Identification of the Predictive Biomarkers of Drug Response using Elastic Net Regression

Brief Introduction

Regression analysis is widely used for drug response prediction tasks. In order to find relation among genetic expression features to drug-response variable, Regression analysis is performed. The assignment problem statement requires the identification of the predictive biomarkers of drug response which is a "large p, small n" case (high-dimensional data with few observations), with 285 cell lines and 13321 genetic features.

A core issue in machine learning is the selection of the optimal method for a given problem. Cross-Validation is a standard resampling procedure used for model selection in order to estimate the error of a predictive model on a test set. Ideally we would like to have a new data observations but it is not possible so Cross-Validation is a clever device to use the same training data to tell how well the prediction method works. The motivation for cross-validation is to create a number of partitions of sample observations, called the validation sets, from the training data set. The central idea is to fit a model on to the training data, measure the performance on each validation set, so as to assess the prediction performance of the model for unseen data points. The numbers of slices of data is usually determined based on the number of observations in the sample data set, Biasness is reduced with increasing in number of partitions in cross-validation.

Overfitting in regression models leads to erratic predictions and unreliable results for unseen data. Hence, shrinkage techniques are one of many ways which assign weights (coefficients) to the features. Ridge regression is one technique to prevent overfitting. It does this by penalizing the L2 norm of the coefficient vector which results in “shrinking” the beta coefficients. The potency of the penalty is controlled by another hyper- parameter called lambda λ. Lasso regression is another method that uses regularization approach. In place of L2 norm, L1 norm of the coefficient vector is penalized. Because it uses the L1 norm, some of the coefficients will shrink to zero while lambda increases. (Jerome H. Friedman, 2010-02-02)

Elastic net regression is a hybrid method that uses both penalization of the L2 and L1 norms. It penalizes the complexity of the model along with assigning zero values to the non-significant features. The value of the α hyper-parameter is between 0 and 1 that affects L2 and L1 penalization. Another hyper-parameter “lambda” is selected through cross-validation i-e either by minimizing the cross-validated mean squared prediction error or maximizing the correlation values. (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf, 2016).

Among 13321 genes, identifying the genes which are predictive to drug-response is the basic theme of this assignment. I have chosen Elastic Net as some genes are more responsible for the drug-response then others so in order to nullify the weights of those genes which do not have direct impact on the drug response, elastic net regularization technique is used and the experimental set is explained in the next section.

Data Pre-processing

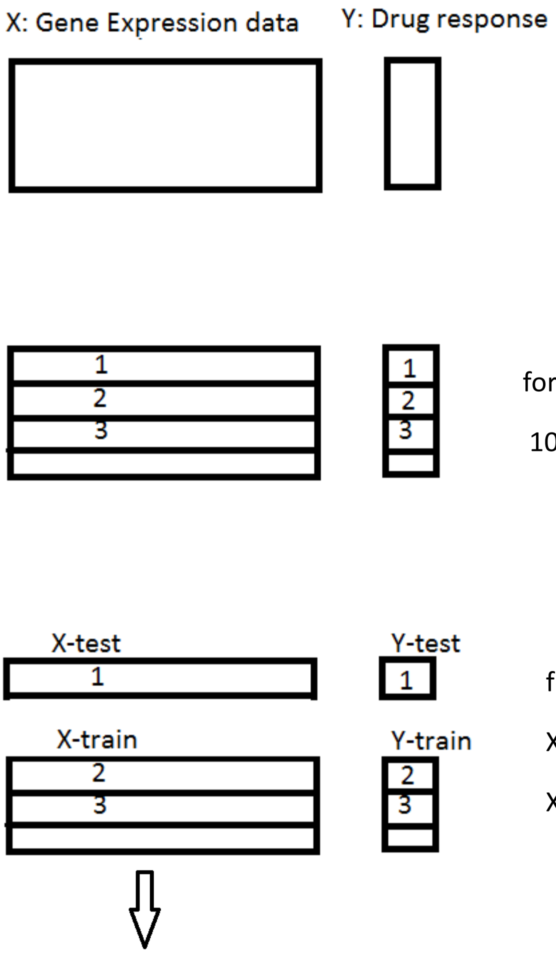
The expression data is count data, Log transformation was performed so as to enhance the signal in the data. Assuming all genes are equally relevant Z-transformation is used to normalize the data. For each gene mean=0 and sd=1 is done, in addition to the drug responses.

Experiment Set-Up

Nested 10 fold cross-validation is used for the model selection in this problem. Data is sliced in different folds to generate training sets and test set. There are two hyper-parameters which need to be optimized i-e alpha and lambda. Model is evaluated with 10 fold C.V on training data, using 100 values (0.01 to 0.99) of alpha and optimal lambda is chosen with internal 10 fold C.V scheme. The trained model is used to make predictions on the held out unseen test dataset. Predicted drug response values are compared with the "Observed" drug response values and the mean square error is calculated. Spearman correlation is also calculated between the prediction values and the real values of drug response.

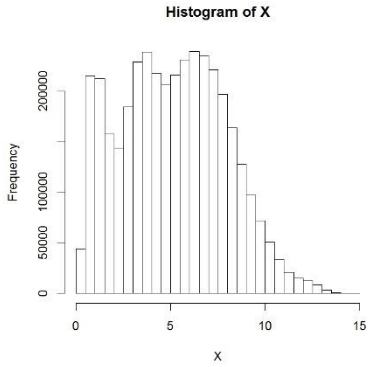
As illustrated in the fig 2 below, the alpha value with the highest spearman value is selected to be used for the feature selection on whole data. After plugging the selected alpha value in the model, 10 fold cross-validation is done on the whole data to choose the lambda value. The glmnet function uses the best alpha value along with the chosen lambda value to assign weights/coefficients to the predictors (genes). The experiment is repeated four times so that common features can be selected. The common features along with weights from one experiment result are reported in table 1.

MODEL SELECTION ALGORITHM



Complete data :

log transformation of data to enhance the signal



for α (0.01………0.99)

10-fold CV of X and Y

for fold =1

X-train , Y-train

X-test , Y-test

Normalize

mean = 0 , sd =1 (Column wise)

STEPS INVOLVED

- Internal Cross-validation on training data for alpha value selection
- Lambda value of with 1 S:E is used to train the model

- Trained model is used to predict on the Test data
- MSE, Spearman correlation, pearson correlation values are computed

FEATURE IDENTIFICATION ALGORITHM

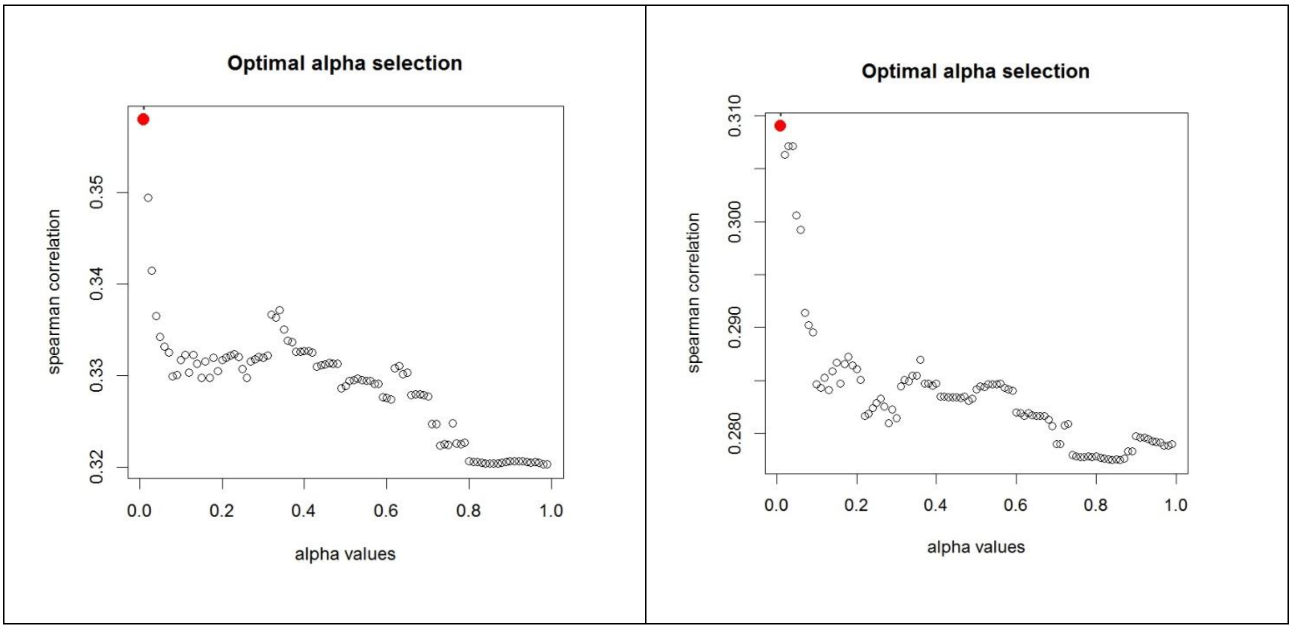
STEPS INVOLVED

- Optimal value of alpha is chosen (alpha value that has the highest spearman correlation)
- full data normalization X , Y
- 10 fold cross-validation is used and lambda.1S.E is plugged to glmnet function to get coefficients (weights of features)

fig 1 : The method adopted in order to select the model is shown in this fig. The algorithm used for the selection of hyper-parameters in model selection is also shown. Steps involved in the Algorithm for feature selection are also listed in the fig.

Prediction Correlation performance: 10- fold Cross-validation

Following figures are generated by repeating experiments four times i-e by selecting different seeds for the fold generation. The optimal value of alpha parameter selected is represented by red dot. It indicates that the value with the highest spearman correlation is selected from the grid of 0.01 to 0.99 value. For selecting genes, the model is trained on the full data using this indicated red dot value of alpha, while learning the optimal lambda value through nested cross-validation.



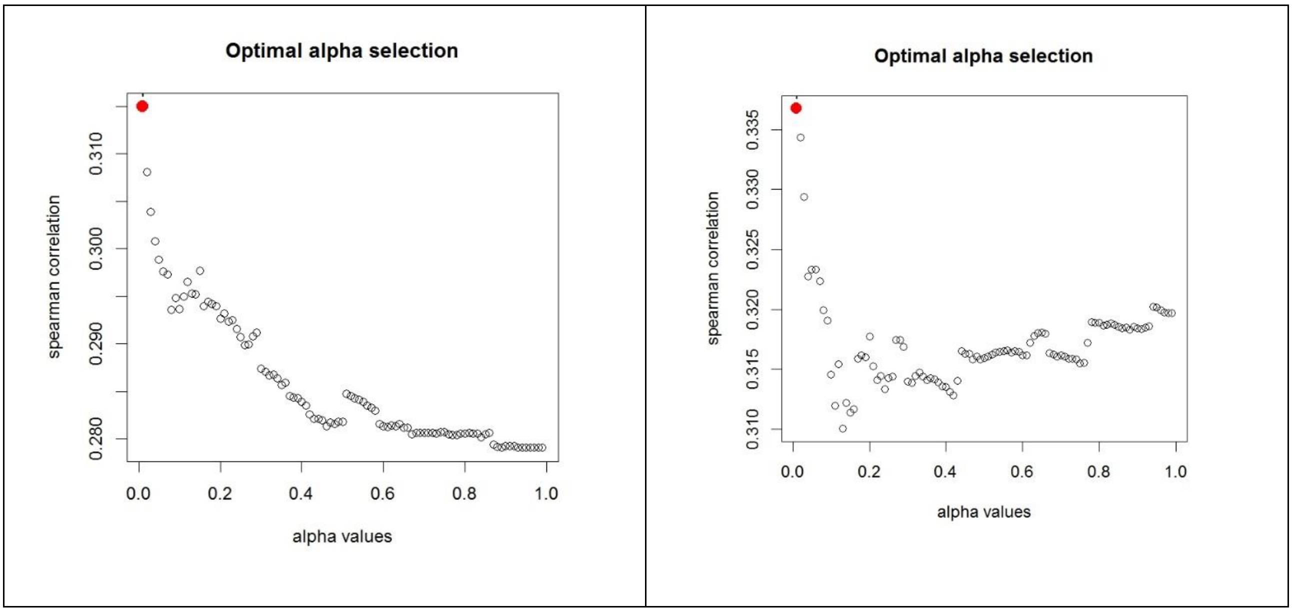


fig 2: Spearman correlation values and the alpha values are represented in the scatter plots shown in the figure. The red dot indicates the optimal alpha value which is selected for feature selection.

List Of Genes Predictive Of Drug Response

The genelist which was consistent in four instances after repeating experiment. However, the weights are represented using one of the seed results.

Table: Genelist which consistently appeared in four repetitions of the experiment

SKP1A	RAI14	PPIA
-3.719565e-03	-8.564554e-04	-1.795188e-03
PRG1	RPS4Y1	IGFBP7
5.415390e-04	2.673845e-03	-2.636723e-04
IGFBP6	MAGEC1	MAGEA6
-2.108700e-04	3.000264e-04	4.272892e-04
ABCC3	CA2	CLU
-5.353537e-04	-2.404547e-03	-2.389802e-04
CLDN1	TPD52L1	CORO1A
-1.630618e-03	-8.745201e-04	1.021151e-03
IGHM	CD24 /// LOC647456	CDKN2A
1.349088e-03	-1.408576e-03	1.410948e-03
UGT1A10 /// UGT1A8 /	KRT5	OLFML2A
-3.841307e-03	-5.189662e-04	-7.015786e-04
GABRE	KIF21B	MATR3
-1.280122e-03	3.116570e-04	-4.076261e-04
SLC25A3	PTRF	DUSP4
-5.770608e-04	-1.372640e-03	8.020176e-04
FGFR3	CXCR4	EGFR
-2.444947e-03	5.086651e-03	-1.193024e-03
YAP1	MYB	AKR1B10
-1.133681e-03	8.035161e-04	-3.889741e-05
TNFSF10	C14ORF139	UGT1A6
-3.771515e-04	2.205851e-04	-5.268476e-03
IFITM3	KRT6A /// KRT6C ///	PRSS23
-6.659226e-04	-7.520602e-04	-2.635727e-03
FER1L3	DKK1	KRT19
-3.262694e-03	-2.017807e-03	-2.406013e-03
KRT17	EFEMP1	PDLIM1
-1.137932e-03	-7.406573e-03	-6.896117e-04
TPBG	AKR1C3	AKR1C2
-4.615864e-04	-7.576607e-03	-2.698007e-03

AKR1C1	DOCK2	ANXA8 /// LOC653107
-1.666337e-03	4.055046e-04	-3.216914e-03
KHDRBS3	SERPINB5	TACSTD2
6.264386e-04	-1.827649e-04	-1.023422e-03
LCN2	ALDH1A1	COL4A5
-2.467341e-03	-4.712877e-03	-1.143552e-03
LOC642299 /// LOC651	LAPTM5	TRIM29
-2.402502e-04	1.460630e-03	-6.089209e-05
AXL	PADI3	LEPREL1
-1.237670e-03	-7.199191e-04	-1.771827e-03
ANXA1	ANXA3	
-1.379233e-03	-2.053380e-03	

Obervations

I have observed that the genes such as DUSP4, CXCR4 has a positive weight in all the experiment runs where as genes like EFEMP1, EGFR have shown to have consistently negative weights.

References

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf. (2016, 12 14). Retrieved from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf

Jerome H. Friedman, T. H. (2010-02-02). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Joural of statistical software*.