International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Sparse Bayesian Machine Learning with application to NBA Data (Python & R)

Yu Wang[1], Yaoxuan Luan[2]

[1]College of Business, North Dakota State University, 811 2nd Ave N, Fargo, ND 58102, USA

[2]Department of Computer Science and Software Engineering, Auburn University, 23, Samford Hall, Auburn, AL 36849, USA

Abstract: ***Machine Learning is one of the hot searches in the search engine and is very useful in a variety of areas and subjects. The definition of Machine learning given by Wikipedia is the study of computer algorithms that improve automatically through experience. Mathematical equations and statistical computations also play crucial roles in the entire machine learning process. Statistical models like regression and classification help in Supervised Learning. Other processes will work for Unsupervised Learning and Reinforcement Learning. Other than pure statistics, which focus more on understanding data in terms of models, Machine Learning focus higher on prediction. This article focusses on the prediction perspective of the Machine Learning process while considering the dimension reduction using the sparsity property. The LASSO (Tibshirani, 1996) method provides a sharp power in selecting significant explanatory variables and has become very popular in solving big data problems. A simulation study was conducted to test the power of the model. For application, NBA data was considered. A prediction of the 2019 postseason bracket is given by learning the historical postseason team statistics. The accuracy of the bracketing could be evaluated.***

Keywords: Machine Learning, Bayesian, Sparsity, Dimension Reduction

1. Introduction

One important area in machine learning is the prediction. Researchers are working on selecting the better model in reducing the prediction variance. Ordinary linear regression is one of the classic models in the supervised learning process. However, the ordinary linear model will encounter the overfitting problem due to the least-squares approach.

The overfitting problem is expressed in Figure 1. The black dots are the data points. The red least-squares line was fitted to describe the linear relationship between the response ($y.2$) and the explanatory variable ($x.2$). If the intention in this machine learning process was to predict over the new dataset marked in green dots, the red regression line is overfitting and yields a high prediction variance for these new dates (green dots). We anticipate a model that was fit over the training dataset (black dots) can provide a blue fitted line, which can also have a relatively small variance toward the testing dataset (green dots).

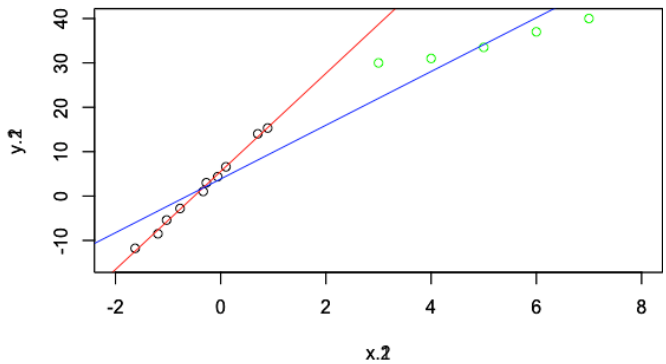

Figure 1: OLS overfitting

One classic approach to reach the blue fitted line is the LASSO model. LASSO stands for the least absolute shrinkage and selection operator. Tibshirani (1996) proposed this LASSO model which minimizes

$$(y - X\beta)'(y - X\beta) + \lambda\|\beta\|_1 \quad (1.1)$$

where λ is the tuning parameter form 0 to $+\infty$. The tuning parameter is the coefficient that controls the power of the shrinkage. $\|\beta\|_1$ is the same as L_1 norm such that $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$. Note that there are a total of p covariates in the model.

Due to LASSO's desired geometric property, it can significantly reduce the dimension of the covariates. Sometimes the LASSO model will even work when $p \gg n$, the number of covariates is larger than the sample size. Figure 2 shows the geometric property of Lasso under the constraint of two dimensions L_1 norm. We can also interpret Figure 2 as $|\beta_1| + |\beta_2| \leq t$ for some $t > 0$.

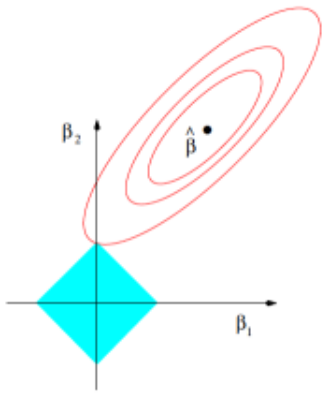

Figure 2: The geometric expression for two dimensions β

The choice of λ is usually through cross-validation. If λ goes to 0, the penalty portion will go to 0. Then there is no difference between the LASSO model and the OLS (ordinary least squares) model. If λ goes to $+\infty$, the penalty portion will dominate the whole equation. The estimated coefficients

will all go to 0. R package like "glmnet" will provide the estimated tuning parameter.

A penalized logistic regression is considered to get the prediction of the binary response data. Figure 3 shows the logistic curve, which is bounded between 0 and 1. The curve's property makes the logistic regression a nature fit for binary response data.

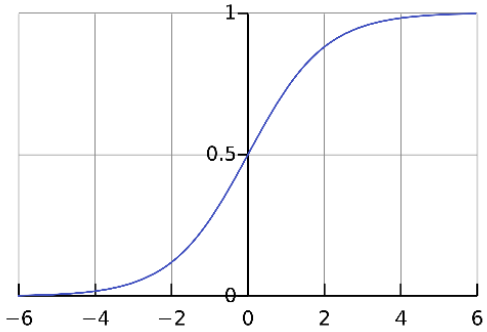

Figure 3: The standard logistic function

Penalized logistic regression imposes a penalty to the logistic model for having too many covariates. This process results in shrinking the coefficients of the less contributive variables toward zero. If LASSO is incorporated in this process, the coefficients of some less contributive variables are forced to be exactly zero. The other popular definition of this process is dimension reduction.

Combine the LASSO logistic regression with the Bayesian informative prior model to directly apply to NBA postseason data. An evaluation method can be applied to check the prediction accuracy of the 2019 NBA postseason bracket.

2. Methodology

The whole methodology part can be separated into two sections. We first start with the LASSO logistic regression. It is treated as the model selection process. Once we get the final reduced model from this LASSO logistic process, we can then fit a Bayesian informative prior model to get the predicted bracket. A comparison between the real bracket and the prediction bracket will get this whole process adequately evaluated.

2.1 LASSO Logistic Regression

We intend to keep a model with only the most significant variables or covariates. The ordinary logistic regression is of the following format

$$log(\frac{p}{1-p}) = \boldsymbol{X\beta} \quad (2.1)$$

With the penalty involved, logistic LASSO is an effective method to reduce large dimensional covariates. Given the logistic model (2.1), the negative log-likelihood with L_1 regularization takes the form

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i(X\boldsymbol{\beta}) - log(1+e^{X\boldsymbol{\beta}})\} + \lambda\|\beta\|_1 \quad (2.2)$$

From the form of the log-likelihood function (2.2), we see that the maximization of the log-likelihood function, a monotone function, is the same as minimizing the negative log-likelihood. The rest part is the same as the basic LASSO regression, λ controls the magnitude of the penalty term and is estimated by cross-validation. Figure 4 shows an example of the selection process of the best-estimated λ.

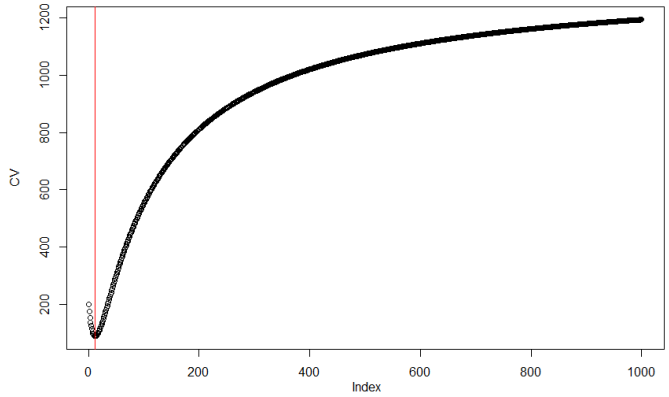

Figure 4: Example CV values with $\lambda \in (0, +\infty)$

Once λ is determined, the estimated coefficients will be reported with the lowest predicting variance. Several coefficients will be set exactly to zero due to the sparsity of LASSO.

2.2 Bayesian Model with Informative Prior

For the Bayesian model, two parts need to be considered, the prior and the likelihood. The posterior is proportional to the product of prior and the likelihood. For the likelihood, since the research interest is the binary response, the setting transfers to a random variable y, taking values 0 or 1, follows a Bernoulli distribution with probability p. Refer to the logistic regression (2.1) and let $\eta_i = X\beta$, the probability p can be expressed as the following equation:

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \quad (2.3)$$

The distribution of $f(y_i|p_i)$ is as follows:

$$f(y_i|p_i) = (p_i)^{y_i}(1-p_i)^{1-y_i}$$
$$f(y_i|p_i) = \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)^{y_i}\left(\frac{1}{1+\exp(\eta_i)}\right)^{1-y_i} \quad (2.4)$$

For a full Bayesian approach, the posterior is stated as $f(\beta \mid data) \propto f(\beta) \times f(data \mid \beta)$. As to the likelihood, $f(data \mid \beta)$, it is determined with the product of the distribution $f(y_i|p_i)$. For the prior $f(\beta)$, we need to involve the indirect prior information based on the historical probability information. Using the delta method, we can have the following distribution for the vector β:

$$\beta \sim N\left[(X'X)^{-1}X'log\left(\frac{\hat{p}}{1-\hat{p}}\right), (X'X)^{-1}X'n\hat{p}(1-\hat{p})X(X'X)'\right] \quad (2.5)$$

Note that the above Normal distribution is a multivariate normal distribution and X refers to as the design matrix.

With all these setups, we can sample from the posterior distribution which is

$$f(\beta|data) = MVN \times \prod_{i=1}^{n} \left(\frac{exp(\eta_i)}{1+exp(\eta_i)}\right)^{y_i} \left(\frac{1}{1+exp(\eta_i)}\right)^{1-y_i} \quad (2.6)$$

The estimated coefficients can then be determined. For the application to the real data, a predicted bracket can be provided by involving the current season's team statistics.

3. Simulation

A simulation process is provided to check the performance of the LASSO logistic regression. The simulation work involved fifteen covariates; hence there will be fifteen $\beta's$ related to the covariates. Ten out of the fifteen covariates will be set to zero. The covariates' data will come from a variety of distributions.

3.1 Simulation on X's

The data should differ from a variety of distributions, and the mean of these data should be away from zero. If the mean of these data is around zero, then it is difficult to distinguish whether it is due to the LASSO sparsity that the later estimated coefficients shrink to zero.

$x_1 \sim N(3,\ 1^2)$
$x_2 \sim N(X_1,\ 1^2)$
$x_3 \sim N(X_2,\ 2^2)$
$x_4 \sim Unif(5,\ 10)$
$x_5 \sim Unif(x_4, x_4 + 3)$
$x_6 \sim N(3.5,\ 1^2)$
$x_7 \sim N(X_6,\ 1^2)$
$x_8 \sim x_4 + x_7$
$x_9 \sim Unif(x_8, x_8 + 3)$
$x_{10} \sim Unif(x_9, x_9 + 1)$
$x_{11} \sim N(5,\ 1^2)$
$x_{12} \sim N(x_{11},\ 1^2)$
$x_{13} \sim N(x_{12},\ 2^2)$
$x_{14} \sim Unif(5,10)$
$x_{15} \sim Unif(x_{14}, x_{14} + 3)$

3.2 Simulation on β's

Table 1: Assumed true coefficient values

Coef	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True Value	4	-4	5	7	-6	0	0	0
Coef	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	
True Value	0	0	0	0	0	0	0	

3.3 Simulation on p's

The following equation calculates the "true" probabilities:

$$p_i = \frac{e^{X\beta}}{1+e^{X\beta}} \quad (3.1)$$

3.4 Simulation on y's

Since we need the binary response, we will set $y = 1$ or $y = 0$ based on the Bernoulli trial with the above probability. The sample size is taken to be 1000.

3.5 Simulation Results

Table 2: Comparison of the coefficient values

Coef	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True Value	4	-4	5	7	-6	0	0	0
Estimated Value	2.5	-2.3	3.1	4.2	-3.6	0	0	0.04
Coef	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	
True Value	0	0	0	0	0	0	0	
Estimated Value	0	-0.1	0	-0.1	-0.02	0	0.13	

We can see that the logistic LASSO model successfully shrinks the unnecessary coefficient down to zero hence reach a dimension reduction effect.

4. Application

NBA postseason team statistics are studied to apply this whole process to live data. The training dataset includes the postseason team statistics from the year 2012 playoffs to the year 2018 playoffs. Table 3 shows all team statistics included in this study.

Table 3: Team statistics list

Fouls Per Game	Games Played	Average Field Goals Made	Average Field Goals Attempted	Field Goal Percentage	Average 3-Point Field Goals Made
Average 3-point Field Goals Attempted	3-Point Field Goal Percentage	Average Free Throws Made	Average Free Throws Attempted	Free Throw Percentage	Offensive Rebounds Per Game
Defensive Rebounds Per Game	Rebounds Per Game	Assists Per Game	Steals Per Game	Blocks Per Game	Turnovers Per Game

With the help of the logistic LASSO regression, some of the team statistics could be removed. This LASSO process will lead to a final reduced model with only important covariates. Table 4 shows the removed team statistics in red color.

Table 4: Team statistics with removed covariates in red

Fouls Per Game	Games Played	Average Field Goals Made	Average Field Goals Attempted	Field Goal Percentage	Average 3-Point Field Goals Made
Average 3-point Field Goals Attempted	3-Point Field Goal Percentage	Average Free Throws Made	Average Free Throws Attempted	Free Throw Percentage	Offensive Rebounds Per Game
Defensive Rebounds Per Game	Rebounds Per Game	Assists Per Game	Steals Per Game	Blocks Per Game	Turnovers Per Game

The estimated tuning parameter λ = 0.0023. The dimension of the covariates reduced from eighteen to seven.

The playoff bracket is seeded based on the regular-season performance. Seeds will be number from 1 to 8 by East or

West teams. Figure 5 is the example bracket for 2019 playoffs (Wikipedia).

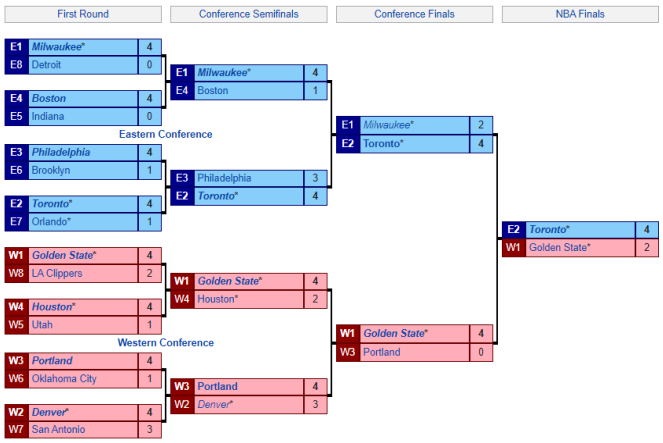

Figure 5: Bracket for 2019 playoffs

For the first round, Milwaukee plays with Detroit. Milwaukee rank seed one and Detroit rank seed eight. The winning probability for Milwaukee based on the historical seeding is involved. The historical winning probabilities are determined by all games matched with those seeding combination. If seed one and seed eight played 20 games at the postseason, and seed one won 19 of those games, then seed one is defined to have 19/20 winning probability over seed eight. The historical winning probabilities will then transfer to the β prior using Delta Method. Table 5 is a comparison of the perdition and the real bracket of the playoffs 2019.

Table 5: NBA 2019 playoffs bracketing

Team	R1	R2	R3	R4
Milwa	Milwa (√)	Milwa (√)	Milwa (Toronto)	Milwa (Toronto)
Detroit				
Boston	Boston (√)			
Indiana				
Phila	Phila (√)	Phila (Toronto)		
Brookln				
Toronto	Toronto (√)			
Orlando				
Golden S	Golden S (√)	Houston (Golden St)	Houston (Golden St)	
LA clippers				
Houston	Houston (√)			
Utah				
Portland	Portland (√)	Portland (√)		
Oklaho				
Denver	San Anto (Denver)			
San Anto				

The prediction accuracy can be evaluated by a single scoring system and as well as a double scoring system (Shen, 2015). For single scoring system, the accuracy is 8/15= 53.33%. For double scoring system, the accuracy is 11/32=34.38%.

5. Conclusions

From the simulation and also the application, it is easy to detect that the sparse machine learning model is beneficial for the prediction or decision making. Furthermore, for the current big data world, it also plays a significant role. When reaching the high dimension problem, the sparsity property can significantly reduce the size of the dimension and yield a more reasonable model.

Current common sparse machine learning models include Ridge Regression, LASSO regression, and Elastic Net. L_q norm could also be considered from a different type of data. Figure 6 is the geometric representation of the typical values of q.

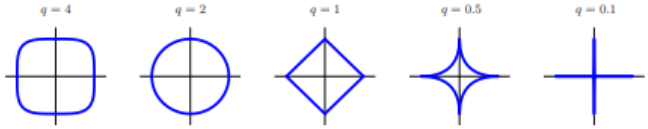

Figure 6: Geometric representation of L_q norm

References

[1] Friedman, J. (2019). R Package 'glmnet'.
[2] Gao, D., Liu, X., Scariano, S. (2020). Ridge and Lasso Regression for Undergraduate Research with Simulation in R. International Journal of Science and Research. Volume 9 Issue 1, Jan 2020.
[3] Hua, S. (2015). Comparing Several Modeling Methods on NCAA March Madness. Unpublished Thesis Paper, North Dakota State University, Fargo, ND.
[4] Oehlert, G. W. (1992), A Note on the Delta Method, The American Statistician, Vol. 36, No. 1, p. 27-29.
[5] Park, T., Casella, G., (2008). The Bayesian Lasso. Journal of the American Statistical Association. Vol. 103, 2008, pp. 681-686.
[6] Shen, G., Gao, D., Wen, Q., Magel, R. (2016). Predicting Results of March Madness Using Three Different Methods. Journal of Sports Research. 2016, Vol. 3, Issue 1, pp. 10-17.
[7] Shen, G., Hua, S., Zhang, X., Mu, Y., Magel, R. (2015). Prediction Results of March Madness Using the Probability Self-Consistent Method. International Journal of Sports Science, 5(4), p. 139-144.
[8] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, pp. 267-288.