Statistical Classification for Diagnosis of Cirrhosis Patients

Natchalee Srimaneekarn
Wei Liu

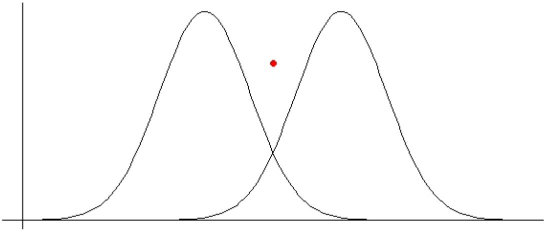
Mathematical Sciences,
University of Southampton, UK

10 December 2015

Table of contents

Research Problem

To classify a patient in term of whether he or she has a disease, based on some diagnostic measurements, is an important problem of statistical classification.



Research Aim

The aim of our research is to study suitable methods for classifying a new patient. Four classification methods for classification into two classes have been studied. They are

- Logistic regression
- Classification tree
- Bayes decision theory
- The new confidence set method.

Data

Two data sets which each observation was classified were used in each methods.

- The Fisher's Iris data set
  - Two (three) classes with four measurements
  - Leave one out method for evaluation
- A data set for classifying patients as normal or having cirrhosis
  - Two classes with 14 measurements on blood samples
  - Data was divided in to two group for construction and evaluation
  - TME and SEN for evaluation criteria

Evaluation Criteria

- Total misclassification error (TME)

$$\text{TME} = \frac{\text{number of incorrect classified}}{N_0 + N_1}$$

- Sensitivity or true positive rate (SEN)

$$\text{SEN} = \frac{\text{number of correct classified in disease group}}{N_1}$$

where N_0 is the total cases in normal group 0,
and N_1 is the total cases in disease group 1

Logistic Regression

Logistic Regression is a classification method using conditional probability for predicting a new case. The dependent variable is the probability that the case belongs to a particular group.

$$log \frac{P(\theta = 1|\boldsymbol{x})}{1 - P(Y = gr.1|\boldsymbol{x})} = \boldsymbol{\beta x},$$

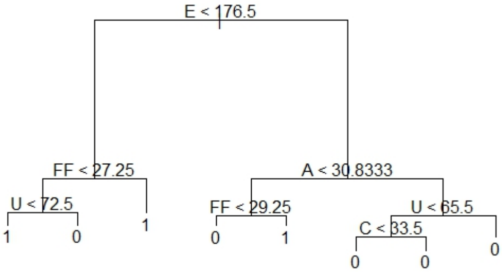
$$P(\theta = 1|\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta x}}}{1 + e^{\boldsymbol{\beta x}}},$$

where $\boldsymbol{\beta x}$ is a linear combination of the predictors.

Then the cut point is applied for classifying a new case.

Classification Tree

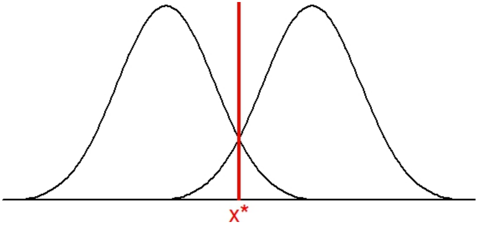
Binary recursive partitioning



Bayes Decision Theory

Classify a case using the highest posterior probability

$$P(\theta = i|\mathbf{x}) > P(\theta = j|\mathbf{x}), \quad i,j = 0,1 \ , \quad i \neq j$$



Decision boundary

$$P(\theta = 1|\mathbf{x}^*) = P(\theta = 0|\mathbf{x}^*)$$

The New Confidence Set Method

Theorem (Lehmann, 1986):

Let data $X \sim f(x;\theta)$ where $\theta \in \Theta$. For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be an acceptance set of size α for testing $H_0 : \theta = \theta_0$, that is,

$$P_{\theta_0}\{X \in A(\theta_0)\} \geq 1 - \alpha.$$

For each X, we can construct $C(X)$ as

$$C(X) = \{\theta_0 : X \in A(\theta_0)\} \subseteq \Theta.$$

Then $C(X)$ is confidence set for θ of confidence level $1 - \alpha$.
That is

$$P\{\theta \in C(x)\} \geq 1 - \alpha.$$

The New Confidence Set Method: one measurement

Using the theorm, the confidence set for the true class can be constructed for a new case $X \sim N(\mu, \sigma^2)$ by inverting the acceptance sets.

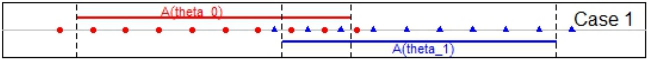
Given $A(\theta_0)$ be an acceptance set for $H_0 : \theta = 0$,
and $A(\theta_1)$ be an acceptance set for $H_0 : \theta = 1$.

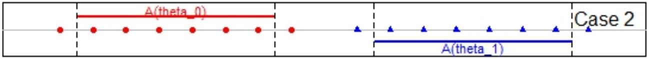
The confidence set is

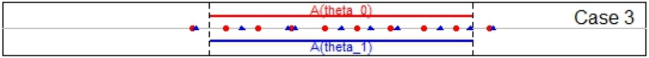
$$C(X) = \{\theta = 0 \text{ or } 1 : X \in A(\theta)\}$$

The New Confidence Set Method: one measurement

$$C(X) = \{\theta = 0 \text{ or } 1 : X \in A(\theta)\}, \ C(X) \in \{\{0\}, \{1\}, \{0, 1\}, \phi\}$$







The New Confidence Set Method: multiple measurements

We have $\boldsymbol{X} \sim \boldsymbol{N_p}(\boldsymbol{\mu}, \Sigma)$. Given

$$A(\theta_0) = \left\{ \boldsymbol{X} : \frac{n_0 - p}{p(1 + 1/n_0)} (\boldsymbol{X} - \bar{\boldsymbol{x}}_{\boldsymbol{0}})' A_0^{-1} (\boldsymbol{X} - \bar{\boldsymbol{x}}_{\boldsymbol{0}}) < f_{p, n_0 - p}^{1-\alpha} \right\},$$

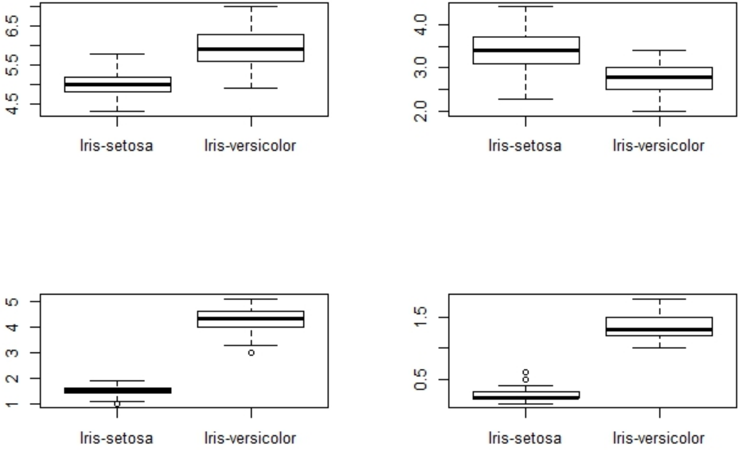
and

$$A(\theta_1) = \left\{ \boldsymbol{X} : \frac{n_1 - p}{p(1 + 1/n_1)} (\boldsymbol{X} - \bar{\boldsymbol{x}}_{\boldsymbol{1}})' A_1^{-1} (\boldsymbol{X} - \bar{\boldsymbol{x}}_{\boldsymbol{1}}) < f_{p, n_1 - p}^{1-\alpha} \right\}.$$

The confidence set is $C(\boldsymbol{X}) = \{\theta : \boldsymbol{X} \in A(\theta)\}$.

The possible confidence sets are $\{0\}$, $\{1\}$, $\{0, 1\}$ and ϕ.

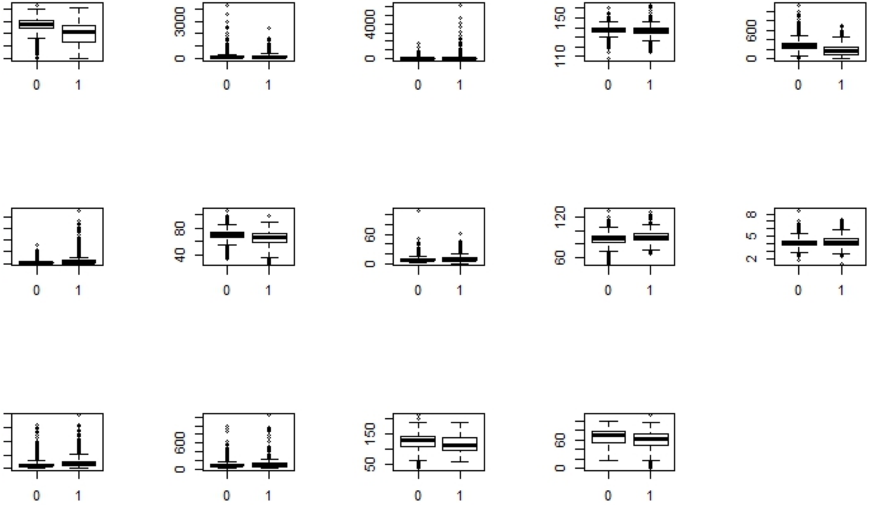
Iris Data


4.5
5.5
6.5
Iris-setosa
Iris-versicolor
2.0
3.0
4.0
Iris-setosa
Iris-versicolor
1
2
3
4
5
Iris-setosa
Iris-versicolor
0.5
1.5
Iris-setosa
Iris-versicolor

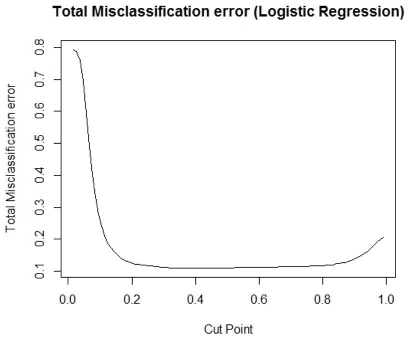
Iris Data: The New Confidence Set Method

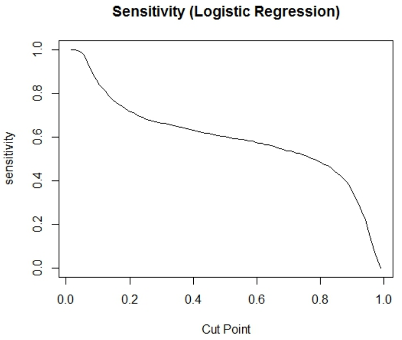
		Predicted Class		
		$\theta = 0$	$\theta = 1$	ϕ
Observed Class	$\theta = 0$	47	0	3
	$\theta = 1$	0	48	2

Cirrhosis Data



Cirrhosis Data: Logistic Regression

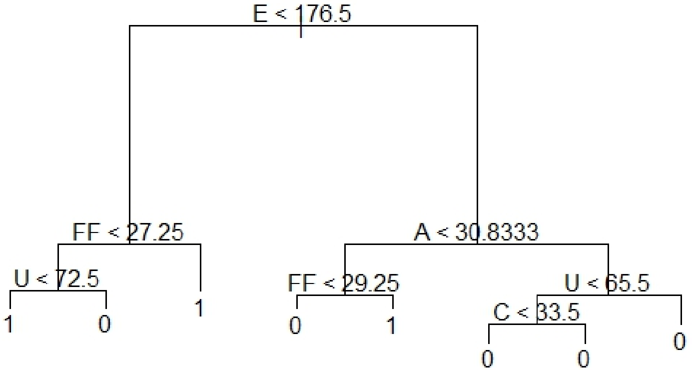




Cirrhosis Data: Logistic Regression

TME	Cut Point	TME	SEN
min	0.4005	0.1095	0.6329
<5% increased	0.2824	0.1141	0.6702
<10% increased	0.2332	0.1200	0.6951

Cirrhosis Data: Classification Tree



Cirrhosis Data: Bayes Decision Theory

		Predicted Class	
		$\theta = 0$	$\theta = 1$
Observed Class	$\theta = 0$	5813	364
	$\theta = 1$	738	869

Cirrhosis Data: The New Confidence Set Method

		Predicted Class			
		$\theta = 0$	$\theta = 1$	ϕ	$\theta \in \{0,1\}$
Observed Class	$\theta = 0$	137	162	123	5755
	$\theta = 1$	3	308	56	1240

Conclusion

Methods	TME	SEN
Logistic Regression	0.1141	0.6702
Classification Tree	0.1345	0.5135
Bayes Decision Theory	0.1416	0.5408
The New Confidence Set Method	0.0442	0.1917

- Logistic regression is the best method for this data set.
- The new confidence set method give the lowest and controllable TME.

References

- Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis.* 3rd ed. Wiley-Blackwel, New York.
- Crawley, M.J. (2012). *The R Book.* 2nd ed. John Wiley and Sons Ltd, West Sussex.
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* Vol.7, 179-188.
- Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). *Applied Logistic Regresion.* 3rd ed. John Wiley and Sons Inc., New Jersey.
- Lehmann, E.L.(1986). *Testing Statistical Hypotheses.* 2nd ed., Duxbury Press, California.
- Martinez, W.L. and Martinez, A.R. (2008). *Computational Statistics Handbook with MATLAB.* 2nd ed. Chapman and Hall/ CRC, Boca Raton.

Thank You for Your Attention

Any Questons?