ON THE EFFECTIVENESS OF LASSO DISCRIMINANT ANALYSIS(LADA) IN VARIABLE SELECTION

JUDE CHUKWURA OBI
Department of Statistics, Chukwuemeka Odumegwu Ojukwu University,
Anambra State, Nigeria

ABSTRACT
Lasso, a multiple regression variant, has been shown to be an effective classification tool. The sparse property of the classifier (Lasso Discriminant Analysis) has been explored further to ascertain if the classifier can be an effective variable selection tool. In the ensuring analysis, it is clear that the classifier equally doubles as a valid tool for variable selection.

Keywords: Binary Classification, Fisher's Discriminant Analysis, Regularized Discriminant Analysis, Support Vector Machines.

1. INTRODUCTION

The Lasso Discriminant Analysis (LaDA) is an offshoot of Regression Discriminant Analysis (RDA). The regression discriminant analysis has been shown (Ye, 2007; Duda *et al.*, 2012; Jude, 2017) to be a valid tool for classification based on the multiple regression. The argument in favor of LaDA, which was clearly illustrated by (Jude, 2017), is that since Lasso is a multiple regression variant, and a Regression Discriminant Analysis exist (Jude *et al.*, 2017), then Lasso is an undisputable valid classification tool. The same author clearly provided ample illustration in his work.

The main concern here is that LaDA brings about sparsity, and it refers to the presence of fewer explanatory variables in the predictive model than in the input data. The presence of sparsity is acknowledged when a classifier sets some coefficients to zero, meaning that the explanatory variables connected to the coefficients may not be of any predictive relevance. This way, LaDA can be compared to the work of (Clemmensen *et al.*, 2011), which is also about sparse discriminant analysis.

One question to address here is that since sparsity leads to fewer explanatory variables in the predictive model, which invariably connotes variable selection, can it then be argued that variables discarded by LaDA are not of any predictive relevance? This, however constitutes the main focus of this paper; among other things, to assess the effectiveness of LaDA as a tool for variable selection. In doing this, this research work seeks to find out if variables discarded by LaDA do not actually consist of any predictive relevance when a different classifier is used, or do we conclude eventually that the variables LaDA selects is exclusively useful to the classifier?

2. THEORETICAL BACKGROUND

The theory behind the use of LaDA as a tool for variable selection develops from the theory behind the use of RDA. The major argument of RDA is that the least squares vector of coefficients $\hat{\beta}$ is proportional to γ, where γ is the weight vector of FDA's classification function given that $y \in (+1,-1)$. A further detailed review of the theory is contained in (Jude *et al.*, 2017; Jude, 2017). The authors averred that a classification function due to RDA is

$$g(x) = \hat{\boldsymbol{\beta}}^T \left(\mathbf{x} - \mathbf{x}_{av}\right), \tag{1.1}$$

where $\hat{\beta}$ is a vector of coefficients given that $y \in (+1,-1)$, x_{av} is unweighted average and x is given dataset. We focus on the vector of coefficients $\hat{\beta}$, and argue that since $\hat{\beta}$ is based on the multiple regression, a replacement with Lasso coefficients given that $y \in (+1,-1)$, will make (1) a classification function due to Lasso, hence a LaDA classifier. Furthermore, the authors gave empirical illustrations to show that LaDA competes effectively with notable classifiers like Regularized Fishers Discriminant Analysis (RFDA) and FDA.

3. METHODOLOGY

LaDA classifier will be used as a classification tool on 15 different datasets. The variables used by LaDA in the classification in each case shall be noted. Next, FDA will be used as a classification tool on the same datasets utilizing all the variables on one hand, and variables selected using LaDA on the other hand. If the variables selected using LaDA contain useful predictive information, the error rates arising from the use of full variables and variables selected using LaDA will be the same or nearly the same. The differences in the two error rates will be tested for significance or otherwise using the Wilcoxon rank sum test. If the test is significant, the null hypothesis will be rejected. Here, the null hypothesis (H_0) states that the two error rates are the same. The alternative hypothesis (H_1), on the other hand, argues that the error rates are not the same. If the null hypothesis is not rejected, it stands to reason the two error rates are statistically the same. Hence, the use of variables selected using LaDA, is as good as good as using all the variables in a classification problem.

4. EMPIRICAL INVESTIGATION

In this investigation, two different classifiers namely, LaDA and the Fishers Discriminant Analysis (FDA) are going to be used. In the first place, LaDA will be used to carry out classification. The variables used by LaDA in the classification and the error rate will be noted and thereafter, FDA will be used to carry out classification on the same dataset considering all the variables on one hand, and only the variables selected using LaDA on the other hand. If the variables discarded by LaDA do not contain useful information, we would expect the difference in the error rate for FDA using full variables, and using reduced variables to be zero or approximately zero. Otherwise, we shall compare the error rates of LaDA and FDA given full variables only and if no significant difference exist, a conclusion will be reached that LaDA variables are useful only for itself.

4.1 Datasets

Appendicitis

The data represents 7 medical measures taken over 106 patients on which the class label represents whether the patient has appendicitis (class label +1) or not (class label −1). We have a total of 21 samples in class +1 whereas class −1 consists of 85 samples. The dataset was sourced from KEEL dataset repository.

Australia

The Australia dataset concerns credit card applications, and all attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. It has dimensions 690×14, with two classes representing approved and not approved. The data source and website are (Alcala´ *et al.*, 2010), and http://sci2s.ugr.es/keel/dataset.php?cod=53 respectively.

CoIL 2000

The dataset was used in CoIL 2000 challenge, and contains information on customers of an

insurance company. It is a binary classification dataset, and consists of 85 variables including product usage data, and socio-demographic data. The number of samples involved is 9822, with a total of 9236 in class +1 and 586 in class −1. It was sourced from the UCI Machine Learning Repository.

Handheight

The Handheight dataset is two dimensional, and consists of heights and stretched hand span of 167 male and female college students. Each student decided which of their hands to measure. Class +1 has 89 samples whereas class −1 consists of 78 samples. The source of the data is (Utts and Heckard, 2011).

Heart

This is a real world binary classification heart disease dataset, and the task is to detect the absence (−1) or presence (1) of heart disease. It contains 270 samples and 13 features, with 120 samples in class +1 and 150 samples in class −1. The data was sourced from the UCI Machine Learning Repository.

Heberman

This dataset contains cases from a study that was conducted between 1958 and 1970, at the University of Chicago's Billings Hospital, on the survival of patients who had undergone surgery for breast cancer. The task is to determine if the patient survived 5 years or longer (positive) or if the patient died within 5 year (negative). The sample size is 306 with 3 features, and class +1 has 225 samples whereas class −1 contains 81 samples. The dataset was sourced from the KEEL dataset repository.

Hepatitis

Hepatitis is a real world dataset; it contains a mixture of integer and real valued attributes, with information about patients affected by the hepatitis disease. It consists of 80 samples and 19 features. Class +1 has 67 samples whereas class −1 has 13 samples, and the task is to predict if these patients will die (−1) or survive (1). It was sourced from the UCI Machine Learning Repository.

Hill valley with noise (HVWN)

The hill valley with noise dataset consists of 606 instances, and 100 features for both training and test sets. Noise contamination of the dataset is retained, thereby differentiating it from the hill valley without noise dataset. The data was sourced from the UCI Machine Learning Repository.

Ionosphere Dataset

This is a radar dataset collected by a system in Goose Bay, Labrador. The system consists of a phased array of 16 high-frequency antennas with a total transmitted power of the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. The dataset is contained in mlbench package R.

Magic

This dataset was used to simulate registration of high energy gamma particles, in a ground-based atmospheric Cherenkov gamma telescope, using the imaging technique. The dataset was generated by a Monte Carlo program (Bretz *et al.*, 2009), and the task is to discriminate statistically images generated by primary gammas, from the images of hadronic showers initiated by cosmic rays in the upper atmosphere. It contains 19020 samples and 10 features; the source is the UCI Machine Learning Repository.

Mammographic

This dataset was used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the

patient's age and three BI-RADS attributes together with the ground truth (the severity field, which is the target attribute). The dataset was collected at the Institute of Radiology of the University ErlangenNuremberg between 2003 and 2006. It has dimensions 830×5, and the source is the KEEL dataset repository.

Parkinsons

The Parkinsons dataset is of dimension 195 × 23, and involves a range of biomedical voice measurements of some people with and without Parkinson's disease (PD). It was sourced from the UCI Machine Learning Repository. Documentation on the dataset shows that each column is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals.

Ringnorm

Ringnorm is a 20 dimensional, 2 class classification dataset. Each class is drawn from a multivariate normal distribution, and class 1 has mean 0 and covariance 4 times the identity. Class 2 has mean$\sqrt{}$ $(a,a,\cdots,a)$ and unit covariance ($a = 2/\sqrt{20}$). The number of instances is 7400, and like most simulated datasets, the dataset is useful for testing performances of binary classifiers. The source is the KEEL dataset repository.

Saheart

The Saheart dataset pertains to a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. The saheart data were taken from a larger dataset described in (Rousseauw *et al.*, 1983). The class label indicates whether the person has a coronary heart disease: negative (−1) or positive (+1). The dataset has dimensions 462×9, and is contained in the ElemStatLearn package in R.

Wisconsin Diagnostic Breast Cancer (WDBC) Dataset

WDBC is a real world dataset, and contains 30 features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The number of instances is 569 and the task is to determine if a tumor found is benign or malignant (−1 = malignant, and 1 = benign). It was sourced from the UCI Machine Learning Repository.

5. DATA ANALYSIS/RESULT

The error rates are contained in Table 1. Prior to obtaining them, a two-fold cross validation was used in all the datasets except HVWN dataset. The two-fold cross validation entails using 70% of each dataset as training set and the remaining 30% used as testing set. In the case of HVWN dataset, both training and testing sets were provided by the data source.

With reference to Table 1, negative difference means that the error rate using all the variables are smallerthan the error rates using LaDA selected variables. Where there is positive difference, the opposite applies. In six of the datasets used, the difference in the two error rates are all zeroes, meaning that the performances of the reduced variables gave exactly the same results compared to where all the variable were used. Further, it seems that where differences exist in the error rates, they are insignificant. However, on the basis of the number of datasets used in the study, a non-parametric Wilcoxon test will be used to investigate the significance or otherwise, of the differences obtained. In other words, the Wilcoxon test will help to determine if there is any significant difference in error rates between using all the variables and using only the variables selected using LaDA.

Table 1: Error Rates Of FDA Using Both Full Variables And Variables Selected Using LADA On Different Datasets

S/No	Name of Dataset	Dimensions	No. of LaDA Var	FDA (Full Var Error)	FDA (LaDA Var Error)	Diff
1	Appendicitis	106×7	5	0.1250	0.1250	0
2	Australia	689×14	8	0.1353	0.1353	0
3	Coil2000	9822×85	50	0.2538	0.2633	-0.0095
4	Handheight	167×2	2	0.18	0.18	0
5	Heart	270×13	11	0.1489	0.1489	0
6	Heberman	306×3	2	0.2418	0.2418	0
7	Hepatitis	80×20	3	0.1667	0.1944	-0.0277
8	HVWN	1212×100	30	0.396	0.33	0.066
9	Ionosphere	350×32	20	0.181	0.1619	0.0191
10	Magic	19020×10	9	0.2064	0.207	-0.0006
11	Mammographic	830×5	2	0.2088	0.2209	-0.0121
12	Parkinsons	195×23	20	0.1379	0.1207	0.0172
13	Ringnorin	7400×20	20	0.2177	0.2477	0
14	Saheart	462×9	7	0.3381	0.3453	-0.0072
15	WDBC	569×30	25	0.0234	0.0175	0.0059

.

The Wilcoxon Test

In this test, the hypotheses of interest are:

$H_0 : d = 0$

$H_1 : d \neq 0$,

where d is the difference in error rate between using the full variables and using only variables selected by LaDA in classification. If the null hypothesis is not rejected, for instance, it shows that the full variables and variables selected using LaDA are equally effective. Otherwise either the full variables is better than the variables selected using LaDA or vice versa.

The application of the Wilcoxon test shows that at a p-value of 0.8385, the null hypothesis cannot be rejected. Hence, both full and reduced variables are equally effective in classification.

CONCLUSIONS

In line with the foregoing, LaDA is arguably an effective tool for variable selection. This position is reinforced by the fact that H_0 could not be rejected, meaning that variable selected using LaDA contain useful information for classification purposes. On this account, it is my opinion that in addition to using LaDA as a tool for discriminant analysis, it should also be used as a valid variable selection tool.

REFERENCES

Alcala, J, A Fern´ andez, J Luengo, J Derrac, S Garc´ ´ıa, L Sanchez, and F Herrera (2010). "KEEL´ data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework". In: *Journal of Multiple-Valued Logic and Soft Computing* 17.2-3, pp. 255–287.

Bretz, Thomas, Daniela Dorner, Robert M Wagner, and Peter Sawallisch (2009). "The drive system of the major atmospheric gamma-ray imaging Cherenkov telescope". In: *Astroparticle Physics* 31.2, pp. 92–101.

Clemmensen, Line, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll (2011). "Sparse Discriminant Analysis". In: *Technometrics* 53.4, pp. 406–413.

Duda, Richard O, Peter E Hart, and David G Stork (2012). *Pattern Classification*. John Wiley & Sons.

Jude, C. Obi (2017). "Regression Discriminant Analysis (RDA) Variants". In: *International Journal of Scientific and Research Publications* Volume 7. URL: http://www.ijsrp.org/researchpaper-0817.php?rp=P686716.

Jude, C. Obi, Thwaites Peter, and Kent John (2017). "On the Regression Discriminant Analysis (RDA), and its Identical Relationship to the Fisher's Discriminant Analysis". In: *International Journal of Scientific and Research Publications*. URL: http://www.ijsrp.org/researchpaper-0717.php?rp=P676658.

Rousseauw, J, J Du Plessis, a Benade, P Jordann, J Kotze, P Jooste, and J Ferreira (1983). "Coronary Risk Factor Screening in three Rural Communities". In: *South African Medical Journal* 64.430436, p. 216.

Utts, Jessica and Robert F Heckard (2011). *Mind on Statistics*. Cengage Learning.

Ye, Jieping (2007). "Least Squares Linear Discriminant Analysis". In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 1087–1093.