 ISSN: 978-602-71169-7-9

Analyze of Classification Accaptence Subsidy Food Using Kernel Discriminant

Alan Prahutama[1,a], Moch. Abdul Mukid[1,b]

[1]Statistics Department, Faculty of Sains and Mathematics, Diponegoro University

[a]alanprahutama@undip.ac.id, [b]mamukid@yahoo.com

Abstract. Subsidy food is government program for social protection to poor households. The aims of this program are to effort households from starve and to decrease poverty. Less precisely target of this program has negative impact. So that to successful program, it's important to know accuracy classification of admission subsidy food. The variables classification are number of household members, number of household member in work, average expenditure capita, weighted household, and floor area. Discriminant analysis is a multivariate statistical technique which can be used to classify the new observation into a specific group. Kernel discriminant analysis is a non-parametric method which is flexible because it does not have to concern about assumption from certain distribution and equal variance matrices as in parametric discriminant analysis. The classification using the kernel discriminant analysis with the normal kernel function with optimum bandwidth 0.6 gives accurate classification 75.35%.

Keywords: Discriminant kernel, Bandwidth, Subsidy Food

Introduction

Economic imbalances in Indonesia resulted in increasing poverty. Furthermore Poverty is an issue that can not be solved for developing countries. Poverty reduction has also become one of the main goals *Millenium Development Goals* (MDGs) in 2000-2015 years. In 2013 Indonesia has been implemented several programs in an effort to overcome the problems of poverty which was impacted by fuel subsidy reduction. The programs has been implemented are Acceleration and Expansion program for Social Protection which consist of subsidy food, subsidy for poor student and family expectaion program. In addition, the government has also implemented special programs such as direct subsidy while for people and Infrastructure Development Acceleration and Expansion program. Subsidy Food is a program of social protection for poor households and in an effort to protect household from the threat of malnutrition.

Discriminant analysis is a method to determine the linier combination of features that best separates two or more groups of objects [2]. It seeks to find the projection directions along which the gropus are best separated [12]. Discriminant analysis usefull to classify object into groups based on variables. Discriminant analysis procedure is to establish a discriminant function which is used for classification. Discriminant analysis is very tight with the assumption, such as multivariate normal distribution of data and variance covariance matrix is homogeneous [8]. If the assumption is not met, then the model is not accurate for classifying new observations [14]. Another opinion expressed that althought discriminant analysis is conceptually simple and has been used in many application areas, its normality assumption limits its performance. Furthermore, as a linier method, it does not capture nonlinierity clustered structures [11]. Predictions and classifications with nonparametric techniques is one of the statistical analysis that is quite popular and widely used in various studies. Nonparametric discriminant analysis is a method that is flexible because it does not have to fulfill certain assumptions as to the parametric discriminant analysis. One nonparametric discriminant analysis developed at this time is the kernel discriminant analysis. Some studies have linked the kernel discriminant analysis has been done such as [10] regarding the selection of kernel parameters at the kernel discriminant analysis. [7] Conduct research kernel discriminant analysis using fuzzy membership functions. [9] Use kernel discriminant analysis for fault diagnosis. [5]

Use kernel discriminant embedding in face recognition. [2] Use kernel discriminant for classification data.

Based on the exposure to be observed regarding the accuracy of the classification of subsidy food distribution using kernel discriminant and determine variables that influence of subsidy food distribution. Varibles to be observed regarding the subsidy food distribution such number of household members (X_1), number of household members who work (X_2), Average Expenditure Per Capita (X_3), Weighing Household (X_4) and Floor Area (X_5)

Materials and Methods

Discriminant Kernel

Kernel Discriminant Analysis is discriminant analysis method which extracts nonlinier discriminant features by generalizing discriminant function to a high dimension kernel space [1]. If density function of $f_i(x)$ unknown or in other words, distribution model of X unknown which $f_i(x)$ assumsed that in smooth function, then nonparametric approch should be use for estimation $f_i(x)$ [4]. The principle of kernel discriminant analysis is that, after the data are nonlinierly mapped to high dimension kernel space, then the data can be analyzed by discriminant linier. In this procces, discriminant linier analysis convert linier to non linier space. Given sample data x_{ij} where i groups (biner) dan j observation with $(i=1,2;\ j=1,2,\ldots,n)$. Discriminant kernel function is:

$$f(x) = g(x|h) + \varepsilon \tag{1}$$

where $f(x)$ can be estimated by:

$$g(x \mid h) = \frac{1}{n}\sum_{j=1}^{n} K(x \mid x_j, h) \tag{2}$$

where $K(x \mid x_j, h)$ is kernel density function, and h smoothers parameters (*bandwidth*). To estimate density function in Eq. (1) can be written by [14]:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right) = \frac{1}{n}\sum_{j=1}^{n} K\left(x-x_j\right) \tag{3}$$

Given H is a matrix bandwidth symmetric and definite positive *dxd* and $\int K(x)\,dx = 1$. If the kernel density function by following a multivariate normal distribution for each group

$$K(x) = \frac{1}{(2\pi)^{d/2} h^d \,|\Sigma|^{1/2}} \exp\left(\left(-0.5(x-x_j)^T \Sigma^{-1}(x-x_j)\right)/h^2\right)$$

substituted into the Eq. (3) then obtained form:

$$\hat{f}(x;H) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{(2\pi)^{d/2} h^d \,|\Sigma|^{1/2}} \exp\left(\left(-0.5(x-x_j)^T \Sigma^{-1}(x-x_j)\right)/h^2\right) \tag{4}$$

So that Eq. (4) for *K(x)* normal multivariate distribution with mean 0 and variance covariance H can be written by:

$$\hat{f}(x;H) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{(2\pi)^{d/2} h^d \,|\Sigma|^{1/2}} \exp\left(\left(-0.5(x-x_j)^T H^{-1}(x-x_j)\right)/h^2\right) \tag{5}$$

Determine H is important because it can determine function of $\hat{f}$.

Kernel Function

Kernel function is one approach that can be used to estimate the probability density function of an observation. Kernel function is a nonparametric density function for estimation method that developed from the histogram estimator [6]. Kernel method is a nonparametric approach to classify the data that have nonlinear patterns [13]. Some of kernel function such as:

 ISSN: 978-602-71169-7-9

Table 1. Kernel function density

Kernel	K(x)
Uniform	$K(x)=\begin{cases}\dfrac{\frac{1}{\pi^{d/2}h^2}}{\Gamma\left(\frac{d}{2}+1\right)}\lvert\Sigma\rvert^{d/2} & jika\ (x-x_j)^T H^{-1}(x-x_j)\le h^2\\ 0\ ;lainnya\end{cases}$
Normal	$K(x)=\dfrac{1}{c_{0(t)}}\exp\left(\left(-0.5(x-x_j)^T H^{-1}(x-x_j)\right)/h^2\right)$
Epanechnikov	$K(x)=\begin{cases}c_{1(t)}\left(1-(x-x_j)^T H^{-1}(x-x_j)/h^2\right), if\ (x-x_j)^T H^{-1}(x-x_j)\le h^2\\ 0\ ;others\end{cases}$
Biweight	$K(x)=\begin{cases}c_{1(t)}\left(1-(x-x_j)^T H^{-1}(x-x_j)/h^2\right)^2, if\ (x-x_j)^T H^{-1}(x-x_j)\le h^2\\ 0\ ;others\end{cases}$
Triweight	$K(x)=\begin{cases}c_{1(t)}\left(1-(x-x_j)^T H^{-1}(x-x_j)/h^2\right)^3, if\ (x-x_j)^T H^{-1}(x-x_j)\le h^2\\ 0\ ;others\end{cases}$

where

$$c_{0(t)}=(2\pi)^{d/2}h^d\lvert\Sigma\rvert^{1/2};\ c_{1(t)}=\left(1+\frac{d}{2}\right)c_{0(t)};\ c_{2(t)}=\left(1+\frac{d}{4}\right)c_{1(t)};\ c_{3(t)}=\left(1+\frac{d}{6}\right)c_{2(t)}$$

APER (Apparent Error Rate)

APER is stating the proportion of samples incorrectly classified. It uses confusion matric to calculate the number of classify [8]. Confusion matrix is the form of an observation classification results. Can be shown in Table 2, n_1 number of classification in group π_1 and n_2 number of classification in group π_2 then it can be conduct as well as:

Table 2. Confusion Matric

		Groups prediction	
		π_1	π_2
Groups	π_1	n_{11}	n_{12}
	π_2	n_{21}	n_{22}

n_{11} is number of members in group π_1 that exactly classify into π_1, while n_{22} is number of member in group π_2 that exactly classify into π_2. While n_{12} is umber of member group π_1 that unexactly classify into π_2, then n_{21} is number of member π_2 that unexctly classify into π_1. Therefore, statistic test for APER can be shown:

$$APER=\left(\frac{n_{12}+n_{21}}{n_1+n_2}\right)\times 100\% \qquad (6)$$

Results and Discussion

The data used in this research is subsidy food data Kendal Regency in 2013. Based on the data can be shown:

Table 3. Descriptive of Subsudy Food in Kendal Regency

Status	Total	Precentage
Receiver subsidy food	566	71.60%
Unreceiver subsidy food	225	28.40%
Total	791	100%

Based on Table 3. From the data we know that, 71.6% is receiver subsidy food, then 28.4% unreceiver subsidy food. It show that, almost receiver data as a describe in this research for classification than unreceiver.

Table 4. Statistics Descriptive for Variables

Variable	Minimum	Maximum	Average	Std. deviation
number of household members (X_1)	1	14	3.5209	1.50754
number of household members who work (X_2)	0	6	1.8633	0.97823
Average Expenditure Per Capita (X_3)	149407	3709412.1	660898.2	1351406
Weighing Household (X_4)	2.76	723.32	317.3316	121.397
Floor Area (X_5)	6	728	85.2832	53.60187

Table 4 show that the variables for classifying subsidy food. the average number of household members ranges 3 until 4 person. Whereas in a houshold, the number of who work ranges 1 until 2 person. While Average Expenditure Per Capita is Rp 660. 898,- for a month. Average of Weighing household is 317.3316. Average of floor area is 85.28 m^2.

Table 5. Proportion of Training Data and Testing Data

Proportion	Training	Testing	Total
90/10	712	79	791
80/20	633	158	791

Based on Table 5, it show that allocation of training data and testing tested three times. It aims to look at the level of accuracy.

Tabel 6. Allocation the details of the training data and testing data

Proportion		90/10		80/20	
Groups	Total	Train.	Test.	Train.	Test.
1	566	509	57	453	113
2	225	203	23	180	45
Total		712	79	633	158

Table 8 show that the allocation amount for each training data and testing is in proportion to the amount of data.

Table 7 show that classification with discriminant kernel method using normal kernel function. The optimum bandwidth is 1 with error 0.26685 for allocation 90: 10. While allocation 80: 20, the optimum bandwidth is 0.6 with error 0.24645.

Based on Table 8 with Epachenikov Kernel Function we can describe about the number of bandwidth . It is so different with normal kernel function

 ISSN: 978-602-71169-7-9

Table 7. Classification Use Discriminant Kernel with Normal Kernel Function

90: 10				80:20			
Bandwidth	Error	Bandwidth	Error	Bandwidth	Error	Bandwidth	Error
0.1	0.375	1	0.26685	0.1	0.372828	0.5	0.252765
0.01	0.394663	1.2	0.272472	0.01	0.375987	0.6	0.24645
0.9	0.269663	1.5	0.273876	0.05	0.374408	0.7	0.249605
0.5	0.269663	1.7	0.271067	0.25	0.338073	0.8	0.248025
0.3	0.289326	2	0.275281	0.4	0.259084	0.9	0.255924

The range optimum bandwidth of normal kernel function ranges 0.01 until 1. Whereas Epachenikov kernel function, the optimum bandwidth is started by 3 until 15. It optimum bandwidth is 7 with error 0.27949. The optimum bandiwdth for allocated 80: 20 is 3 with error 0.26066.

Table 8. Classification Use Discriminant Kernel with Epachenikov Kernel Function

90:10:00				1/3/1900 8:20			
Bandwidth	Error	Bandwidth	Error	Bandwidth	Error	Bandwidth	Error
7	**0.27949**	8	0.2809	5	0.27646	7	0.2812
7.2	0.27949	8.5	0.2809	4	0.27014	7.5	0.27962
7.5	0.2809	9	0.2823	**3**	**0.26066**	10	0.27804
7.7	0.27949	9.5	0.28371	6	0.27804	15	0.28436

Based on Table 9, it show that discriminant kernel use triweight kernel function. The optimum bandwidth for allocated 90: 10 is 7 with error 0.27528. And for 80:20, the optimum bandwidth 3 with error 0.25434. The characteristic optimum bandwidth of Epachenikov and triweight kernel function is similar.

Table 9. Classification Use Discriminant Kernel with Triweight Kernel Function

90:10				80:20			
Bandwidth	APER	Bandwidth	APER	Bandwidth	APER	Bandwidth	APER
7	**0.27528**	7.8	0.27809	**3**	**0.25434**	6	0.27804
7.2	0.27669	8	0.27669	4	0.2654	7	0.27646
7.3	0.27669	8.2	0.27949	5	0.27014	8	0.27646
7.5	0.27809	8.5	0.2809	5.5	0.27804	8.5	0.2812

Based on Table 7, Table 8 and Table 9, it can be concluded the optimum bandwidth is 0.6 with error 0.24645, that in allocated training and testing is 80:20 with Normal kernel function.

Then we conduct significance variables which influence of subsidy food. The method that used is binary logistic regression. The results of this method can be shown in Table 9.

Based on Table 10, it show that the variables have influence for subsidy food with significant level 1% are Average Expenditure Per Capita (X_3), Weighing Household (X_4) and Floor Area (X_5).

Table 10. Significant of Parameter use Logistic Biner Regression

Variable	Coefecient	Standard error	Wald Test	df	Sig.	Conclusion
X_1	-,021	,083	,064	1	,800	Not Significant
X_2	-,114	,121	,899	1	,343	Not Significant
X_3	,000	,000	16,389	1	,000	Significant
X_4	,000	,000	3,715	1	,054	Significant
X_5	,026	,004	34,985	1	,000	Significant

Conclusion

Bandwidth ranges for different kernel functions that produce a range of different bandwidth values. The bandwidth optimum is 0.6 with APER 0.24645 for allocate training and testing is 80:20 with normal function kernel. The accuration of subsidy food is 75.35%. The variables have influence for subsidy food with significant level 1% are Average Expenditure Per Capita (X_3), Weighing Household (X_4) and Floor Area (X_5).

References

[1] Baudat, G., Anouar, F., 2000. Generalized discriminant analysis using a kernel approach. *Neural Comput*. 12 (10), 2385–2404.

[2] Bishop C.M., *Pattern Recognition and Machine Learning*, Springer, Singapore, 2007.

[3] Duong, T., 2007, Kernel Density estimation and kernel discriminant analysis for Multivariate Data in R, *Journal of Statistical Software*, Vol.21, Issue. 7

[4] Hair, J.F.Jr., *et. al*., 2010, *Multivariate Analysis Methods and Applications*. Pearson Education. Inc, New Jersey.

[5] Han, P.Y., Jin, A.T.B., and Ann, T.K., 2011, Kernel discriminant Embedding in Face Recognition, *J. Vis. Commun. Image R.*, Vol. 22, pp: 634-642.

[6] Hardle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, U.K.

[7] Heo, G., and Gader, P., 2011, Robust kernel discriminant analysis using fuzzy memberships, *International Jornal of Pattern Recognition Letters*, Vol. 44, pp: 716-713

[8] Johnson, R.A., and Wichern, D.W., 2007, *Applied Multivariate Statistical Analysis 6th Edition*, Pearson International Edition, New Jersey.

[9] Li, J. and Cui, P., 2009, Improved Kernel Fisher Discriminant Analysis for Fault Diagnosis, *Expert System with Applications*, Vol. 36, pp: 1423-1432.

[10] Liu, J., Zhao, F., and Liu, Y., 2013, Learning kernel parameters for kernel Fisher discriminant analysis, *International Journal of Pattern Recognition Letters*, Vol. 34, pp: 1026-1031.

[11] Park, C., Park, H., 2005. Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition. SIAM J. Matrix Anal. Appl. 27 (1), 87–102.

[12] Scott, D.W., 1992, *Multivariate Density estimation: Theory, Practice and Visualitation*, John Wiley and Sons, New York.

[13] Silverman, B.W., 1986, *Density estimation for Statistics and data Analysis*, Chapman and Hall, London.

[14] Worth, A.P., and Cronin, M.T.D., 2003, The Use of discriminant analysis logistic regression and classification tree analysis in development of classification models for human health effects, *Theochem*, Vol. 622, pp:97-111.