Logistic regression in feature selection in data mining

J.Padmavathi[1],

[1] Computer Science, SRM University,

Chennai, Tamil Nadu, 600 026,India

Padmalaya90@gmail.com

Abstract

Predictive data mining in clinical medicine deals with learning models to predict patients' health. The models can be devoted to support clinicians in diagnostic, therapeutic, or monitoring tasks. Data mining methods are usually applied in clinical contexts to analyze retrospective data, thus giving healthcare professionals the opportunity to exploit large amounts of data routinely collected during their day-by-day activity. Moreover, clinicians can nowadays take advantage of data mining techniques to deal with the huge amount of research results obtained by molecular medicine, such as genetic or genomic signatures, which may allow transition from population-based to personalized medicine. This paper aims at throwing light on the oldest feature extraction method, namely, the Logistic Regression (LR). LR is useful for situations in which we want to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. LR was used in heart attack classification in coronary Heart Disease (CHD). LR is used in Binary classification.

Keywords: *logistic regression, feature extraction, Coronary Heart Disease.*

1. Introduction

Predictions may range from the simple stratification of the patients' population on the basis of known risk factors, such as age or lifestyle, to the forecast of the effect that a treatment or drug may have on a single patient. Generally speaking, in a clinical context, predictions may support diagnostic, therapeutic, or monitoring tasks. Diagnosis is related to the classification of patients into disease classes or subclasses on the basis of patients' data[5].

Logistic regression is used to analyze relationships between a dichotomous dependent variable and metric or dichotomous independent variables. Logistic regression combines the independent variables to estimate the probability that a particular event will occur, i.e. a subject will be a member of one of the groups defined by the dichotomous dependent variable. The variate or value produced by logistic regression is a probability value between 0.0 and 1.0. If the probability for group membership in the modeled category is above some cut point (the default is 0.50), the subject is predicted to be a member of the modeled group. If the probability is below the cut point, the subject is predicted to be a member of the other group. For any given case, logistic regression computes the probability that a case with a particular set of values for the independent variable is a member of the modeled category.

$$Yi = e^u/(1 + e^u)$$

Where Yi is the estimated probability that the ith case is in a category and u is the regular linear regression equation:

$$u = A + B_1X_1 + B_2X_2 + \cdots + B_KX_K$$

2. Related work

The data about Coronary heart disease was collected and used in classifying the patients with heart attack was collected from Neomed Hospital. Attribute entries were recorded from Treadmill test, 3D-CCD and from physician examination recorded in the case history. Table 1, shows the attributes used for the analysis. The stepwise analysis used Exangi, Chol. Level, CP, slope (ST-depression or ST-elevation) as significant variables. Direct analysis used all the 17 parameters. SPSS software is used for analysis. Optimization criteria are taken as likelihood criteria. The single predictor is defined as,

$p(x\ ) = P(Y_j = 1\ X = x\ ) \Rightarrow E(Y_jX = x\ )$.

Instead of the probability of heart disease, we consider the odds as a function of age. Odds range from zero to infinity, so the problem fitting a linear model to the upper asymptote can be eliminated. If we go one step further and consider the logarithm of the odds, we now have a dependent variable that ranges from -1 to +1. We try to fit a linear regression model to the log-odds variable.

Our model would now be ,

logit(p(x)) = ($\log\left(\frac{p(x)}{1-p(x)}\right) = \beta 0 + \beta 1$ …).(1)

If we can successfully fit this linear model, then we also have successfully fit a nonlinear model for p(x), since the logit function is invertible, so after taking logit^{-1} of both sides, we obtain p(x) = logit^{-1} ($\beta_0 + \beta_1 x$)….(2)

where, logit^{-1}(w)= $\exp(w)/(1+\exp(w))$

= $1/(1+\exp(-w)$)…..(3)

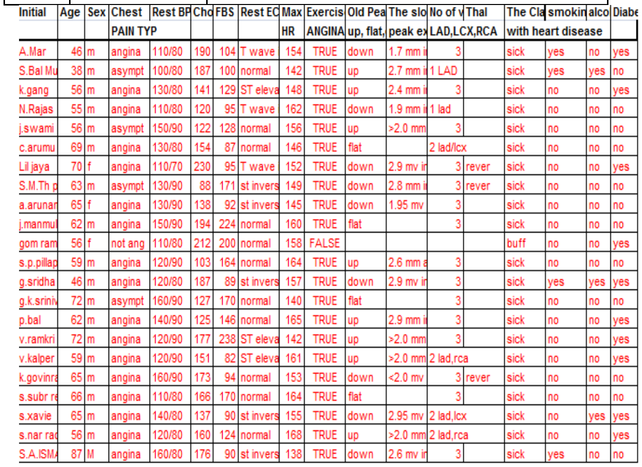
The above system generalizes to more than one predictor, i.e.,p(x)=E(Y | X= x) => logit^{-1}(β'x)….(4).

It turns out that the system we have just described is a special case of what is now termed a generalized linear model. The

Change in probability is not constant (linear) with constant changes in X. This means that the probability of a success (Y = 1) given the predictor variable (X) is a non-linear function, specifically a logistic function.

Table 1: Attributes of Cardiovascular disease dataset

No	***Name***	***Description***
1	Age	Age in years
2	Sex	1 = male, 0 = female
3	Cpain	Chest pain type (1 = typical angina, 2 =atypical angina, 3 = non anginal pain,4 = asymptomatic)
4	RestBP	Resting blood Pressure (in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7	RestECGg	Resting electrocardiographic results(0 = normal, 1 = having ST/ T wave abnormality, 2 = left ventricular hypertrophy)
8	MaxHR	Maximum heart rate
9	Exangi	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1 = up sloping, 2 = flat, 3 =down sloping)
12	Noves.	Number of major vessels colored by fluoroscopy
13	Thal	3 = normal, 6 = fixed defect, 7 = reversible defect
14	class	Class (0 = healthy, 1 = have heart disease)
15	Smoking	(yes/no)
16	Alcohol	(yes/no)
17	Diabetes	(yes/No)



Initial	Age	Sex	Chest	Rest BP	Cho	FBS	Rest EC	Max	Exercis	Old Pea	The slo	No of v	Thal	The Cla	smokin	alco	Diabe
			PAIN TYP					HR	ANGINA	up, flat,	peak ex	LAD,LCX,RCA		with heart disease			
A.Mar	46	m	angina	110/80	190	104	T wave	154	TRUE	down	1.7 mm i	3		sick	yes	no	yes
S.Bal Mu	38	m	asympt	100/80	187	100	normal	142	TRUE	up	2.7 mm i	1 LAD		sick	yes	yes	no
k.gang	56	m	angina	130/80	141	129	ST eleva	148	TRUE	up	2.4 mm i	3		sick	no	no	yes
N.Rajas	55	m	angina	110/80	120	95	T wave	162	TRUE	down	1.9 mm i	1 lad		sick	no	no	no
j.swami	56	m	asympt	150/90	122	128	normal	156	TRUE	up	>2.0 mm	3		sick	no	no	no
c.arumu	69	m	angina	130/80	154	87	normal	146	TRUE	flat		2 lad/lcx		sick	no	no	no
Lil jaya	70	f	angina	110/70	230	95	T wave	152	TRUE	down	2.9 mv ir	3	rever	sick	no	no	yes
S.M.Th p	63	m	asympt	130/90	88	171	st invers	149	TRUE	down	2.8 mm i	3	rever	sick	no	no	no
a.arunar	65	f	angina	130/90	138	92	st invers	145	TRUE	down	1.95 mv	3		sick	no	no	no
j.manmul	62	m	angina	150/90	194	224	normal	160	TRUE	flat		3		sick	no	no	no
gom ram	56	f	not ang	110/80	212	200	normal	158	FALSE					buff	no	no	yes
s.p.pillap	59	m	angina	120/90	103	164	normal	164	TRUE	up	2.6 mm a	3		sick	no	no	no
g.sridha	46	m	angina	120/80	187	89	st invers	157	TRUE	down	2.9 mv ir	3		sick	yes	yes	yes
g.k.sriniv	72	m	asympt	160/90	127	170	normal	140	TRUE	flat		3		sick	no	no	no
p.bal	62	m	angina	140/90	125	146	normal	165	TRUE	up	2.9 mm i	3		sick	no	no	yes
v.ramkri	72	m	angina	120/90	177	238	ST eleva	142	TRUE	up	>2.0 mm	3		sick	no	no	yes
v.kalper	59	m	angina	120/90	151	82	ST eleva	161	TRUE	up	>2.0 mm	2 lad,rca		sick	no	no	yes
k.govinra	65	m	angina	160/90	173	94	normal	153	TRUE	down	<2.0 mv	3	rever	sick	no	no	no
s.subr re	66	m	angina	110/80	166	170	normal	164	TRUE	flat		3		sick	no	no	no
s.xavie	65	m	angina	140/80	137	90	st invers	155	TRUE	down	2.95 mv	2 lad,lcx		sick	no	yes	yes
s.nar rad	56	m	angina	120/80	160	124	normal	168	TRUE	up	>2.0 mm	2 lad,rca		sick	no	no	yes
S.A.ISMA	87	M	angina	160/80	176	90	st invers	138	TRUE	down	2.6 mv ir	3		sick	yes	no	no

Level of measurement requirements

Logistic regression analysis requires that the dependent variable be dichotomous. Logistic regression analysis requires that the independent variables be metric or dichotomous. If an independent variable is nominal level and not dichotomous, the logistic regression procedure in SPSS has an option to dummy code the variable for you. If an independent variable is ordinal, we will attach the usual caution [4].

Assumption

Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables. Because it does not impose these requirements, it is preferred to discriminant analysis when the data does not satisfy these assumptions [6].

Sample size

The minimum number of cases per independent variable is 10, using a guideline provided by Hosmer and Lemeshow, authors of *Applied Logistic Regression*, one of the main resources for Logistic Regression. For preferred case-to-variable ratios, we will use 20 to 1 for simultaneous and hierarchical logistic regression and 50 to 1 for stepwise logistic regression.

Methods for including variables.

There are three methods available for including variables in the regression equation:

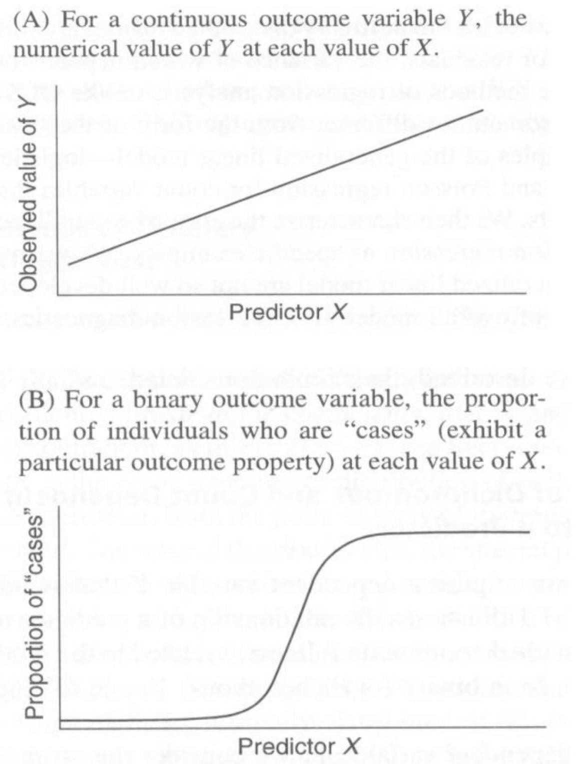
- The simultaneous method in which all independents are included at the same time.
- The hierarchical method in which control variables are entered in the analysis before the predictors whose effects we are primarily concerned with.
- The stepwise method (forward conditional in SPSS) in which variables is selected in the order in which they maximize the statistically significant contribution to the model.

Forward selection consists in choosing the most predictable variable and then checks for a second variable that is added to the first, most improves the model. This process is repeated until either all variables have been selected or no further improvement is made. Backward stepwise feature selection is the reverse process. It starts with all the variables and then removes a variable at each stage which less degrades the model. Forward selection is faster but may miss key variables if they are independent. Backward stepwise selection does not suffer this problem, but is time consuming at the beginning of the process due to evaluation of whole set of variables.

For all methods, the contribution to the model is measured by model. Chi-square is a statistical measure of the fit between the dependent and independent variables, like R^2(regression analysis, the relative amount of variance of dependent variable)

Assessing the Model using Multiple regression

Multiple regression uses the least-squares method to find the coefficients for the independent variables in the regression equation, i.e. it computed coefficients that minimized the residuals for all cases. Logistic regression uses maximum-likelihood estimation to compute the coefficients for the logistic regression equation. This method finds attempts to find coefficients that match the breakdown of cases on the dependent variable. The overall measure of how well the model fits is given by the likelihood value, which is similar to the residual or error sum of squares value for multiple regression. A model that fits the data well will have a small likelihood value. A perfect model would have a likelihood value of zero. Maximum-likelihood estimation is an iterative procedure that successively tries and works to get closer and closer to the correct answer.



(A) For a continuous outcome variable Y, the numerical value of Y at each value of X.

(B) For a binary outcome variable, the proportion of individuals who are "cases" (exhibit a particular outcome property) at each value of X.

In multiple linear regression, the residual sum of squares provides the basis for tests for comparing mean functions. In logistic regression, the residual sum of squares is replaced by the deviance, which is often called G^2. Suppose there are 'k' data groupings based on n_i ; i = 1; : : : ; k binomial observations. The deviance is defined for logistic regression

$$G^2 = 2\sum_{i=1}^{k} y_i \log\left(\frac{Y_i}{Y'_i}\right) + \left(N_i - \frac{y_i}{N_i} - Y'_i\right)$$

The degrees of freedom associated with the analysis is the number of groupings 'n' used in the calculation minus the number of free parameters in β that were estimated. Comparing models in logistic regression is similar to regular linear regression [8].

For two nested models, the difference in deviances is treated as a chi-square with degrees of freedom equal to the difference in the degrees of freedom for the two models.

4.0 Conclusion

Logistic regression is a type of multivariable analysis used with increasing frequency in the health sciences because of its ability to model dichotomous outcomes. Proper use of this powerful and sophisticated modeling technique requires considerable care both in the specification of the form of the model and in the calculation and interpretation of the model's coefficients[7][8]. The coefficients of the predictor variables are interpreted as signifying the relative contribution of their respective variables toward the predicted probability of a positive outcome. The criteria considered in this article can affect the regression coefficients, in different ways and at different stages of the model-building process. Although many parts of the process have been effectively automated, the authority of the final model depends on the attempts by investigators to rule out sources of bias or inaccuracy toward which each of the criteria contributes.

Future Work

The future work is to design a hybrid model, which would help in predicting heart attack. Feature extraction is done with logistic regression and the data is fed into RBF neural network.

References

[1] Vollmer RT. Multivariate statistical analysis for pathologists. Part I, The logistic model. Am J Clin Pathol 1996;105:115–26.

[2] Lemeshow S, Hosmer DW. Logistic regression. In: Armitage P, ColtonT, Eds. Encyclopedia of Biostatistics. New York: J. Wiley, 1998.p. 2316–27.

[3] Glantz SA, Slinker BK. Primer of applied regression and analysis of variance. New York: McGraw-Hill, Inc., 1990.

[4] Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989

[5]Abu-Hanna, A., & de Keizer, N. (2003). Integrating Classification trees with local logistic regression in intensive care prognosis. Artificial Intelligence in Medicine, 29, 5-23.

[6]Khemphila, A.; Boonjing, V., "Comparing performanceof logistic regression, decision trees and neural networks for classifying heart disease patients". Proceedings of International Conference on Computer Information System and Industrial Management Applications 2010, pp. 193 –198.

[7] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[8] Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application,Vol3,2008,pp.366-374.

Appendix

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	177	100.0
	Missing Cases	0	.0
	Total	177	100.0
Unselected Cases		0	.0
Total		177	100.0

Classification Table(a,b)

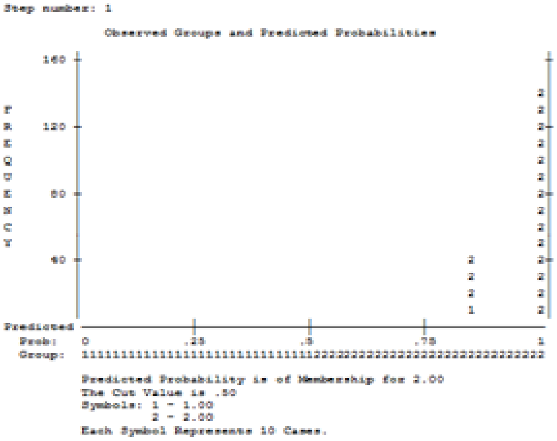
Observed			Predicted		
			class		Percentage Correct
			1.00	2.00	1.00
Step 0	class	1.00	0	7	.0
		2.00	0	170	100.0
	Overall Percentage				96.0

a Constant is included in the model.

b The cut value is .500

Forward-Stepwise Method

With Exercise Angina and Chest pain as variables.

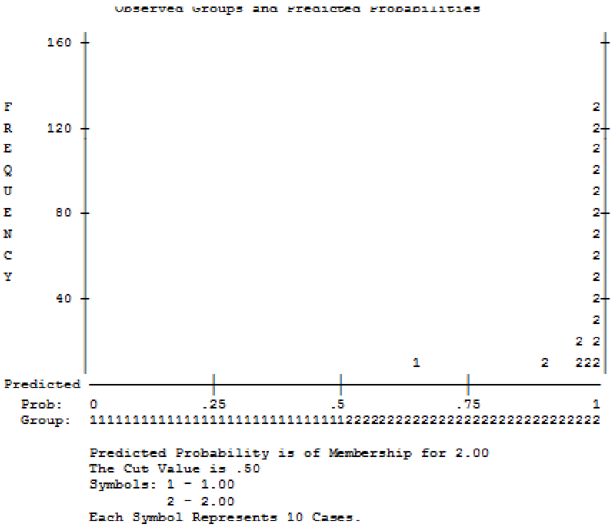


```
Step number: 1

          Observed Groups and Predicted Probabilities

Predicted Probability is of Membership for 2.00
The Cut Value is .50
Symbols: 1 - 1.00
         2 - 2.00
Each Symbol Represents 10 Cases.
```

Classification based on No. Of vessels



```
Observed Groups and Predicted Probabilities

Predicted Probability is of Membership for 2.00
The Cut Value is .50
Symbols: 1 - 1.00
         2 - 2.00
Each Symbol Represents 10 Cases.
```