Discussiones Mathematicae
Probability and Statistics 25 (2005) 103–113

STACKED REGRESSION WITH RESTRICTIONS

Tomasz Górecki

Faculty of Mathematics and Computer Science,
Adam Mickiewicz University,
Umultowska 87, 61–614 Poznań, Poland

e-mail: drizzt@amu.edu.pl

Abstract

When we apply stacked regression to classification we need only discriminant indices which can be negative. In many situations, we want these indices to be positive, e.g., if we want to use them to count posterior probabilities, when we want to use stacked regression to combining classification. In such situation, we have to use least squares regression under the constraint $\beta_k \geq 0, k = 1, 2, \ldots, K$. In their earlier work [5], LeBlanc and Tibshirani used an algorithm given in [4]. However, in this paper we use a more general algorithm given in [6].

Keywords: stacked regression, regression with restrictions, mixed regression.

2000 Mathematics Subject Classification: 62H30, 62J05.

1. Introduction

Wolpert [7] presented an interesting idea of combining classifiers known as "stacked generalization". He was not searching for the best classifier in the set of all classifiers but their linear combination. Since each single one has some advantages, their combination is reasonable. Wolpert's proposal was translated into the language of the statistics by Breiman [2]; he called it "stacked regression". Then LeBlanc and Tibshirani [5] used it to construct a combined classifier in discriminant analysis. A combined classifier is a linear combination of estimated posterior probabilities; coefficients of this combination are estimated by the stacked regression. These coefficients may be negative and so may discriminant indices. Very often we need posterior probabilities, so we can use another method to estimate regression coefficients.

Suppose that a training sample $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$ has been collected by sampling a population P consisting of K subpopulations or classes $G_1, \ldots, G_K$. The ith observation in $\mathbf{z}$ is a pair denoted by $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is a p-dimensional feature vector and y_i is the label for recording class membership. The corresponding pair for an unclassified observation is denoted by $\mathbf{z}_0 = (\mathbf{x}_0, y_0)$. In this case $\mathbf{x}_0$ is observed but the class label y_0 is unobserved. The object of classification is to construct a classification rule for predicting the membership of an unclassified feature vector $\mathbf{x}_0 \in P$. An automated classifier can be viewed as a method of estimating the posterior probability of membership in G_k. The classification rule assigns $\mathbf{x}_0$ to the group with the largest posterior probability estimate. We denote the posterior probability of membership in G_k by

$$p_k(\mathbf{x}_0) = P(y_0 = k|\mathbf{x}_0), k = 1, 2, \ldots, K.$$

Let us assume that we have c different classifiers. An estimate of $p_k(\mathbf{x}_0)$ obtained by jth classifier is denoted by

$$\hat{p}_k^j(\mathbf{x}_0), k = 1, 2, \ldots, K; j = 1, 2, \ldots, c.$$

Stacked regression is a generalization of the sum rule. We have c classifiers and K classes so we have Kc estimates which are arranged in the vector:

$$\hat{\mathbf{p}}(\mathbf{x}_0) = \left(\hat{p}_1^1(\mathbf{x}_0), \ldots \hat{p}_K^1(\mathbf{x}_0), \ldots, \hat{p}_1^c(\mathbf{x}_0), \ldots, \hat{p}_K^c(\mathbf{x}_0)\right)'.$$

These estimates are being arranged to the stack as rows of the $\mathbf{P}$ matrix. Let $\mathbf{u}_h$ be a vector having a 1 in the ith position if the observation falls in class h and 0 otherwise, so

$$u_{i,h} = \begin{cases} 1, & \text{if } y_i = h, \\ 0, & \text{if } y_i \neq h. \end{cases}$$

The stacked regression model has the form:

$$\mathbf{u}_h = \mathbf{P}\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_h,$$

where $\boldsymbol{\beta}_h$ is a $Kc \times 1$ vector of unknown stacked regression coefficients and $\boldsymbol{\varepsilon}_h$ a N-vector of errors assumed to follow a normal distribution with mean zero and common covariance matrix. Least squares estimate of $\hat{\boldsymbol{\beta}}_h$ can be obtained by solving the following equation:

$$\mathbf{P}'\mathbf{P}\boldsymbol{\beta}_h = \mathbf{P}'\mathbf{u}_h$$

with respect to $\boldsymbol{\beta}_h$.

Estimates of posterior probability received from the classifiers are summed up to one, so

$$\forall j = 1, 2, \ldots, c, \quad \sum_{k=1}^{K} \hat{p}_k^j = 1.$$

Hence columns of the $\mathbf{P}$ matrix are undergoing c linear constraints, $\mathbf{P}$ is not a full column rank and $\mathbf{P}'\mathbf{P}$ is a singular matrix. We can use the Moore - Penrose generalized inverse of the matrix $\mathbf{P}'\mathbf{P}$ denoted by $(\mathbf{P}'\mathbf{P})^+$ and

$$\hat{\boldsymbol{\beta}}_h = (\mathbf{P}'\mathbf{P})^+\mathbf{P}'\mathbf{u}_h.$$

Given the estimates $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K$ we classify $\mathbf{x}_0$ using the scalar product:

$$\hat{u}_{0,h} = \hat{\mathbf{p}}'(\mathbf{x}_0)\hat{\boldsymbol{\beta}}_h.$$

We are choosing the group with the largest values of $\hat{u}_{0,h}$. These scalar products are called discriminant indices because they are not summing up to one.

2. Stacked regression with restrictions

In many situations we want discriminant indices to be nonnegative, e.g., if we intend to use them as a posterior probability. We can replace negative values by 0 and calculate probabilities with the followig formula:

$$\tilde{u}_{0,h} = \frac{\hat{u}_{0,h}}{\sum\limits_{j=1}^{K} \hat{u}_{0,j}},\ h = 1, 2, \dots, K.$$

The values $\tilde{u}_{0,h}$ are posterior probabilities.

On the other hand, we can use restricted regression. LeBlanc and Tibshirani [5] used an algorithm from [4]. We used a model proposed by Toutenburg and Roeder [6]. This model is more general than the model proposed in [5] and if we choose a large value of p in this algorithm we can obtain values of parameters as in the model described by Lawson and Hanson [4]. Our experiments suggest that $p = 5$ is sufficient. On the other hand, in stacked regression we do not need positive values of $\boldsymbol{\beta}_h$ parameters to receive positive values of discriminant indices. Experiments suggests that is better to choose smaller values for the p parameter.

Assuming there is prior information that the coefficients satisfy the interval constrains

$$r_{L_i} \leq r_{U_i}.$$

Toutenburg and Roeder [6] considered a procedure giving lower bounds for the probabilities of $\beta_i \in [r_{L_i}, r_{U_i}],\ i = 1, \dots, k$ (in our application, taking $p = 5$, the lower bounds will equal 96%, as we shall see). Then the constrains will be expressed as the following stochastic restrictions

$$\mathbf{r} = \boldsymbol{\beta} + \mathbf{v},$$

where $\mathbf{v}$ is a random vector with mean $\mathbf{0}$ and the covariance matrix $(4p^2\mathbf{H})^{-1}$, $\mathbf{H}$ being a $k \times k$ diagonal matrix with principal elements $(r_{U_i} - r_{L_i})^{-2}$, and the components of $\mathbf{r}$ are $(r_{L_i} + r_{U_i})/2$. Thus, for the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is a T-vector of observations on the dependent variable, $\mathbf{X}$ – a $T \times k$ full rank matrix of observations on k independent variables, $\boldsymbol{\beta}$ – a k-vector of coefficients and $\boldsymbol{\varepsilon}$ is a T-vector with distribution $N(\mathbf{0}, \sigma^2\mathbf{W})$ ($\mathbf{W}$ known, σ^2 unknown), we get the mixed regression estimator (MRE):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X} + 4p^2\sigma^2\mathbf{H}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} + 4p^2\sigma^2\mathbf{H}\mathbf{r}\right).$$

Toutenburg and Roeder [6] showed that $\hat{\boldsymbol{\beta}}$ is unbiased.

Now we apply this method to stacked regression. We know that $\mathbf{P}$ matrix is not a full column rank, so it does not fulfill the assumption in the model. Fortunately, we can modify it. When we create the $\mathbf{P}$ matrix, we are using all posterior probabilities and this creates difficulties. We do not have to use all the probabilities because the last one is a linear combination of previous probabilities. In this way, we are reducing the $\mathbf{P}$ matrix to $N \times (Kc - c)$ matrix $\mathbf{P}_1$. This matrix is a full column rank and we can use it.

Now we can formulate a model for stacked regression with restrictions:

$$\mathbf{u}_h = \mathbf{P}_1\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_h,$$

where $\mathbf{u}_h$ is an N-vector of observations on the dependent variable, $\mathbf{P}_1$ is a $N \times c(K-1)$ full rank matrix of observations on independent variables, $\boldsymbol{\beta}_h$ – a $c(K-1)$-vector of coefficients and $\boldsymbol{\varepsilon}_h$ is an N-vector with distribution $N(\mathbf{0}, \sigma_h^2\mathbf{W})$($\mathbf{W}$ known, σ_h^2 unknown). Under our specifications we can formulate constraints in the interval $[0, a]$. Hence

$$0 \le \beta_{hi} \le a, \ i = 1, 2, \ldots, c(K-1).$$

We have

$$\mathbf{r} = \frac{a}{2}\mathbf{1},$$

$$\mathbf{H} = a^{-2}\mathbf{I}$$

and

$$\hat{\boldsymbol{\beta}}_h = \left(\mathbf{P}_1'\mathbf{W}^{-1}\mathbf{P}_1 + 4p^2\sigma_h^2a^{-2}\mathbf{I}\right)^{-1}\left(\mathbf{P}_1'\mathbf{W}^{-1}\mathbf{u}_h + 2p^2\sigma_h^2a^{-1}\mathbf{1}\right).$$

Hence in the stochastic model we have:

$$E(\mathbf{v}) = \mathbf{0},$$

$$Var(\mathbf{v}) = \frac{a^2}{4p^2}\mathbf{I},$$

$$E(\mathbf{r}) = \boldsymbol{\beta}_h,$$

$$Var(\mathbf{r}) = \frac{a^2}{4p^2}\mathbf{I}.$$

By Chebyshev inequality we have:

$$P\left(|r_i - \beta_{hi}| \leq \frac{a}{2}\right) \geq 1 - \frac{1}{p^2}.$$

We can change the constraints but the dependent variable is 0 or 1 so estimations will be in the $[0,1]$ interval. Chaturvedi and Wan [3] proposed using the following estimator of σ_h^2 :

$$\hat{\sigma}_h^2 = \frac{\mathbf{u}_h^{'}\left(\mathbf{W}^{-1} - \mathbf{W}^{-1}\mathbf{P}_1\left(\mathbf{P}_1^{'}\mathbf{W}^{-1}\mathbf{P}_1\right)^{-1}\mathbf{P}_1^{'}\mathbf{W}^{-1}\right)\mathbf{u}_h}{N - c(K-1)}. \tag{1}$$

We used the estimator (1) but the results were not good. The efficiency of regression with restrictions was comparable with the efficiency of regression without restrictions since we got estimators $\hat{\sigma}_h^2 \approx 0$.

As an alternative, since $\boldsymbol{W}$ is unknown, we took

$$\mathbf{W} = \mathbf{I},$$

thus obtaining

$$\hat{\boldsymbol{\beta}}_h = \left(\mathbf{P}_1^T\mathbf{P}_1 + 4p^2\hat{\sigma}_h^2\mathbf{I}\right)^{-1}\left(\mathbf{P}_1^T\mathbf{u}_h + 2p^2\hat{\sigma}_h^2\mathbf{1}\right), \tag{2}$$

where

$$\hat{\sigma}_h^2 = \frac{\mathbf{u}_h^T\Big(\mathbf{I} - \mathbf{P}_1\Big(\mathbf{P}_1^T\mathbf{P}_1\Big)^{-1}\mathbf{P}_1^T\Big)\mathbf{u}_h}{N - c(K-1)}.$$

3. Experiments

We have made experiments on real data sets. As individual methods we used the linear discriminant analysis (lda), quadratic discriminant analysis (qda) and J-nearest neighbors method for $J = 2, 4, 6$ (J-nn). Information about the datasets are presented in Table 1. More information we can find in [1].

Table 1. Information about datasets.

Name	Number of features	Number of classes	Number of instances in classes	Number of all instaces
beetles	2	3	21,21,22	64
blood	3	4	20,20,20,20	80
chemistry	3	4	12,14,11,8	45
crude-oil	5	3	7,11,38	56
fish	4	3	12,12,12	36
football	6	3	30,30,30	90
hayes	5	3	51,51,30	132
iris	4	3	50,50,50	150
irradiation	3	4	6,14,15,10	45
risk	2	3	30,28,29	87
school	2	3	31,28,26	85
thyroid	5	3	150,35,30	215
turtles	6	2	24,24	48
wave	21	3	37,43,45	125
wine	13	3	59,71,48	178

To compare stacked regression and stacked regression with restrictions we have carried out experiments. In Table 2 we have results of this comparison. In the first column we have names of dataset, in the second the error rate (in %) for stacked regression and in columns 3–7 we have the error rate for stacked regression with restrictions, whereas in the last column there is p for which we have the smallest error rate for restricted regression. We used the bootstrap technique to estimate the error rate.

Table 2. Comparison of regressions.

Name	Stacked	Stacked regression with restrictions					p
	regression	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	
beetles	2.92	**2.19**	2.82	3.69	4.57	4.69	1
blood	**78.27**	83.28	85.77	86.46	87.32	87.68	0
chemistry	73,37	**68.82**	68.93	69.26	69.49	70.03	1
crude-oil	18.05	**14.07**	17.18	19.86	20.97	21.84	2
fish	54.27	49.25	**46.34**	47.43	47.48	47.55	2
football	43.53	41.23	**38.55**	38.62	39.21	39.67	1
hayes	38.21	**37.32**	39.18	40.08	41.29	42.23	1
iris	2.86	**2.53**	3.27	3.56	3.66	3.70	2
irradiation	73.64	68.68	**67.95**	68.52	68.80	68.84	1
risk	1.82	**0.94**	1.01	1.01	1.01	10.1	1
school	8.43	**7.52**	7.92	9.23	11.26	12.56	1
thyroid	4.18	**3.87**	5.14	6.22	6.72	6.92	1
turtles	19.75	17.47	**16.43**	17.19	17.30	17.31	2
wave	45.35	44.73	39.73	32.09	29.11	**28.30**	5
wine	2.17	1.27	**1.20**	1.63	1.81	1.93	2
mean	31.12	29.54	29.43	29.66	30.00	30.29	

Experimental results with the bootstrap error estimator show that all but one p stacked regression with restrictions give less mean error rate than stacked regression. Only for one dataset "blood" - restricted regression increases the error rate. This method gives us 3.17% reduction of the error rate if we compare stacked regression with stacked regression with restrictions for the best p. It gives reduction of a relative error rate by about 17.34%.

It is interesting how to decrease the percentage of negative values of $\hat{\beta}_{hi}$ coefficients and indices u_{hi}. Experimental results are presented in Table 3 and Table 4.

Table 3. Percentage of negative values of $\hat{\beta}_{hi}$.

Name	Stacked	Stacked regression with restrictions				
	regression	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$
beetles	46.71	24.31	15.73	8.98	3.02	0.71
blood	48.85	29.38	13.93	6.35	2.55	0.50
chemistry	48.78	24.35	7.98	1.85	0.50	0.20
crude-oil	46.27	21.29	8.93	3.82	1.29	0.04
fish	46.53	22.18	5.69	0.71	0.13	0.00
football	45.82	30.00	14.67	5.47	0.89	0.18
hayes	45.60	34.22	18.93	9.91	6.89	2.76
iris	44.67	32.18	15.64	8.31	2.09	0.31
irradiation	48.38	20.45	8.48	3.73	2.35	1.08
risk	46.49	27.07	13.11	5.56	2.09	0.27
school	47.87	31.64	11.42	3.24	0.62	0.04
thyroid	42.58	27.82	18.36	9.33	4.62	1.69
turtles	42.30	23.50	5.50	0.90	0.00	0.00
wave	43.51	23.91	17.29	4.58	1.20	0.18
wine	44.44	35.51	19.56	7.33	2.71	0.84
mean	45.92	27.19	13.01	5.34	2.06	0.59

Table 4. Percentage of negative values of u_{hi}.

Name	Stacked	Stacked regression with restrictions				
	regression	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$
beetles	39.81	0.09	0.00	0.00	0.00	0.00
blood	24.49	0.27	0.00	0.00	0.00	0.00
chemistry	32.51	0.87	0.00	0.00	0.00	0.00
crude-oil	29.44	7.61	0.20	0.00	0.00	0.00
fish	30.62	1.02	0.00	0.00	0.00	0.00
football	24.07	2.16	0.00	0.00	0.00	0.00
hayes	21.22	5.45	0.07	0.00	0.00	0.00
iris	36.88	0.01	0.00	0.00	0.00	0.00
irradiation	34.57	3.14	0.03	0.00	0.00	0.00
risk	34.05	0.00	0.00	0.00	0.00	0.00
school	31.50	1.72	0.00	0.00	0.00	0.00
thyroid	28.11	1.36	0.01	0.00	0.00	0.00
turtles	21.55	0.00	0.00	0.00	0.00	0.00
wave	36.21	5.57	0.00	0.00	0.00	0.00
wine	36.28	1.61	0.11	0.00	0.00	0.00
mean	30.75	2.06	0.03	0.00	0.00	0.00

As expected the percentage of negative values $\hat{\beta}_{hi}$ decreases quite fast and for $p = 5$ we have only 0.59% negative values. We can see that for $p = 3$ we do not have negative values of u_{hi} and for $p = 2$ we have only 0.03%.

4. Conclusion

We can see that stacked regression with restrictions has some advantages:

- it gives a smaller error rate,
- we do not have to use the generalized Moore - Penrose inverse of matrix,

- we have positive values of u_{hi} so we can use them to count posterior probabilities.

It appears that the best choice of p is 1 or 2. For these values we have the smallest error rate and the percentage of positive values of u_{hi} is enough. We can see that for $p = 1$ the percentage of negative values is decreasing from 30.75% to 2.06%.

References

[1] C. Blake and C. Merz, *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu/ mlearn/MLRepository.html, Univeristy of California, Irvine, Department of Information and Computer Sciences.

[2] L. Breiman, *Stacked Regression*, Machine Learning **24** (1996), 49–64.

[3] A. Chaturvedi and A. Wan, *Estimation of Regression Coefficients Subject to Interval Constraints in Models with Non-spherical Errors*, Snakhyā **61** series B (1999), 433–442.

[4] J. Lawson and R. Hanson, *Solving Least Squares Problems*, Prentice-Hall, New Jersey 1974.

[5] M. LeBlanc and R. Tibshirani, *Combining Estimates in Regression and Classification*, JASA **91** (1996), 1641–1650.

[6] H. Toutenburg and B. Roeder, *Minimax-Linear and Theil Estimator for Restricted Regression Coefficients*, Statistics **9** (1978), 499–505.

[7] D. Wolpert, *Stacked Generalization*, Neural Networks **5** (1992), 241–259.

Received 2 March 2005