Somatic Proteomics and Genomics Analysis of TCGA cancer by Regression and Feature Selection

Jing Wang

1 Introduction

Diverse immune response and infiltration in tumor micro-environment contributes to varying tumor development across individuals, often having important implications in predicting response to immuno-therapy. Somatic mutation drivers may affect immune infiltration through (1) being a direct neoantigen (2) increase mutational burdens and total number of neoantigens (3) activate other pathways/tumor cell features specifically triggering immune response.

Leukocyte fraction varied substantially across immune subtypes. Hence, it is important to understand the Leukocyte fraction and their correlation with gene symbols. In this work, we utilize multiple regression methods and feature selection for this problem. On a data set with 8,594 patients, we successfully predict their Leukocyte fractions and achieve mean square error 0.017. In the same time, combining the regression model with feature selection technique, we find the gene symbols most corrected with Leukocyte fraction.

Data set. The TCGA cancer data set has 8,594 patients, each of which is described by 300 features consisting of 299 gene symbols and 1 cancer type. There are totally 30 unique types of cancers in our data set, including OV, GBM, LUAD and so on. The distribution of cancer types are shown in Figure 1. It is shown that the number of patients for each cancer type is in the range of $[34, 911]$. There are an average of 286 patients for each caner type. We use a 30-dimensional zero-one vector to represent the cancer type feature where the position of one indicates the cancer type. In this way, the data set is with 329-dimensional (299+30) feature space.

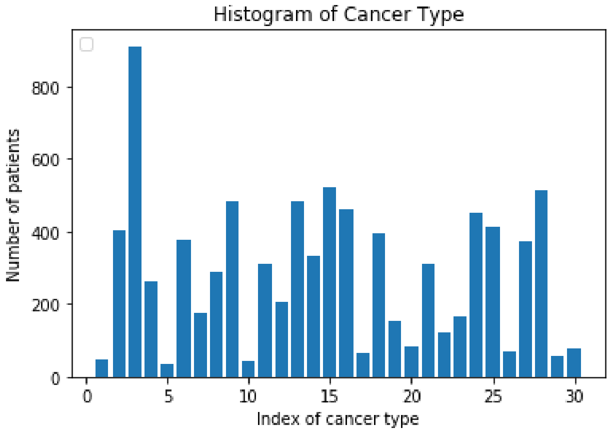

Figure 1: The distribution of patients with different cancer types.

Goal. We aim to predict the Leukocyte fraction based on the gene symbols and cancer type of the patients. As the target variable Leukocyte fraction is a continuous value, the most straightforward way is to utilize regression model for prediction.

2 Method

Notations. Let us instate several pieces of notation that are involved throughout the paper. We write bold capital letters such as $\mathbf{X}$ for matrices and its i-th column vector is denoted as $\mathbf{x}_i$. We use bold lowercase letters, such as $\mathbf{w}$, to denote a vector. The ℓ_1-norm and ℓ_2-norm of a vector $\mathbf{w}$ are denoted by $||\mathbf{w}||_1$ and $||\mathbf{w}||_2$ respectively. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ (where n is the number of data points, and d is the feature dimension) and a target vector $\mathbf{y} \in \mathbb{R}^n$, we aim to learn a coefficient vector $\mathbf{w} \in \mathbb{R}^d$ which leads to an approximation $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ as close to the groundtruth $\mathbf{y}$ as possible.

Regression methods. Here we present a brief introduction of the regression methods that are employed in this paper.

- Ridge regression learns ridge coefficients by minimizing a penalized residual sum of squares,

$$\min \frac{1}{2n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \alpha||\mathbf{w}||_2^2, \tag{1}$$

  where $\alpha > 0$ is a complexity parameter that controls the shrinkage of the ridge coefficients.

- Lasso is a linear model with ℓ_1-norm regularizer on the coefficients. The objective function aims to learn a sparse parameter which minimizes the least-squares penalty as follows,

$$\min \frac{1}{2n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \alpha||\mathbf{w}||_1, \tag{2}$$

  where α is the constant that balances the least-squares loss and ℓ_1-norm constraint on $\mathbf{w}$.

- Elastic Net is a linear regression method that linearly combines the ℓ_1-norm and ℓ_2-norm penalties of the Lasso and ridge regression. It overcomes the limitations of Lasso when dealing with high-dimensional data with few data points. The objective function is given by:

$$\min \frac{1}{2n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \alpha_1||\mathbf{w}||_2 + \alpha_2||\mathbf{w}||_2^2, \tag{3}$$

  where the parameters $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ are tuned by cross-validation.

- Bayesian ridge regression notes that the minimizer of loss function in Equation (1) can be considered as the posterior mean of a model where

$$p(\mathbf{w}|\lambda) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}_d), \tag{4}$$

  where α is chosen from gamma distribution, and $\mathbf{I}_d$ is the identity matrix.

Feature Selection. To explore the correction among gene symbols and the Leukocyte fraction, we employ two feature selection methods to select features with high correlation with the Leukocyte fraction. The feature selection methods are described as follows.

- Variance Threshold removes the features that are either 0 or 1 in more than p percent of the samples.

- F regression computes the correlation between each feature dimension and the target variable, that is

$$\frac{(\mathbf{x}_i - 1/n\sum_{i=1}^n \mathbf{x}_i) \cdot (\mathbf{y} - 1/n\sum_{i=1}^n \mathbf{y}_i)}{var(x_i) \cdot var(y)}, \tag{5}$$

  where $\mathbf{x}_i$ is the ith feature dimension of the data matrix $\mathbf{X}$, var is the standard variance of $\mathbf{y}$.

3 Experiments

We randomly choose two third of patients for training and the rest for testing. We repeat the experiments 50 times and report the average performance as result. Specifically, in each iteration, we run regression methods including Ridge regression, Lasso, Elastic net and Bayesian ridge regression to learn their coefficient vectors. We also implement the feature selection methods such as Variance threshold and F regression to select most important features. The parameter of Variance threshold is set as 0.8, that is, the features with over 80% missing will be removed. The parameter of F regression is set as 10, to select the top-10 most related features.

Evaluation Metric. We use mean squared error and variance score to evaluate the performance of regression methods. The mean square error represents the squared loss between predicted result and groundtruth target variable. The variance score is the proportion of the variance in the dependent variable that is predictable from the independent variables. We and report the average result in Table 1, Figure 2 and Figure 3.

As shown in Table 1, the Elastic Net achieves the best performance in terms of mean squared error 0.0179 and variance score 0.1965.

Table 1: **Experimental results using different regression models.** Smaller mean squared error or larger variance score indicates better performance.

Evaluation Metric	Ridge regression	Lasso	Elastic net	Bayesian ridge regression
Mean Squared Error	0.0185	0.0221	0.0179	0.0192
Variance Score	0.1705	0.0069	0.1965	0.1290

Figure 2 presents the coefficient vectors learned by different regression models. As Lasso employs ℓ_1-norm regularizer on the coefficient vector, most coefficients are zero and there are only three non-zero coefficients. Elastic net perfectly balances the penalties of ℓ_1-norm and ℓ_2-norm which lead to the best performance.

Figure 3 plots the histogram of coefficient vectors learned by different regression models. We also show the coefficients of features selected by Variance threshold (marked in black color) and F regression (marked in light green). We discover that the regression methods assign higher values to the coefficients of features selected by feature selection methods. For example, in terms of Lasso, the feature selection methods Variance threshold and F regression as well as Lasso believe “TNFAIP3” is the most important feature. As the features of our data set is sparse and binary, Variance threshold only selects one important feature “TNFAIP3”. In reach random iteration, the number of samples for training is different, F regression tends to select the same features, including “ATR”, “CARD11”, “CDKN1B”, “CHEK2”, “FAM46D”, “TET2”, “MYD88”, “CTCF”, “MYD88”, “TNFAIP3”. It demonstrates that F regression is a stable algorithm. That is the reason that we choose F regression for our feature selection problem.

4 Conclusion and Future Works

We discovered that among 300 gene symbols, only a small subset of features are highly correlated to the Leukocyte fraction of TCGA cancer in light of the results in Table 1. It also demonstrates that feature selection is of critical importance in understanding the TCGA cancer. By learning a sparse regression model, we obtain perfect prediction of Leukocyte fraction of patients with TCGA cancer. We expect to build a comprehensive understanding of proteomics and gene data. It is interesting to explore more applications of machine learning technique for somatic proteomics and genomics analysis.

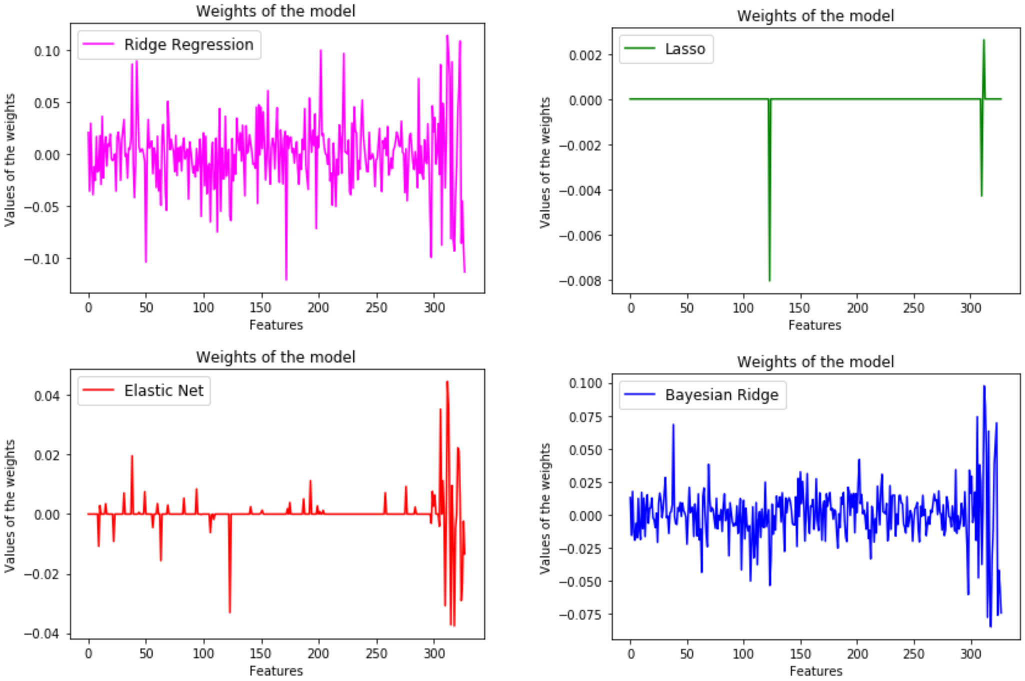

Figure 2: Weights learned by various regression models.

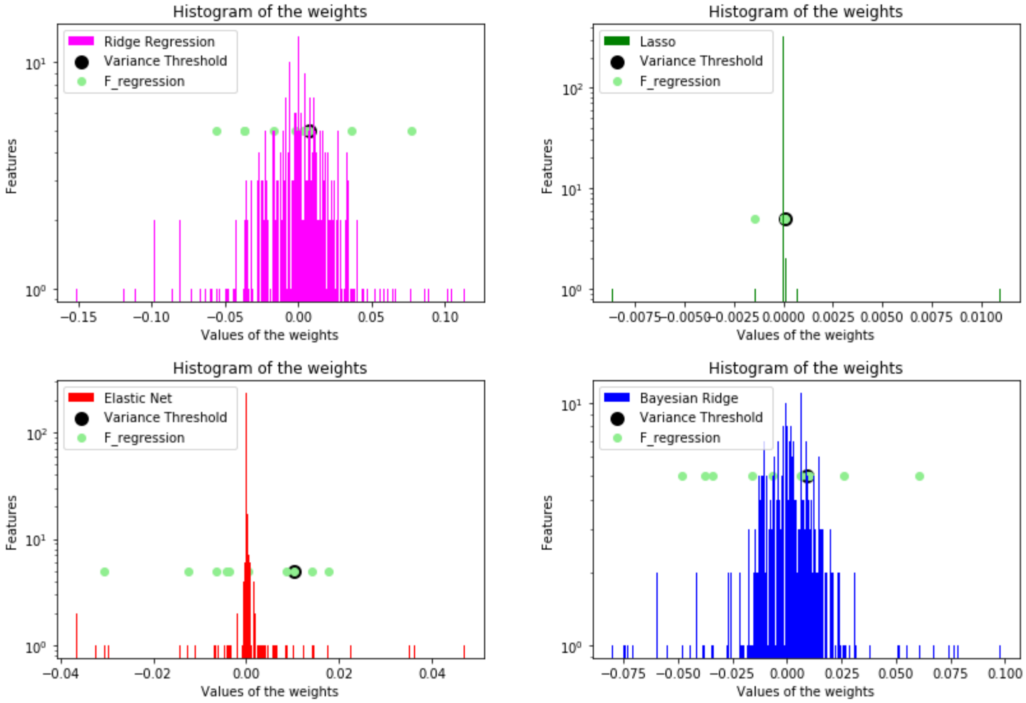

Figure 3: Histogram of weights from multiple models, including Lasso, Ridge regression, Elastic Net and Bayesian ridge regression. The coefficients of important features selected by Variance threshold and F regression are marked in black dot and light green dots.