Empirical Evaluation Of Elastic Net In Cancer Gene Analysis

Bingyi Jiao 20900771 BJIAO@UWATERLOO.CA
Yuqing Zhang 20911078 Y3593ZHA@UWATERLOO.CA

Abstract

We did a **data analytic and applied statistics** project in the context of high-dimensional cancer gene data modeling. Variable selection is of paramount importance in gene data analysis because the data often has extremely high dimensions relatively to the sample size. Among all kinds of variable selection techniques, elastic net is excellent for its parameter estimation ability, the balance of fitting and penalizing, and a unique advantage of grouping effect.

In our project, we trained several gene TP53 inactivation classifiers, compared the performance of lasso, ridge and elastic net in sparse logistic regression and support vector machine. We verified that logistic regression is a better model in this project and elastic net performs better than lasso and ridge regularization, with a higher test AUC (93.6%) and reasonable proportion of selected variable. In addition, we did further analysis to make sure this elastic net model has biological significance and good interpretability.

Keywords: Elastic Net; Gene TP53; Sparse Logistic Regression.

1. Introduction

Recent decades have witnessed a major development in gene sequencing technology, which becomes more scalable, stable and comprehensive. The cost of sequencing has dropped rapidly and analyzing gene sequences is crucial for identifying genetic diseases, preventing disease outbreaks and defining cancer progression drivers.

Nowadays, a large volume of cancer gene data has been provided in open access to researchers. For instance, the famous Tumor Genome Mapping Project (TCGA) was launched by the National Cancer Institute in last decade and provided huge amount of valuable data. How to extract useful information from this large amount of data is of great concern.

Gene data analysis can exert profound impacts. If some particular genes are inactivated, it will indicate a high risk of triggering cancer or other bad diseases. We can build a model using other genetic data to predict if one particular gene is inactivated or not. A good example is a famous gene, the tumor protein p53 (TP53), which acts as a tumor suppressor. Inactivated TP53 is hazardous for DNA repairing and cell division, thus leading to uncontrolled cell division and triggering tumors (Chen et al., 2010).

In statistics, gene data is also instrumental in model building for its inherit high-dimensional characteristics. High-dimensional means the number of variables (the number of genes, p) is much larger than the sample size (n). Furthermore, an important fact is that each cancer is affected by at most one percent of all human genes (Futreal et al., 2004). Putting other irrelevant gene data into the model will contribute nothing but increase unnecessary complexity of the model. In practice, to improve the model interpretability and accuracy, variable selection is the key determinant.

Among all the variable selection methods, the embedded method is a popular one in recent years. It can select variables and construct a model at the same time. Embedded methods such as sparse methods provide less computation than wrapping methods (Guyon and Elisseeff, 2003). For cancer gene selection, sparse logistic regression and sparse support vector machine (SVM) with regularization can simultaneously select variables and estimate parameters by introducing a penalty term to the loss function, which is helpful to early cancer detection in clinical trials and various cancer mechanism research.

Apart from variable selection, another aspect which needs to be taken into consideration is the 'grouping effect' of genes. Research has found that in human body, it is a common phenomena that multiple genes share the same biological 'pathway'. This will bring about the multicollinearity problem. In gene selection, an ideal model should be able to select all the genes in the same pathway.

In our project, we hope to build a competitive gene inactivation status classifier. The input is numerical data of other genes expression and the output is the binary data of whether a particular gene is inactivated. We will explore two kinds of classifier - sparse logistic regression and sparse SVM, with three types of regularization (lasso, ridge and elastic net) respectively. The result is evaluated by computing the AUC, accuracy and comparing the number of genes selected.

In our expectation, Elastic Net should be the best model, as it incorporates the advantages of Lasso and Ridge. Additionally, in the ideal case, genes in the same pathway should be viewed as a group and selected together. With regards to this aspect, the grouping effect of the elastic net can be a bonus.

Data	**Description**	Note
copy_number_loss_status	whether a gene has a copy loss in each sample (binary)	if = 1, the gene is inactivated (Y = 1)
pancan_mutation_freeze	whether a gene is deleterious mutated in each sample(binary)	
pancan_rnaseq_freeze	gene expression level in each sample (numerical)	input data (X)
sample_freeze	disease kind of each sample	for analysis use
mutation_burden_freeze	mutation burden freeze of each sample	for analysis use
vogelstein_cancergenes	gene description	for analysis use

Table 1: data description

In our project, we used the pan-cancer RNA-seq data to build a classifier to predict whether the gene TP53 is inactivated. We managed to conduct sparse logistic regression and sparse support vector machine with three kinds of regularization (Lasso, Ridge and Elastic Net) on the training data. And the optimal values of the hyper-parameters were found using cross validation. Then we tested the models in the test data set and compared their performance (i.e. AUC, accuracy, etc.).

Our results showed logistic regression is a better model and elastic net performs better than lasso and ridge regularization in gene data analysis, with a higher test AUC (93.6%) and reasonable proportion of selected variable (13.2%). In addition, according to our further analysis, this elastic net model has biological significance and good interpretability.

2. Data

2.1 Dataset

We used the pan-cancer RNA-seq data from Pan-Cancer Atlas of TCGA (`https://gdc.cancer.gov/about-data/publications/pancanatlas`), which contains data about the expression level of specified genes of samples. Complementary data of copy number can be accessed through figshare.com (`https://figshare.com/articles/dataset/TCGA_PanCanAtlas_Copy_Number_Data/6144122`). The data description can be seen in Table 1. All the data sheets (except the gene description) have the same samples, representing by the same row names (sample barcode).

The original dataset consists of 9074 samples with 33 kinds of cancers, where each sample contains 20500 genes. Table 2 lists the top 10 frequent diseases, referred in `samplefreeze`. The main data we used is the `pancanrnaseqfreeze` data, recording gene expression level of thousands of genes in each sample. The gene expression is measured by genomic analysis techniques, including high-throughput DNA sequencing and other advanced bioinformatics techniques.

Diseases	Number
BRCA	981
LGG	507
UCEC	507
LUAD	502
HNSC	487
THCA	480
PRAD	479
LUSC	464
BLCA	398
STAD	383

Table 2: Top ten diseases in the data

2.2 Data Preprocessing

We selected 21 diseases (BLCA, BRCA, UCS, LGG, etc.) which have more than 15 samples in each class, and 95% samples in both classes. For variables (genes), we selected 8034 most variably expressed genes by median absolute deviation. To keep the model from relying on the expression of gene TP53, we deleted the RNA-seq data of TP53. We also dropped outlier samples whose mutation burden is over five standard deviations of the samples. Last, we normalized the gene expression to by subtracting the mean then being divided by standard deviation of samples.

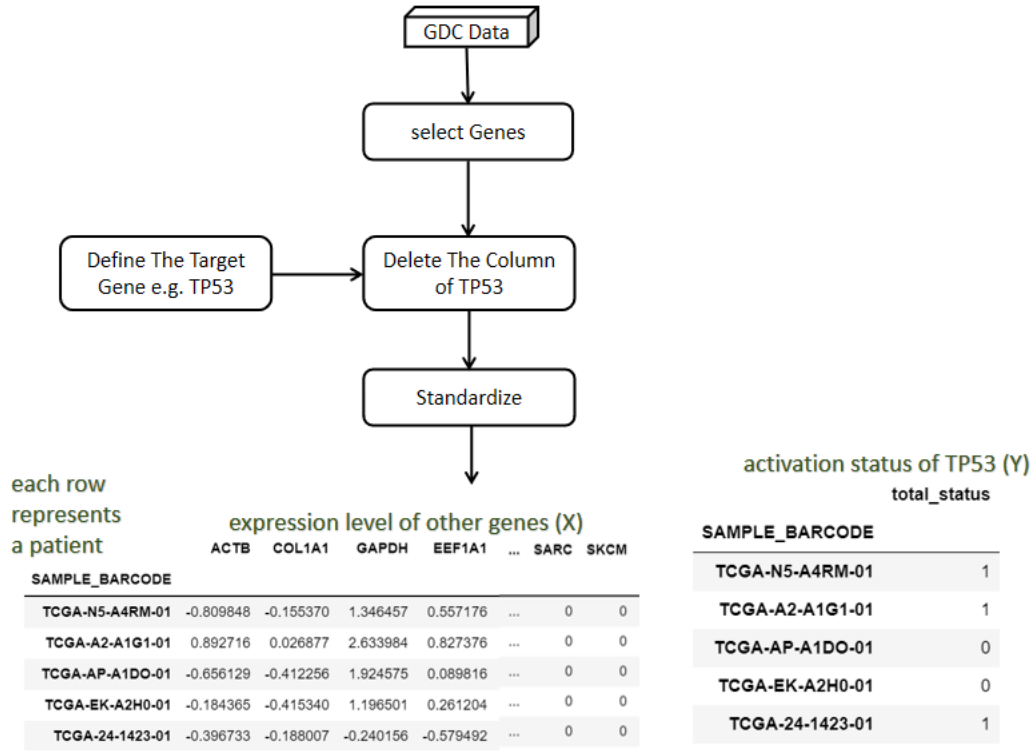
The preprocessed data X contains 6746 rows and 8034 columns. Each column represents the standardized gene expression level of a specific gene. Each row represents the data of one sample (one patient's data). The output is a column of binary values of TP53 inactivation status (1 denotes inactivation and 0 denotes activation).

To clarify, we summarized the preprocessing process into a graph (Fig 1).

2.3 Data Analysis

After preprocessing the data, we calculated the sample average gene expression of each gene (column mean) and plotted the histogram (Fig 2). Notice that there are several outliers (0.2% of all the genes) whose mean is larger than 0.02 and we deleted them when plotting. These outliers are all genes whose standardized expressions are either 0 or 1, such as SARC in the figure. Other than the outliers, the plot is approximately in normal distribution, which indicates our standardization is effective.

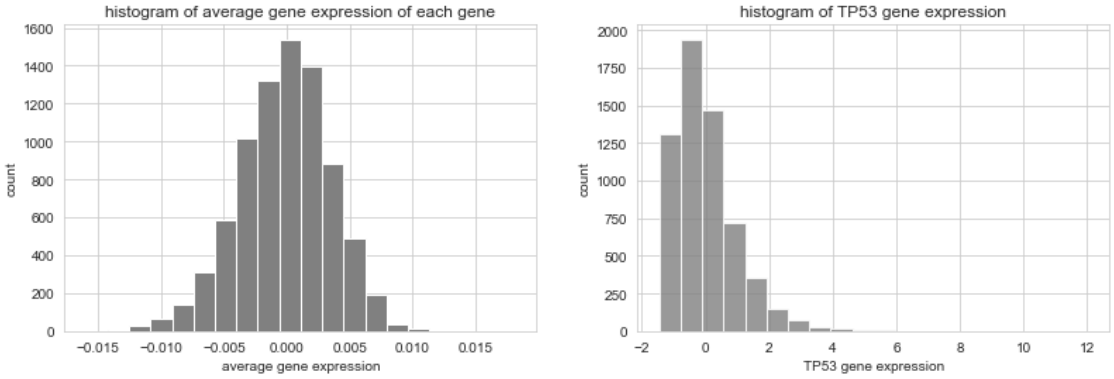
In addition, we plotted the distribution of TP53 gene expression. The distribution is right skewed but mainly concentrated in (-1,1).



SAMPLE_BARCODE	ACTB	COL1A1	GAPDH	EEF1A1	...	SARC	SKCM
TCGA-N5-A4RM-01	-0.809848	-0.155370	1.346457	0.557176	...	0	0
TCGA-A2-A1G1-01	0.892716	0.026877	2.633984	0.827376	...	0	0
TCGA-AP-A1DO-01	-0.656129	-0.412256	1.924575	0.089816	...	0	0
TCGA-EK-A2H0-01	-0.184365	-0.415340	1.196501	0.261204	...	0	0
TCGA-24-1423-01	-0.396733	-0.188007	-0.240156	-0.579492	...	0	0

SAMPLE_BARCODE	total_status
TCGA-N5-A4RM-01	1
TCGA-A2-A1G1-01	1
TCGA-AP-A1DO-01	0
TCGA-EK-A2H0-01	0
TCGA-24-1423-01	1

Figure 1: Data Preprocessing



(a) (b)

Figure 2: Data Analysis

3. Methodology

3.1 Sparse Logistic Regression

Logistic regression is a generalized linear regression analysis model, mostly used in classification problems. It has the advantages of simple design ideas, easy implementation and strong interpretation ability, thus performing well in many tasks.

Logistic regression is used to find the odd ratio in the presence of multiple independent variables x (Sperandei, 2014). It can be writtern as

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_m x_m$$

where π indicates the probability of an event and β denotes the coefficients.

It uses Sigmoid function to map any value to a value between 0 and 1, and then uses a threshold classifier to output a discrete binary result, so that it can predict the probability of event occurrence. In cancer gene analysis, whether a gene is inactivated is predicted according to the probability of logistic regression.

However, in gene analysis, a significant problem to be solved is that the number of genes (p) is much larger than the number of samples (n). Simple logistic regression cannot be performed on cancer gene data due to high-dimensionality of data (Bielza et al., 2011). In contrast, Sparse Logistic Regression, which adds a penalty term to the loss function of the original logistic regression, is capable of performing variable selection and estimating parameters simultaneously (Shevade and Keerthi, 2003).

Suppose our data has n samples and p explanatory variables. The value of x_{ij} represents the value of j-th explanatory variable of the i-th sample. Let $y_i \in \{0:1\}$ be the predicted value for the sample i.

Then the positive probability of one sample is calculated by

$$P\left(y_i = 1 \mid X_i\right) = f\left(X_i\beta\right) = \frac{1}{1+e^{-\beta X_i}}$$

where β is the coefficient vector of each variable. Our goal is to find the β which minimizes the loss function.

The loss function in original logistic regression is

$$L(\beta \mid X) = -\sum_{i=1}^{n} y_i\left(X_i\beta\right) + \left(1-y_i\right)\log\left(1-f\left(X_i\beta\right)\right)$$

In the sparse logistic regression, our loss function becomes

$$L(\beta \mid X) + \lambda u(\beta)$$

where $\lambda u(\beta)$ is the penalty term, in which u is the penalty function and λ is the penalty factor, a hyper-parameter controlling the bias-variance trade-off between fitting and regularization. As the λ increases, the penalty term is of greater importance in the variable coefficient estimation.

Model regularization techniques involve methods such as Lasso, Ridge, and Elastic Net. Among them, Lasso and Elastic Net incorporate embedded-based variable selection in the

process of parameter estimation. It is widely acknowledged that Lasso and Elastic Net can play an important role in the context of cancer gene classification.

3.2 Regularization: Lasso And Ridge

When the penalty term is in the absolute value format (or 1-norm format), it is a Lasso regression estimate and we call the penalty L1 penalty. Alternatively, when the penalty term is in the square format (or 2-norm format, euclidean distance), it is a Ridge regression estimate and L2 penalty is used. The above two regression estimate of a linear model can be defined respectively as

$$\widehat{\beta}^{lasso} := \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$$\widehat{\beta}^{ridge} := \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where the positive penalty factor, $\lambda > 0$, is a tuning parameter which controls the bias-variance trade-off. Ridge regression has a closed form solution for the parameter estimates. It can be considered as an update of least square estimate, shrinking all the coefficients in the original least square estimate but cannot reduce them to zero. It keeps all the variables in the model, which has an incline of overfitting.

Instead, Lasso can perform a direct variable selection. Although it does not have an explicit form solution, it can compress some coefficients to zero. The limitation of Lasso is that it may select too few variables and exclude some important variables out of the model.

In the cancer gene study, for a dataset with n samples and p genes ($p >> n$), Lasso can keep at most n variables. It may probably fail in selecting the genes in the same path together. So Lasso is still not the optimal model, despite its desirable variable selection characteristics.

3.3 Regularization: Elastic Net

In the Elastic Net, the penalty term is the convex combination of the lasso and the ridge regularization.

$$\widehat{\beta}^{EN} := \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left[(1-\alpha)\beta_j^2 + \alpha |\beta_j| \right] \right\}$$

When $\alpha = 0$, the elastic net method becomes the ridge regularization. When $\alpha = 1$, the elastic net method becomes the lasso regularization.

Based on the Theorem 1 mentioned in the paper (Zou and Hastie, 2005), if x_i and x_j are highly correlated, the difference between the coefficient paths of predictor i and predictor j is almost 0 (see Appendix for the complete theorem). The elastic net regularization has a grouping effect, which can select genes of the same path together.

In summary, Elastic Net addresses the problem of high-dimensionality ($p >> n$) and has the unique advantage of the grouping effect. In cancer gene data analysis, we chose the elastic net as the optimal regularization method to build the classification model.

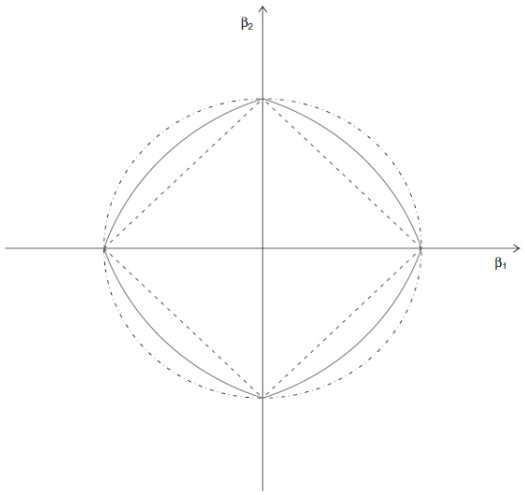


Figure 3: Two-dimensional contour plots. The outside circle is the shape of the ridge penalty; The inside diamond is the contour of the lasso penalty; the middle one is contour of the elastic net penalty with $\alpha = 0.5$): we see that singularities at the vertices and the edges are strictly convex; the strength of convexity varies with α (figure from Zou and Hastie (2005))

From contour plot in the Figure 3, we can see the sharp corner shape of lasso and elastic net provides the characteristic of variable selection. The contour of ridge is too smooth to abandon any variables.

3.4 Sparse SVM

3.4.1 Traditional SVM

Support Vector Machine is a popular supervised model in traditional machine learning.

Suppose we are classifying data into 2 classes. For a p-dimensional data, we need a $p-1$ dimensional hyperplane to divide the data into two group (Fig 4). And we hope to maximize the shortest distance of data points in both sides to the hyperplane. Let $w^{\mathrm{T}}x-b=0$ denote the hyperplane. And $w^{\mathrm{T}}x-b=\pm 1$ is the hyperplane consisting the support vectors.

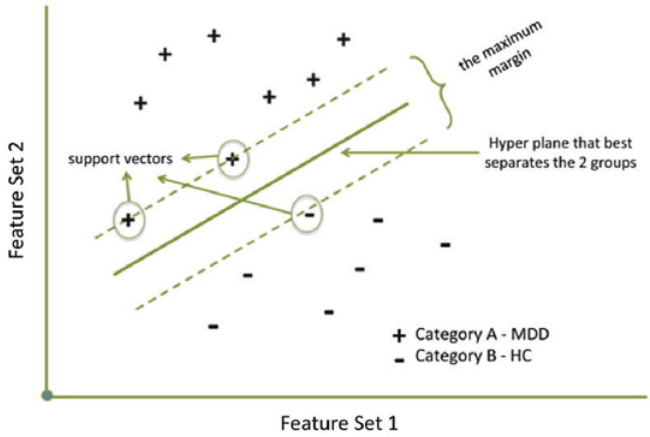


Figure 4: Illustration of SVM (figure from Schnyer et al. (2017))

In order to maximize the distance between the hyperplanes, we just minimize w, that is, minimize $\frac{1}{2}w^2$ equivalently.

3.4.2 SVM WITH HINGE LOSS

In 2002, Schölkopf et al. (2002) first introduces SVM with hinge loss. Including the regularization, the objective loss function becomes

$$\sum h\left(y_i\left(x_i'\mathrm{w}+\mathrm{b}\right)\right)/\mathrm{n}+\lambda^* penalty$$

where h represents the hinge loss function $h = max(0, 1-t)$, and x' is the standardized version of x $x' = (x - mean(x))/sd(x)$.

We can substitutes the penalty function with the same Ridge, Lasso and Elastic Net regularization function as we discussed above. This method is computationally efficient, using Newton's coordinate descent algorithm to compute (Chapelle, 2007). It is widely applied in a large range of statistical learning.

4. Results

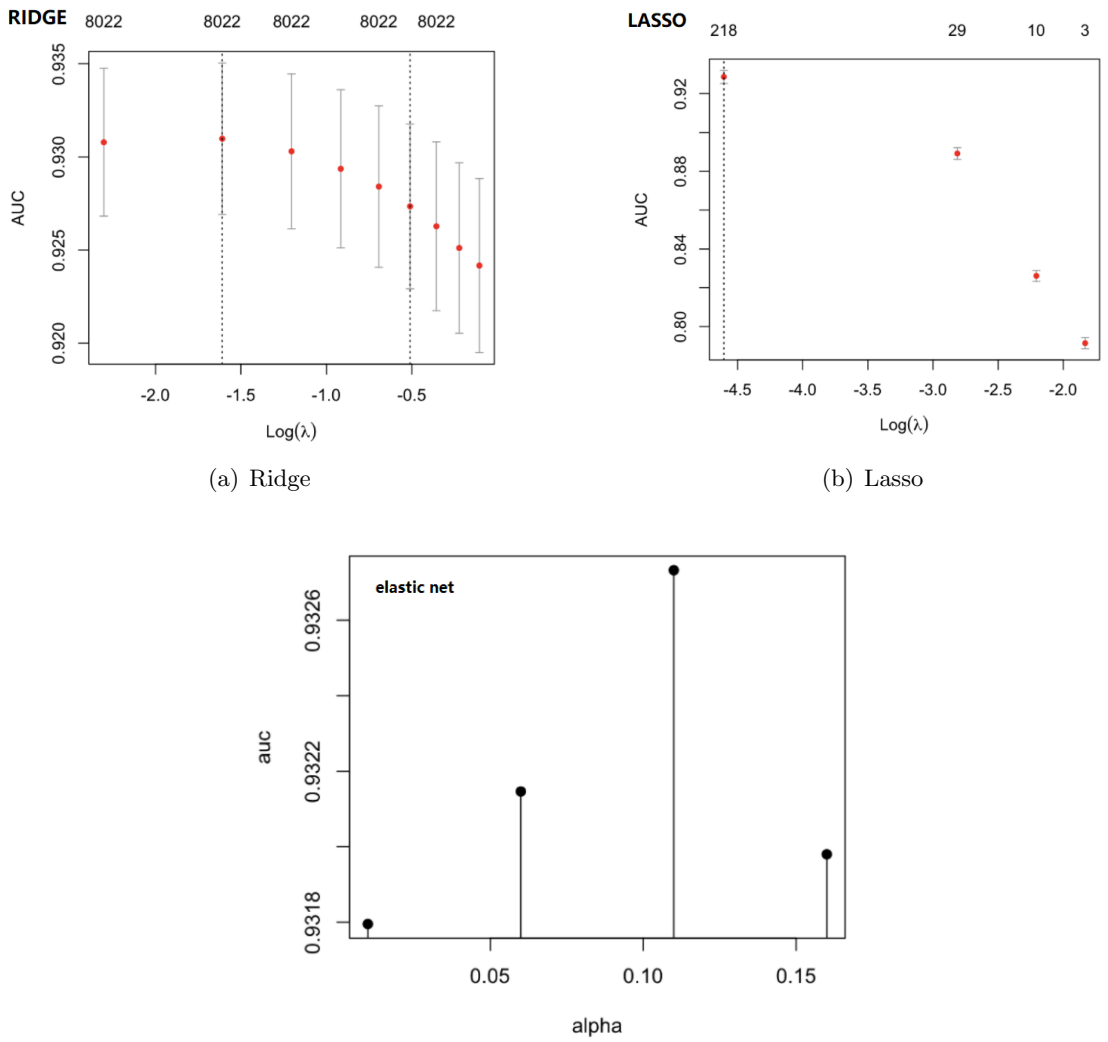
In the context of cancer gene analysis, we did three experiments in total.

4.1 Classification Performance Comparison

In the first experiment, we compared the performance in 6 conditions (sparse logistic regression and sparse SVM with Ridge, Lasso and Elastic Net respectively).

First we used cross validation to explore the optimal value of penalty factor in all the models. Note that we need to search for the optimal values of elastic net factor (α) and penalty factor (λ) at the same time. This involves the idea of grid search.

In `R`, we use package `cv.glmnet` and `cv.sparseSVM` for cross validation. We also tried `Python` with the package `sklearn` and `GridSearchCV` to do the same process.



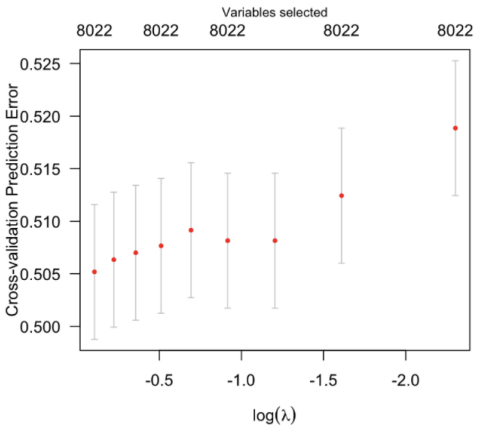
(a) Ridge

(b) Lasso

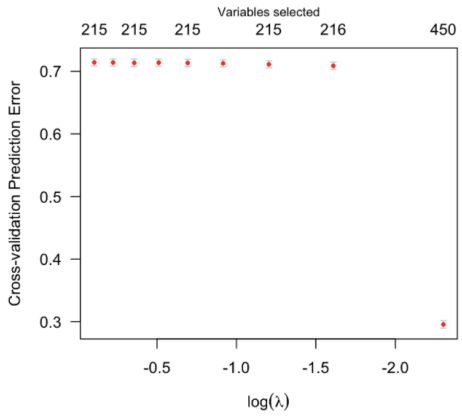
(c) Elastic Net

Figure 5: Logistic Regression: Ridge, Lasso and Elastic Net Cross-Validation Result - the first two are plots of AUC at each lambda along with upper and lower standard error bars, x-axis is the log of the penalty factor lambda; y-axis is the AUC; the top row is the number of variable selected; In the third plot, x-axis is the elastic net parameter, alpha; y-axis is the AUC.

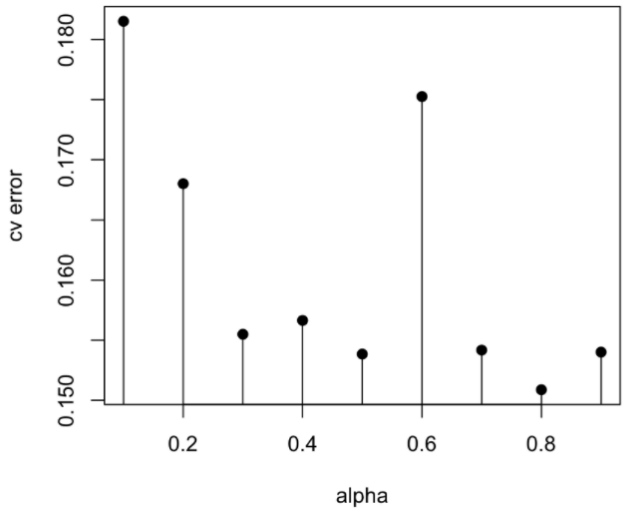
In logistic regression, we use the metric of AUC (the area under the receiver operating characteristic) for cross-validation. We hope to find the parameter which can maximize the AUC. The result is shown in Fig 5. When searching the best parameter pair (λ, α) for elastic net, we first fix the alpha and find the best model for each alpha. Then we record the corresponding optimal lambda and the AUC.



(a) Ridge



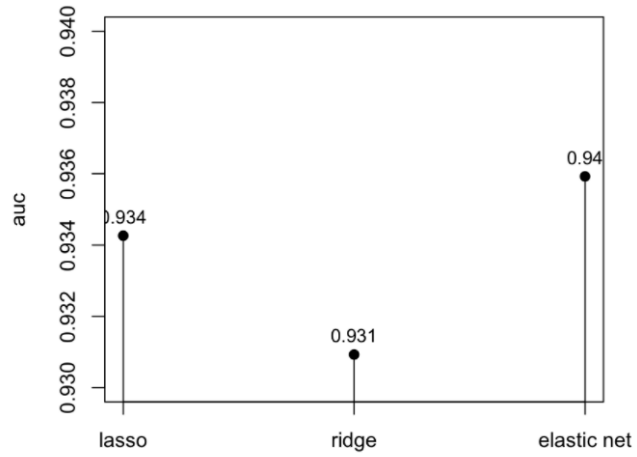
(b) Lasso



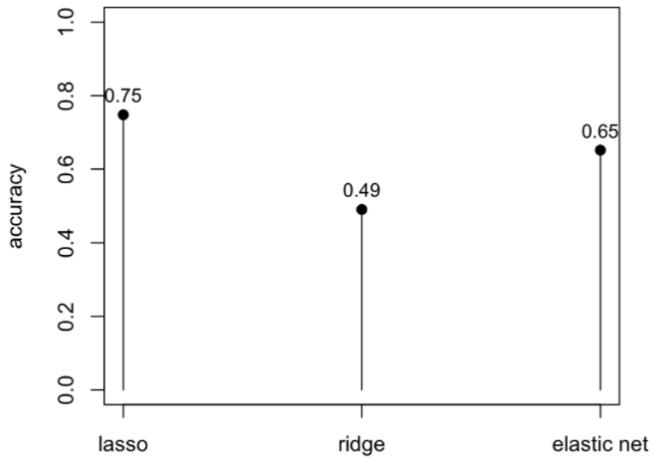
(c) Elastic Net

Figure 6: SVM: Ridge, Lasso and Elastic Net Cross-Validation Result - the first two are plots of mean cv errors at each lambda along with upper and lower standard error bars, x-axis is the log of the penalty factor, lambda; y-axis is the CV error; the top row is the number of variable selected; In the third plot, x-axis is the elastic net parameter, alpha; y-axis is the CV error.

In SVM, due to the limitation of R package, we use the metric of misclassification error instead of AUC for cross-validation. We hope to find the parameter which can minimize the misclassification error. The result is shown in Fig 6. When searching the best parameter pair (λ, α) for elastic net, we first fix the alpha and find the best model for each alpha. Then we record the corresponding optimal lambda and the CV error.



(a) Logistic Regression



(b) SVM

Figure 7: best CV model performance comparison

After finding the optimal parameter(s) for each model, we tested our model on the test dataset. For evaluation, we took AUC and the accuracy for the measurement metrics. The optimal parameter result and the performance of each model are listed here:

Model	Regularization	Optimal Values	Measurement
Logistic Regression	Ridge	lambda = 0.2	AUC=0.9309
	Lasso	lambda = 0.01	AUC=0.9343
	Elastic Net	lambda = 0.0339, alpha = 0.11	AUC=0.9359
SVM	Ridge	lambda = 0.8	Accuracy = 0.490
	Lasso	lambda = 0.1	Accuracy = 0.748
	Elastic Net	lambda = 0.136, alpha = 0.8	Accuracy = 0.652

Table 3: optimal values for parameters and model performance comparison

From the Table 3, we can see that overall Logistic Regression performs much better than SVM. Although using different measurement(AUC in LR and Accuracy in SVM), the difference in figure is large enough to conclude that logistic regression is a better approach for cancer data analysis.

In logistic regression, Elastic Net is the best model with the largest AUC in the test data set. This result agrees with our expectation. Additionally, Ridge is the method with the largest AUC in the training data set. This makes sense because Ridge includes almost all the variables, which contributes to a rather complex model and is easy to overfit. The success of Elastic Net is largely attributable to the good balance of bias and variance trade-off.

We conducted the confusion matrices in SVM models in Fig 9 and compared model performance in statistical figures (accuracy, precision, recall, F1-score) in Table 4. Obviously, the performance of SVM is poor regardless the regularization method in general, especially Ridge, which is affected heavily by the fact of $p >> n$. In sparse SVM, Lasso is the best model with the largest accuracy and F1-score.

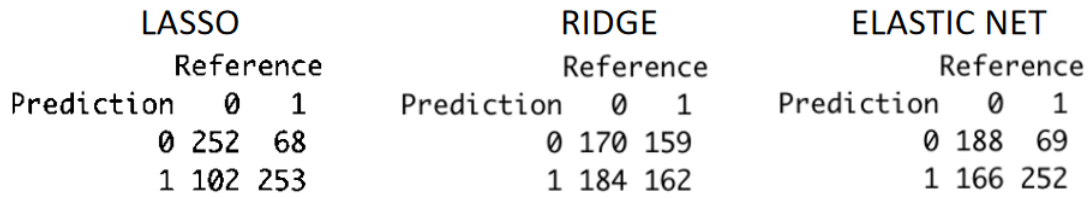
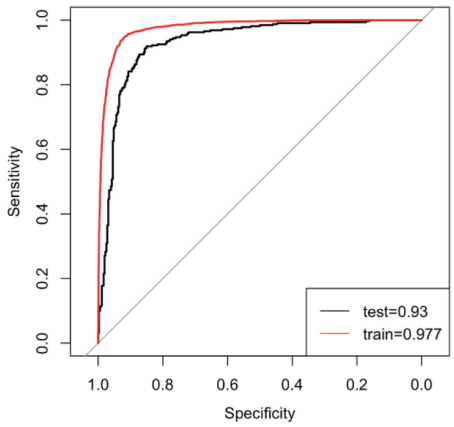


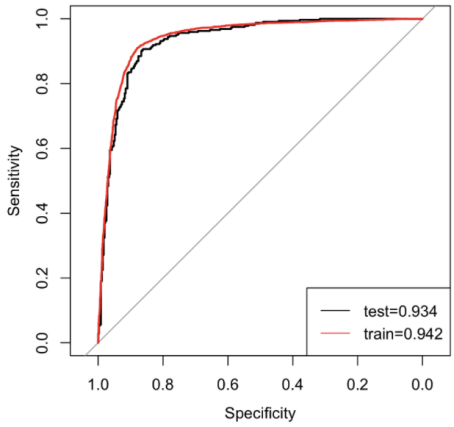
Figure 9: confusion matrices in SVM

	Lasso	Ridge	Elastic Net
Presicion	0.7875	0.5167	0.7315
Recall	0.7119	0.4802	0.5311
F1-score	0.7487	0.4981	0.6233
Accuracy	0.7481	0.4919	0.6519

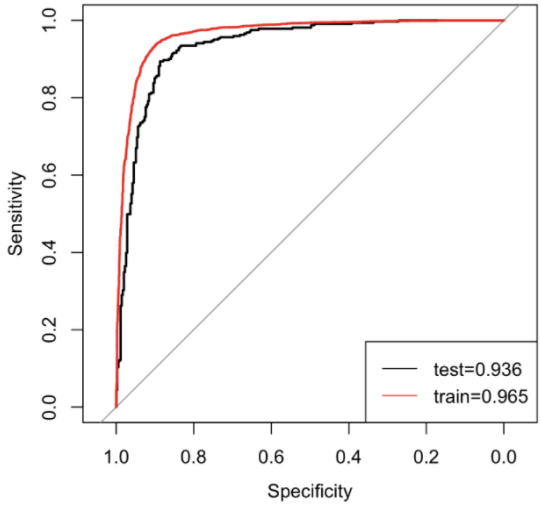
Table 4: model performance in sparse SVM



(a) Ridge



(b) Lasso



(c) Elastic Net

Figure 8: ROC curves of sparse Logistic Regression

In general, the performance of SVM is poor regardless the regularization method.

4.2 Variable Selection Comparison

After obtaining the optimal parameter in each model, we can collect the number of total variable selected by each model. The result can be seen in Table 5.

Clearly, Ridge contain almost all the variables and does not perform a direct variable selection. In contrast, Lasso and Elastic Net can select a portion of relevant variables. However, due to the fact that Lasso can at most keep n (i.e. the number of samples) variables, it may easily lose some valuable variables. For instance, Lasso only keeps approximately 2 percent of all the variables in the sparse logistic regression, which is too risky. In contrast,

Regularization Method	Classification Model	Variables Selected
Ridge	Logistic Regression	8022 (99.8%)
	SVM	8022 (99.8%)
Lasso	Logistic Regression	239 (2.97%)
	SVM	458 (5.70%)
Elastic Net	Logistic Regression	1063 (13.2%)
	SVM	478 (5.94%)

Table 5: The number of total variable selected by each model

Elastic Net remains 13 percent of variables, which is a relatively reasonable amount. As discussed above, sparse logistic regression with the elastic net regularization is the most suitable approach for gene data analysis.

4.3 Related Genes Analysis

In the above experiments, we have discovered the prediction result and the coefficient of each variable (each gene). It is also of great significance to verify that our result has biological significance, even at the molecular level. We need to make sure our model has not only statistical meaning, but also biological meaning in practice.

The top ten genes with the highest absolute weight in the pan-cancer TP53 inactivation classification model (using sparse logistic regression with the elastic net regularization) are gene RPS27L, DDB2, AEN, BAX, MDM2, etc. (Table 6). Among these variables, gene RPS27L, DDB2, AEN, BAX, MDM2, CDKN1A, XPC, FDXR and RRM2B are all direct target genes of TP53. Take gene RPS27L for example, it is found that gene TP53 directly induces the expression of a ribosomal protein, RPS27L, which in turn promotes apoptosis He and Sun (2007). In addition, it is found that the gene deletion removes the entire coding region of MPDU1 along with neighboring genes including TP53 in Kato III human gastric cancer cells Bennett et al. (2018). Consequently, the top ten genes with the highest coefficients are all highly related to gene TP53 at the biological level, which verifies the interpretability of our model.

5. Conclusions

In the above three experiments, we built different models in the context of TP53 inactivation classification. After comprehensive analysis, we made the following conclusion:

- It is highly possible to use the expression level of other genes to predict the activation status of TP53. We can build some traditional machine learning models, such as logistic regression and SVM.
- After comparing the performance of sparse logistic regression and sparse SVM with different regularization methods and we concluded sparse logistic regression is the better model in terms of cancer gene data analysis.

Gene	Weights	Abs
RPS27L	-0.259	0.259
DDB2	-0.239	0.239
AEN	-0.238	0.238
BAX	-0.217	0.217
MDM2	-0.197	0.197
CDKN1A	-0.170	0.170
XPC	-0.133	0.1334
FDXR	-0.123	0.123
MPDU1	-0.111	0.111
RRM2B	-0.105	0.105

Table 6: Top ten related genes and coefficients

- In our project, elastic net is the best regularization technique with the ability of variable selection, the power of parameter estimation and the unique advantage of grouping effect. It can address the dilemma of high dimensional gene data and achieve a good balance in the bias-variance trade-off.
- Variable selection is of paramount significance in gene data study. In our project, keeping only the related genes in the model is of a great concern. This idea is involved in the data preprocessing and the post-experiment analysis.
- Related gene analysis further confirmed the biological significance of our project. Our model has high interpretability and has the potential to exert profound influence in future clinical trials and various cancer mechanism research.

6. Future Work

Admittedly, our work has some inevitable limitation.

Due to the limitation of time and computational resources, when doing the cross validation to search the optimal hyper-parameter value, we can only explore few values in a small range. It is highly possible that our model is not the optimal one. We acknowledge that our final classifier may not be optimal. But we are confirmed that sparse logistic regression with elastic net is an ideal approach for cancer gene data inactivation classification.

Additionally, in this project, we only explored one particular gene, TP53. And our sample size is not large enough for practical application. This limited the generalizability of the results.

There is much work we can do in the future:

- Try different target gene and verify the conclusion in different dataset.
- Try different classification models with different advanced regularization methods (such as adaptive elastic net) to find if there are some better approaches for gene data modeling.
- Elastic net has some weakness that it is trained so slow because it has two parameters to optimize. We can mitigate the negative influence by trying some adaptation of elastic net. E-ENDPP (Xu et al., 2018) is a method worthy trying to implement.

- The data we used includes all kinds of cancers, that is, it is a pan-cancer data. We can filter the data according to the disease and compare the gene inactivation model performance on single-cancer data and pan-cancer data. If the model is performed better on pan-cancer data, then it can be confirmed that that particular gene is more suitable for pan-cancer analysis for its role in multiple cancers.

7. Contribution

We two built a classification model in a regression course :)

Bingyi Jiao: Research; R Coding; Part of Report writing

Yuqing Zhang: Data Analysis; Report writing; Part of Coding

We want to express our thanks for all the amazing work done by the professor and TAs in this course!!

References

Daniel C Bennett, Aurelie Cazet, Jon Charest, and Joseph N Contessa. Mpdu1 regulates ceacam1 and cell adhesion in vitro and in vivo. *Glycoconjugate journal*, 35(3):265–274, 2018.

Concha Bielza, Víctor Robles, and Pedro Larrañaga. Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, 38(5):5110–5118, 2011.

Olivier Chapelle. Training a support vector machine in the primal. *Neural computation*, 19 (5):1155–1178, 2007.

Fang Chen, Wenge Wang, and Wafik S El-Deiry. Current strategies to target p53 in cancer. *Biochemical pharmacology*, 80(5):724–730, 2010.

P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature reviews cancer*, 4(3):177–183, 2004.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

H He and Y Sun. Ribosomal protein s27l is a direct p53 target that regulates apoptosis. *Oncogene*, 26(19):2707–2716, 2007.

David M Schnyer, Peter C Clasen, Christopher Gonzalez, and Christopher G Beevers. Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor mri measures in individuals with major depressive disorder. *Psychiatry Research: Neuroimaging*, 264:1–9, 2017.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1): 12–18, 2014.

Yitian Xu, Ying Tian, Xianli Pan, and Hongmei Wang. E-endpp: a safe feature selection rule for speeding up elastic net. *Applied Intelligence*, 49:592–604, 2018.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

8. Appendix

For the completeness of the report, here we quote the theorem (Zou and Hastie, 2005) which interprets the grouping effect:

Theorem 1 *Given data* (y, X) *and parameters* λ_1, λ_2*, the response* y *is centered and the predictors* X *are standardized. Let* $\hat{\beta}(\lambda_1, \lambda_2)$ *be the elastic net estimate. Suppose that* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$*. Define*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{|y|_1}\left|\hat{\beta}_i(\lambda_1,\lambda_2) - \hat{\beta}_j(\lambda_1,\lambda_2)\right|$$

then $D_{\lambda_1,\lambda_2}(i,j) \leqslant \frac{1}{\lambda_2}\sqrt{\{2(1-\rho)\Big\}}$*. where* $\rho = x_i^T x_j$*, the sample correlation.*

The unitless quantity $D_{\lambda_1,\lambda_2}(i,j)$ describes the difference between the coefficient paths of predictors i and j. If x_i and x_j are highly correlated, i.e. $= 1$, theorem 1 says that the difference between the coefficients paths of predictors i and predictors j is almost 0. The upper bound in the above inequality provides a quantitative description for the grouping effect of the elastic net.