efficient than recursive feature elimination.

3 Results and Discussion

We first compare different regression method and check the effectiveness of dimension reductioon or feature selection methods based on the Mean Square Error through 5-fold cross validation. The training errors are also included to check the overfitting. We then picked several models and trained them on the whole training dataset and submitted them on Kaggle.

From the results in Table1, we can see that the features are redundant and corelated. In PCA and KBest, we reduce the feature dimension to 200 without degrading the predictive performance. The 200 principle components can explain almost 100% of the variance. However, to our experiences, the predictive performance on cross validation is not consistent with the result on Kaggle. We think it is caused by the bias of the test data.

An interesting finding we found is that the variance of IC50 values at 72hrs is about 1. If only mean square error is used as evaluation matric, we can get a result with 1 mean square error by predicting the mean.

train, val	All features	PCA	Kbest	Lasso-error based
LASSO	0.9428, 1.0327	0.4411, 0.9553	0.9504, 1.039	0.9737, 1.0420
Ridge Regression	1e-8, 1.0516	1e-8, 1.0516	0.0036, 1.4114	0.0882, 0.1750
SVR	0.5788, 0.9463	0.4199, 1.0269	0.3111, 0.9262	0.2609, 0.6376
Decision tree	1e-10, 1.675	1e-10, 1.2311	1e-10, 1.4497	4.8e-10, 0.9405
Gradient Boosting	1e-6, 1.077	1e-3, 1.0538	1e-3, 1.0664	1.1e-3, 0.5789

Table 1: Cross validation results measured by mean square error for each method

4 Remark on Contributions

The project is finished under the discussion and close collaboration of our group members. Hao Xin wrote the code skeleton, tried KBest feature selection method, and compared many regression methods by cross validation. Jinxing Yu performed dimension reduction, tried the baseline using gradient boosting without dimension reduction or feature selection, which unexpectedly got best submission score, and wrote the majority of the report draft. Xun Jian proposed the recursive feature selection method and implemented it, which got the best cross validation result.