STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Logistic Regression

Choosing between Logistic Regression and Discriminant Analysis

Daniel Kadyrov
CS559 – Machine Learning
Spring 2020
Presentation #2

Agenda

- Press, S. J., & Wilson, S. (1978). Choosing between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, *73*(364), 699–705. doi: 10.1080/01621459.1978.10480080
  - https://www.math.arizona.edu/~hzhang/math574m/Read/LogitOrLDA.pdf
- Introduction
- Discussion: Logistic Regression Model, Discriminant Function Operators
- Functional Form
- Estimation
- Empirical Applications
  - Example 1
  - Example 2
- Summary and Conclusions

Introduction

Logistic Regression Model

- Two problems are considered:
    1. Relating qualitative dependent variable to at least one independent, not necessarily quantitative, independent variables.
    2. Classification/Discrimination of an object with given characteristics needs to be assigned

Logistic Regression Model

Logistic Regression Model

$$p(\mathbf{x}) \equiv \Pr\{E|\mathbf{x}\} = 1/[1 + \exp\{-\alpha - \beta'\mathbf{x}\}]$$

- When dependent and independent variables are related by logistic distribution
- α and β are the unknown parameters
- Object Classification: *E* can represent a first population event for the object and *x* can represent a profile vector of classification attributes.

Normal Discrimination/Classification:

- Assume that the populations are multivariate normal with equal covariance matrices.
- Data Estimates: θ_1, θ_2 mean vectors of the populations, Σ covariance matrices
- Context Assessment: q_1, q_2 prior classification probabilities

$$(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} + (\tfrac{1}{2})(\boldsymbol{\theta}_2 + \boldsymbol{\theta}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \geq \log(q_2/q_1),$$

Discriminant Function Operators

- Truett, Cornfield, and Kannel 1967 have shown that discriminant function estimators have been used in logistic regression but Halperin, Blackwelder, and Verter 1971, and D'Agostino, 1978 found that they were inferior.
- Reverse Taylor Series Approximators and Conditional Estimators (Nerlove and Press 1973).

$$F(x) = 1/[1 + e^{-(a+bx)}], \quad b \neq 0, \quad -\infty < x < \infty .$$

$$F(x) = \left\{\frac{1}{1 + e^{-(a+b\bar{x})}} - \frac{b\bar{x}e^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2}\right\} + \left\{\frac{be^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2}\right\} x + R(x)$$

$$b = B/[(A + B\bar{x})(1 - A - B\bar{x})]$$

$$a = -b\bar{x} - \log\left(\frac{1}{A + B\bar{x}} - 1\right)$$

$$A = \frac{1}{1 + e^{-(a+b\bar{x})}} - B\bar{x}, \quad B = \frac{be^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2}.$$

Functional Form

- Assume explanatory variables are multivariate, normally distrusted with equal covariance matrices.
  - Variables are independent and dichotomously zero-or-one (two parts, binary)
  - Some variables are multi-variate normal and some dichotomous

- Logistic model for discriminant analysis (rather than a linear discriminant function) is relatively robust; i.e., many types of underlying assumptions lead to the same logistic formulation
- Alternative to contingency table analysis (Gordon (1974))
  - Major role in biological and medical applications where cross-classified tables with large numbers of cells (and usually too few observations per cell) are typically replaced by a logistic or log-linear relationship among the variables, thus obviating (removing) the need for the table

Estimation

- Halperin, Blackwelder, and Verter (1971) used an IBM-360-50 and -65 and found that "the times required for compilation and execution of the programs were higher for the maximum likelihood method than for the discriminant function method by factors ranging approximately from 1.3 to 2”
- Efron (1975) has shown that logistic regression estimators are between one-half and two-thirds as efficient as discriminant function estimators when the data are multivariate normal with equal covariance matrices.
- If the data is strictly normal with equal covariance matrices, linear discriminant function estimators are more economical to calculate and are more efficient than logistic regression MLEs (maximum likelihood estimation)

IBM System/360 Model 50 - Wikipedia

Estimation

Against Use of Discriminant Function Estimators

1. When the explanatory variables don't follow a multivariate normal distribution with equal covariance matrices for each state of the dependent variables, discriminant function estimators of the slope coefficients in the logistic regression will not be consistent
   - No guarantee for fit on large datasets
2. Discriminant function estimation can give mis-leading results regarding significance of the logistic regression coefficients when the normality condition is violated.
3. Halperia, Blackwelder, and Verter (1971) "the maximum likelihood method usually gives slightly better fits to the model, as evaluated from observed and expected numbers of cases per decile of risk." "there is a theoretical basis for the possibility that the discriminant function will give a very poor fit, even if the [logistic regression] model holds."
4. Use of discriminant function estimators tends to mask the troublesome cases by not providing danger signals.
5. The logistic regression model is well-known to have enough statistics associated with it. The MLEs are functions of the sufficient statistics, while the discriminant function estimators are not. smaller mean squared error achieved using estimators based on sufficient statistics (when they exist, as they do here) than by using estimators not based upon sufficient statistics.
   - Rao-Blackwell theorem
6. Maximum likelihood estimation of the logistic regression model forces the expected number of cases to equal the observed number of cases
7. There is some evidence that use of discriminant function estimators may tend to generate substantial bias in some applications.

Empirical Applications

Example 1

- Data collected for certain breast cancer patients initially treated at the British Columbia Cancer Institute between 1955 and 1963
- The variables for the study were mixed, continuous, and discrete. Many of the variables were binary (or turned into binary)
- 173 of the female breast cancer patients were randomly divided into two groups.
  - 115 patients was used as the training set for the classification procedures.
  - 58 patients was used to cross-validate the classification functions.
- The binary grouping variable was defined to be 0 if the lymph nodes were not involved with metastatic carcinoma, and 1 if the nodes were involved.
- The independent variables were number of births, a history of hysterectomy (binary), a history of benign breast disease during lactation (binary), presence of nipple changes as the first disease symptom (binary), and duration of symptoms in months.
- The discriminant analysis was performed using the computer program BMD Stepwise Discriminant Analysis. The logistic regression was performed with the program listed in Nerlove and Press (1973)
- Computations were done on an IBM 370-168
- Computing time for logistic regression was found to be 1.38 times longer than that for discriminant analysis, but this may primarily reflect the computational algorithms that were used.

Empirical Applications

Example 1

1. Summary of Classifications of Breast Cancer Patients by Logistic Regression and Discriminant Function Methods

Case	Actual group	Discriminant classification			Logistic regression classification		
		0	1	Classification rate (%)	0	1	Classification rate (%)
Training set	0	71	5		65	11	
	1	33	6	67	22	17	71
	Total	104	11		87	28	
Validation set	0	31	0		25	6	
	1	24	3	59	16	11	62
	Total	55	3		41	17	

$$U(\mathbf{X}) = .362 - .251X_1 - 1.245X_2 + 1.104X_3 - .036X_4 + 2.114X_5$$

Discriminant

Logistic Regression

$$Y(\mathbf{X}) = .058 - .233X_1 - 1.096X_2 + .713X_3 - .028X_4 + .995X_5$$

- X1is number of births, X2 is hysterectomy, X3 is benign breast disease, X4 is duration of symptoms, and X5 is nipple change symptom.

Empirical Applications

Example 2

- Population change and demographics census data were collected for the 50 states of the U.S.
- Percent change between 1960 and 1970 was coded into binary based on if the population increased or decreased from median.
  - Dependent variable for analysis
- Independent variables (attributes) were per capita income, birthrate, death rate, urbanization, and coastline presence
- The 50 states were randomly assigned to five groups of ten states each.
  - Estimation procedures were performed on 40 states and then validated on the remaining ten states
  - This was done five times with a different group as the validation set

- Computation time was longer for logistic regression, but it provided better discrimination for both the training set and the validation set
- Discriminant analysis misclassified four cases that logistic regression classified correctly.
- Logistic regression misclassified two cases that discriminant analysis had correct.

Empirical Applications

Example 2

3. Raw Data for Example 2

State	Population change	Income	Births	Coast	Urban	Deaths
			a. Set I			
Arkansas	0	2.878	1.8	0	0	1.1
Colorado	1	3.855	1.9	0	1	.8
Delaware	1	4.524	1.9	1	1	.9
Georgia	1	3.354	2.1	1	0	.9
Idaho	0	3.290	1.9	0	0	.8
Iowa	0	3.751	1.7	0	0	1.0
Mississippi	0	2.626	2.2	1	0	1.0
New Jersey	1	4.701	1.6	1	1	.9
Vermont	1	3.468	1.8	0	0	1.0
Washington	1	4.053	1.8	1	1	.9
			b. Set II			
Kentucky	0	3.112	1.9	0	0	1.0
Louisiana	1	3.090	2.7	1	0	1.3
Minnesota	1	3.859	1.8	0	0	.9
New Hampshire	1	3.737	1.7	1	0	1.0
North Dakota	0	3.086	1.9	0	0	.9
Ohio	0	4.020	1.9	0	1	1.0
Oklahoma	0	3.387	1.7	0	0	1.0
Rhode Island	0	3.959	1.7	1	1	1.0
South Carolina	0	2.990	2.0	1	0	.9
West Virginia	0	3.061	1.7	0	0	1.2
			c. Set III			
Connecticut	1	4.917	1.6	1	1	.8
Maine	0	3.302	1.8	1	0	1.1
Maryland	1	4.309	1.5	1	1	.8
Massachusetts	0	4.340	1.7	1	1	1.0
Michigan	1	4.180	1.9	0	1	.9
Missouri	0	3.781	1.8	0	1	1.1
Oregon	1	3.719	1.7	1	0	.9
Pennsylvania	0	3.971	1.6	1	1	1.1
Texas	1	3.606	2.0	1	1	.8
Utah	1	3.227	2.6	0	1	.7
			d. Set IV			
Alabama	0	2.948	2.0	1	0	1.0
Alaska	1	4.644	2.5	1	0	1.0
Arizona	1	3.665	2.1	0	1	.9
California	1	4.493	1.8	1	1	.8
Florida	1	3.738	1.7	1	1	1.1
Nevada	1	4.563	1.8	0	1	.8
New York	0	4.712	1.7	1	1	1.0
South Dakota	0	3.123	1.7	0	0	2.4
Wisconsin	1	3.812	1.7	0	0	.9
Wyoming	0	3.815	1.9	0	0	.9
			e. Set V			
Hawaii	1	4.623	2.2	1	1	.5
Illinois	0	4.507	1.8	0	1	1.0
Indiana	1	3.772	1.9	0	0	.9
Kansas	0	3.853	1.6	0	0	1.0
Montana	0	3.500	1.8	0	0	.9
Nebraska	0	3.789	1.8	0	0	1.1
New Mexico	0	3.077	2.2	0	0	.7
North Carolina	1	3.252	1.9	1	0	.9
Tennessee	0	3.119	1.9	0	0	1.0
Virginia	1	3.712	1.8	1	0	.8

4. Coefficients for Classification Equations of Example 2

Case	Classification Method	Constant	Income	Births	Coast	Urban	Deaths
1	Discriminant analysis	−10.500	+1.610	+3.060	+1.118	−0.360	−1.830
	Logistic regression	−6.238	+1.388	+2.484	+0.874	−0.579	−4.046
2	Discriminant analysis	−7.000	+1.110	+2.240	+0.972	+0.710	−2.240
	Logistic regression	+1.918	+0.580	+0.560	+0.706	+0.249	−5.910
3	Discriminant analysis	−10.700	+1.960	+2.100	+1.482	−0.250	−1.020
	Logistic regression	−6.655	+1.399	+1.894	+0.841	−0.436	−2.428
4	Discriminant analysis	−17.400	+3.550	+5.100	+1.966	−1.890	−5.800
	Logistic regression	−15.162	+3.432	+4.378	+1.391	−1.900	−6.037
5	Discriminant analysis	−13.600	+2.300	+3.610	+0.284	−0.013	−1.870
	Logistic regression	−6.854	+1.542	+2.728	+0.437	−0.452	−4.120
Mean of 5 cases	Discriminant analysis	−11.840	+2.106	+3.222	+1.164	−0.361	−2.552
	Logistic regression	−6.598	+1.668	+2.409	+0.850	−0.634	−4.508

Summary and Conclusion

- Logistic regression with maximum likelihood estimation vs. linear discriminant analysis.
  - Classification Problem
  - Relating qualitative to explanatory variables.
- Logistic regression with MLE outperformed classical linear discrimination in both problems but not by a large amount.

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu