Support vector machine Variable selection with Lasso and Fused lasso

Prof. Dr. Tahir R. Dikheel* , Zahraa K. Aswad

Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq

Abstract

The support vector machine (SVM) is a highly accurate and adaptable binary classification technique. It has had considerable success, but its performance can suffer if too many covariates are included. In this article, we used lasso penalizes least squares regression by adding the absolute values of the coefficients ($L_1 - norm$). This penalty's structure promotes sparse solutions (with many variables coefficients equal to 0). We propose the fused lasso as a generalization designed for situations like this. Both the simulation study and colon cancer data example show that proposed methods outperform the other existing methods.

Keyword: support vector machine; variable selection; lasso; fused lasso

Introduction

As a result of the recent advent of new data collection and storage technology, we have seen an explosion in data complexity in a variety of study fields such as genomics, imaging, and finance. As a result, the number of forecasters increases dramatically. However, there are just a few examples available for investigation.(Donoho et al., 2000) in tumor categorization using genomic data; for example, tens of thousands of gene expression levels are employed. Are accessible, although the number of arrays is usually in the tens. The classification of high dimensional data presents several statistical problems, necessitating the development of novel approaches and theories.

In this post, we will look into high-dimensional classification, where the number of variables is large. Diverges with sample size and may be considerably bigger than the sample size. The support vector machine (SVM) is a strong binary classification technique developed by (Vapnik, 1996) excellent precision and adaptability. It has found success in a variety of applications. However, one significant disadvantage of the conventional SVM is that its performance might suffer as a result. If the decision rule has a large number of redundant variables (Friedman et al., 2001).

Variable selection is crucial in the construction of a support vector machine. This approach gives shrinkage for appropriate estimating parameters, good production, and identification of the key variables. The supply of interpretable models distinguishes statistical techniques for variable selection. Variable selection approaches, such as stepwise and best subset selection, may be unstable. (Tibshirani, 1996) presented the least absolute shrinkage and selection operator to solve this problem (**lasso**). (Tibshirani et al. 2005) proposed the fused lasso, , it may be used to functional data, which can be thought of as multivariate data having order on its dimensions. Variables, on the other hand, can be grouped into strongly correlated groups and then a single representative covariate retrieved from each cluster.

Support vector machine:

The support vector machine (SVM) is a big margin classifier that distinguishes between two classes by maximizing the margin between them.When dealing with non-separable data, the soft – margin SVM employs the slack variable to regulate the upper bound of the misclassification error. SVM is a classification approach that employs multidimensional hypotheses and is powered by an optimization algorithm developed from statistical learning theory (Vapnik, 1974). SVM has numerous benefits in handling nonlinear and high-dimensional classification problems, and it has shown good results in pattern recognition, function approaches, and probability density (Shi, 2012). SVM training is a computationally demanding operation, owing mostly to the curved quadratic programming problems associated with the dense Hessian Matrix used during optimization (Godwin, 2013).The SVM method is briefly discussed here "(Burgers, 1998), (Huang, Chen, and Wang, 2006), (SchÖlkopf and smola, 2002)". Let $(\boldsymbol{x_i}, \boldsymbol{y_i}), \mathbf{1} \leq \boldsymbol{i} \leq \boldsymbol{N}$, represent a collection of training data. Where **N** is the quantity of training data. Each datum must the xi∊ $\mathbf{R^d}$ And $\boldsymbol{y_i \epsilon \{-1, 1\}}$ where d is dimension count in terms of input data. SVM tries to locate a hyper plane, it serves as a dividing plane for data categorization in a multi-dimensional space w and b are parameters provided by

$$(\langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle + \boldsymbol{b}) = \boldsymbol{0}, \qquad (\boldsymbol{1}) \qquad \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{N}$$

If there is a hyper plane that meets **Eq** $(\boldsymbol{1})$,

Then there is linear separation.. **W** and **b** may be rephrased as follows in this situation **Eq**$(\boldsymbol{1})$ is transformed into

$$\underset{1\leqslant i\leq N}{\boldsymbol{min}}\, \boldsymbol{y_i}(\langle \boldsymbol{w} \cdot \boldsymbol{x_i} \rangle + \boldsymbol{b}) \geqslant \boldsymbol{1}, \quad (\boldsymbol{2}) \qquad \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{N}$$

Assume that the distance between the data point and the hyper planes $\boldsymbol{1}/||\boldsymbol{w}||$. There is one optimum separating hyper plane (OSH) a monq separating hyper planes. and the distance between them is between two support vector points located on opposite ends of this .The hyper plane is the most extensive because the distance between two support vector points equals $\boldsymbol{1}/||\boldsymbol{w}||^{\boldsymbol{2}}$ may be used to calculate the shortest distance to OSH. $||\boldsymbol{w}||^{\boldsymbol{2}}$ a separating hyper plane’s margin computed as the generalization of the hyper plane is determined by $\boldsymbol{2}/||\boldsymbol{w}||$ ability. Among separating hyper planes, the OSH has the greatest margin **Eq (2)**And Lagrange’s polynomial are used to minimize $||\boldsymbol{w}||^{\boldsymbol{2}}$.Let a stand in for $(\boldsymbol{a_1}, \dots., \boldsymbol{a_n})$. Combining the polynomial of Lagrange ("in the order of N") with $\boldsymbol{Eq}$ $(\boldsymbol{2})$ generates the following maximizing equations

$$\boldsymbol{W}(\boldsymbol{a}) = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{a_i} - \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{a_i a_j y_i y_j x_i x_j} \qquad (\boldsymbol{3})$$

Where $\boldsymbol{a_i} \geq \boldsymbol{0}$ and $\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{y_i a_i} = \boldsymbol{0}$ under constraint .To do this, the quadratic programming approach might be used solve the maximizing issue described above. If a vector is $\boldsymbol{a_0} = (\boldsymbol{a_{10}} \dots \boldsymbol{a_{n0}})$ in maximizing, if **N** satisfies $\boldsymbol{Eq}$ $(\boldsymbol{3})$,then the OSH in terms of $(\boldsymbol{w_0}, \boldsymbol{b_0})$ may be written as follows

$$\boldsymbol{w_0} = \sum_{i=1}^{n} \boldsymbol{a_i^0 y_i x_i} \qquad (\mathbf{4})$$

Where the support vector points must satisfy $\boldsymbol{a_{i0}} \geq \mathbf{0}$ and $\boldsymbol{Eq}$ $(\mathbf{2})$.When looking at expansion in constraint $\boldsymbol{Eq}$ $(\mathbf{4})$ the hyper plane determinant function is written as follows

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{sig\,n}\left(\sum_{i=1}^{n} \boldsymbol{a_i^0 y_i x_i x} + \boldsymbol{b_0}\right) = \mathbf{0} \qquad (\mathbf{5})$$

Most of the time, the data are not linearly separable and must be transferred to a higher dimensional feature space. As a result, if the data cannot be properly categorized .The SVM will then map in current dimensional space .They are then classified in a higher dimensional space .The input data is assigned to a higher dimensional feature by drawing a curve that is not linear in space curve . The OHS is built in the feature space $\boldsymbol{\phi}(\boldsymbol{x})$ can be used in limited $\boldsymbol{Eq}$ $(\mathbf{3})$ by first defining the feature space. As shown below

$$\boldsymbol{W}(\boldsymbol{a}) = \sum_{i=1}^{n} \boldsymbol{a_i} - \frac{\mathbf{1}}{\mathbf{2}} \sum_{i,j=1}^{n} \boldsymbol{a_i a_j y_i y_j \phi}(\boldsymbol{x_i})\boldsymbol{\phi}(\boldsymbol{x_j}) \qquad (\mathbf{6})$$

The existence of Mercer’s theorem may be inferred from asymmetric and positive kernel function $\boldsymbol{K}$ $(\boldsymbol{x}, \boldsymbol{y})$ .As a result, $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})\,\boldsymbol{\phi}(\boldsymbol{y})$ .Assuming that the kernel function K fulfills Mercer’s theorem, and the derived training method is ensured for minimizing

$$\boldsymbol{W}(\boldsymbol{a}) = \sum_{i=1}^{n} \boldsymbol{a_i} - \frac{\mathbf{1}}{\mathbf{2}} \sum_{i,j=1}^{n} \boldsymbol{a_i a_j y_i y_j \phi}(\boldsymbol{x_i})\boldsymbol{\phi}(\boldsymbol{x_j}) \qquad (\mathbf{7})$$

The decision function is expressed as follows:

$$\boldsymbol{f}(\boldsymbol{x}) = \mathbf{sig\,n}\left(\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{a}_{\boldsymbol{i}}\boldsymbol{y}_{\boldsymbol{i}}\boldsymbol{K}\left(\boldsymbol{x}_{\boldsymbol{i}} \cdot \boldsymbol{x}_{\boldsymbol{j}}\right) + \boldsymbol{b}\right) \qquad (\mathbf{8})$$

Kernel function helps the **SVM** in determining the best answer .The polynomial, sigmoid, and radial basis kernels are the most often utilized kernel functions (**RBF**) is the most often used classification method since it can categorize multidimensional data.

Least absolute shrinkage and selection operator (lasso):

(Tibshirani, 1996) presented the least absolute shrinkage and selection operator to solve this problem. This approach offers shrinkage coefficients toward zero and makes certain coefficients precisely zero, attempting to maintain the key variables with substantial impacts. A penalty function was added to the least squares loss function, as seen in the equation below.

$$\boldsymbol{\beta}_{\boldsymbol{lasso}} = \boldsymbol{arg\ min} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{\beta})^{\mathbf{2}} + \boldsymbol{\lambda} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{k}}|\boldsymbol{\beta}_{\boldsymbol{k}}| \qquad (\mathbf{9})$$

$\boldsymbol{\lambda} \geq \mathbf{0}$ Controls the strength of penalty

Thus, they dealt with the issues that arose in the work of the lasso technique, they examined the lasso, and pointed out numerous lasso problems as follows:

1- When $\boldsymbol{p} > \boldsymbol{n}$, the lasso picks approximately n variables.

2- If there is a group of tightly linked variables, lasso will select only one from this grouping and disregard the remaining variables.

Studies have found that the lasso estimator is sometimes inefficient, and the results of the variables selection are inconsistent (Fan & Li, 2001; Zou, 2006).

To overcome this problem, Zou (2006) proposed the adaptive least absolute shrinkage and selection operator (alasso), which penalizes different regression coefficients by different weights. These penalties reflect the size of the coefficient to define the correct model.

Fused lasso:

(Tibshirani et al. 2005) proposed the fused lasso, an expanded variant of the lasso that penalizes least squares regression based on the sum of absolute values ($\boldsymbol{L_1 - norm}$) of the coefficients. The fused lasso penalizes the $\boldsymbol{L_1 - norm}$ of both coefficients and their subsequent differences, as shown below.

$$\widehat{\boldsymbol{\beta}} = \mathbf{argmi\,n} \sum_{i} \left(y_i - \sum_{j} x_{ij}\boldsymbol{\beta}_j \right)^2 \qquad (\mathbf{10})$$

$$\textbf{subject to} \sum_{j=1}^{p} \left|\boldsymbol{\beta}_j\right| \leq s_1 \textbf{ and} \sum_{j=2}^{p} \left|\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}\right| \leq s_2 \qquad (\mathbf{11})$$

Features at the $\boldsymbol{i - th}$ observation, $\boldsymbol{i = 1, \ldots, n}$, and $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are tuning parameters. The fused lasso follows feature ordering. As a result, it may be used to functional data, which can be thought of as multivariate data having order on its dimensions. Variables, on the other hand, can be grouped into strongly correlated groups and then a single representative covariate retrieved from each cluster.

"Consider how many degrees of freedom"are employed in a "fused lasso fit" $\boldsymbol{\hat{y} = X\hat{\beta}}$

A $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are changed. "(Efron et al., 2002)" examined a concept number of "degrees of freedom" based on Stein's (1981) formula:

$$\mathbf{df}(\hat{\boldsymbol{y}}) = \frac{\mathbf{1}}{\boldsymbol{\sigma}^2}\sum_{i=1}^{n} \mathbf{cov}(\boldsymbol{y_i}, \hat{\boldsymbol{y}}_i) \qquad (\mathbf{12})$$

Where $\boldsymbol{\sigma^2}$ denotes $\boldsymbol{y_i}$variance with $\mathbf{X}$ held constant, while cov signifies covariance with X held constant. df ($\hat{\boldsymbol{y}}$)reduces to p for a conventional multiple linear regression with $\boldsymbol{p < n}$ predictors. Now consider the situation of an orthonormal design. The lasso estimators are essentially soft threshold estimations. of $(\boldsymbol{X^T X = I})$, and "(Efron et al., 2002)" Demonstrated that the "degrees of freedom" are proportional to the number of coefficients that are not zero. They also demonstrated this is applicable to the "LAR and lasso estimators under the positive cone situation", implying that the estimations are monotone as a function of the $\boldsymbol{L_1 - bound\ s1}$. In the orthonormal situation, the evidence is straightforward: it use Stein's formula.

$$\frac{\mathbf{1}}{\boldsymbol{\sigma}^2}\sum_{i=1}^{n} \mathbf{co\,v}(\boldsymbol{y_i}, \boldsymbol{g_i}) = \boldsymbol{E}\left\{\sum_{i} \frac{\boldsymbol{\partial g(y)}}{\boldsymbol{\partial y_i}}\right\} \qquad (\mathbf{13})$$

Where $\boldsymbol{y = (y_1, y_2, \dots, y_n)}$ "is a multivariate normal vector with mean and covariance I", and $\boldsymbol{g(y)}$ is an estimator, a nearly distinguishable "function from $\mathbf{R}^{\mathrm{N}}$ to $\mathbf{R}^{\mathrm{N}}$". We rotate the basis for the lasso with orthonormal design so that $\boldsymbol{X = I}$, and therefore from equation

$$\boldsymbol{g(y) = sgn(y_i)(|y_i| - \gamma 1\,)} \qquad (\mathbf{14})$$

If the i-th component is non-zero, the derivative $\boldsymbol{\partial g.(y)|\partial y_i}$ equals 1; otherwise, it equals to $\mathbf{0}$. As a result, the degrees of freedom are equal to the number of non-zero coefficients.

Simulation:

In this part, we will do a simulation study to demonstrate our behavior.

The suggested model was created by fusing Lasso with the R package. And compare it to existing models such as Lasso and Fused lasso . Our comparison is based on the average sum of errors (MSE) criterion and the categorization error criterion. In addition, we utilized the mean to assess the performance of prediction accuracy for several models. Where samples with a volume of (n=100,150,200) were created for the purpose of generating data in the following format.

$$\mathbf{y_i = sign(b + x\ .\ w + error)} \qquad (\ 15\ )$$

(b) denotes a constant bias amount equal to (3). Where (x) is the result of a multivariate normal distribution. And (w) is the weight vector, where (k = 7) and =0.25 are the values. The random error term was produced using the conventional normal distribution. The results provided in the theoretical section were achieved and compared using the MSE and MIS. The results are stable after 1000 replications of the experiment. The number of factors that influence the outcome (p = 81,100,256). The sample count (n) was divided into two groups (g1,g2). We chose the best approach using the MSE and MIS criteria, with the least valuable way for the MSE and MIS criteria being the best.

Table (1): explains the results of MSE when $\rho = 0.25$ and K=3

n	P	MSE

		Lasso	**Fused lasso**
100	81	1.345	0.224
	100	1.268	0.029
	256	0.930	0.712
150	81	1.639	0.482
	100	1.906	0.440
	256	1.126	0.520
200	81	1.960	0.717
	100	1.674	0.581
	256	1.311	0.459

Table (2): explains the results of MIS when $\rho = 0.25$ and K=3

n	**P**	MIS	
		lasso	**Fused lasso**
100	81	0.138	0.101
	100	0.129	0.101
	256	0.142	0.134
150	81	0.166	0.098
	100	0.173	0.118
	256	0.117	0.103
200	81	0.169	0.109
	100	0.173	0.127
	256	0.135	0.107

Table (1) and table (2) show that the results when $\rho = 0.25$, n=100,150,200, and k=3 with its three weights (0.5,1.5,1). We note that the proposed method gives a best results compering with other methods depend on the values of (Mse and MIS), especially that the method fused lasso has the smallest values for MSE and MIS .

Conclusion:

In this paper, It is proposed fused lasso with svm . The methods are illustrated using a simulation study . The results showed that the proposed method is more stable than lasso method in comparison. Thus, this proposed method is capable of handling this data. Method fused lasso with p = 81,100.256 and sample size n=100,150,200 gave the best results compared with method lasso.

Reference

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745-6750.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, *1*(2000), 32.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2002). Least angle regression (Technical Report). Statistics Department, Stanford University..

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning, volume 1 Springer series in statistics Springer.

Caruana, G., Li, M., & Liu, Y. (2013). An ontology enhanced parallel SVM for scalable spam filter training. *Neurocomputing*, *108*, 45-57.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, *33*(4), 847-856.

Schölkopf, B., Smola, A. J., & Bach, F. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shi, T. (2012). Research on the application of e-mail classification based on support vector machine. In *Frontiers in Computer Education* (pp. 987-994). Springer, Berlin, Heidelberg.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1135-1151.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91-108.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Vapnik, V., & Chervonenkis, A. (1974). Theory of pattern recognition.

Zhang, H. H., Ahn, J., Lin, X., & Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *bioinformatics*, *22*(1), 88-95.

Zhang, X., Wu, Y., Wang, L., & Li, R. Variable selection for support vector machines in high dimensions. *Europe PMC free article*.