Issues using Logistic Regression for Highly Imbalanced data

Yazhe Li, Niall Adams, Tony Bellotti

Imperial College London

y.li16@imperial.ac.uk

Credit Scoring and Credit Control conference, Aug 2017

Acknowledgments

Yazhe Li is a Ph.D student from Department of Mathematics, Imperial College London.

This work is supervised by Dr Tony Bellotti (Imperial College London) and Professor Niall Adams (Imperial College London).

Binary Classification Problems

- Credit risk modelling originated in calculating the probability that a debtor will default. Two-class (default, non-default) classification problems are common in the credit risk industry.
- High imbalance (one class is rare) is a common problem in the credit risk industry.
- For example, mortgage default rate could be as low as 0.5% in some data sets.

Logistic Regression

Logistic regression:

- is designed for modelling the posterior probabilities of each class.
- is robust and computational fast.
- has a strong theoretical underpinning, thus making it the most commonly used method for building credit risk model [3].

Introduction

In this talk, we would like to address **highly imbalanced behavior** characteristics of logistic regression and its extensions to penalized logistic regression.

These problems are becoming more pronounced because of increased available data for credit scoring models.

We illustrate these characteristics and their consequences using simulated and real data sets.

Notation

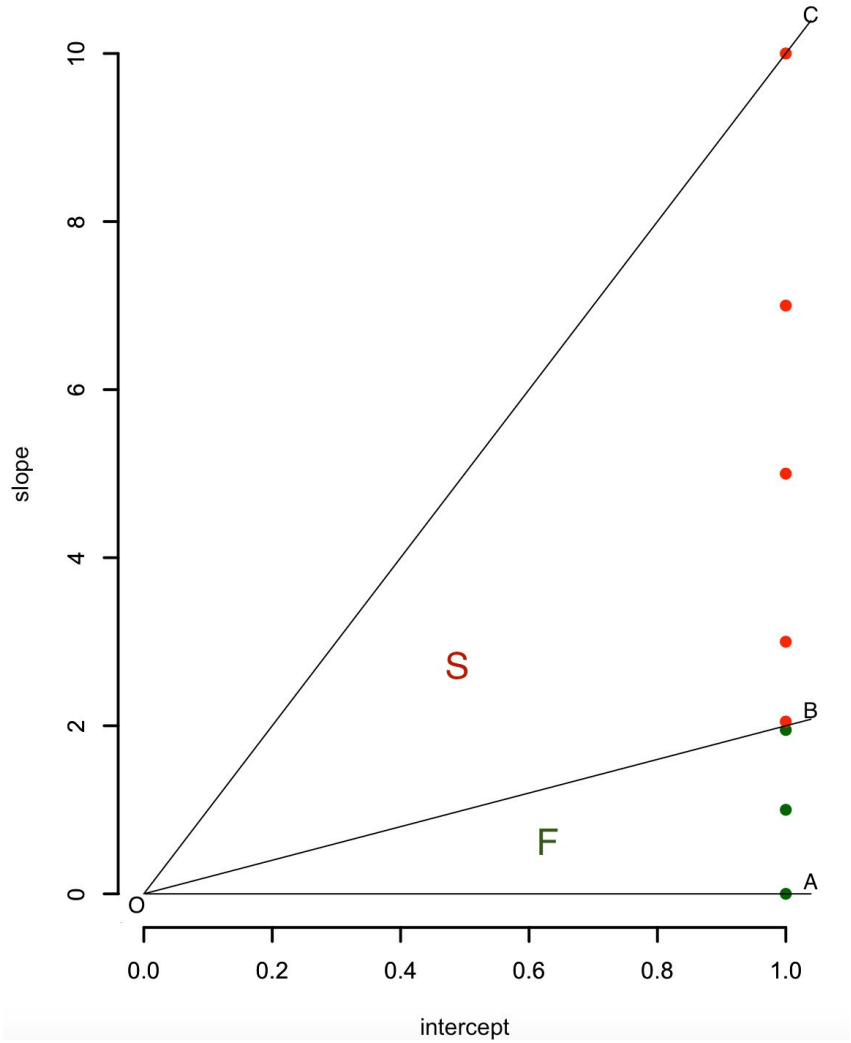
- p-dimensional vector x when $y = 1$ (default) are $x_{11}, \cdots, x_{1n}$.
- p-dimensional vector x when $y = 0$ (non-default) are $x_{01}, \cdots, x_{0N}$.
- Thus, we have $n + N$ p-dimensional observations.
- S, F are the two relative interiors of the convex cones (sub vector space spanned by different class vectors)

$$S = \left\{ \sum_{i=1}^{n} k_i x_{1i} | k_i > 0 \right\} \text{ and } F = \left\{ \sum_{i=1}^{N} k_i x_{0i} | k_i > 0 \right\}. \quad (1)$$

Notation

Example from Silvapulle [2]:

Intercept	Score	Default status
1	0	0
1	1	0
1	2	0
1	2	1
1	3	1
1	5	1
1	7	1
1	10	1



Existence of MLE

Silvapulle has investigated the existence of maximum likelihood estimation (MLE) for logistic regression [2].

Theorem

For data as described above, assume that the $n + N$ by $p + 1$ data matrix (include constant vector 1 to accommodate the intercept) has rank $p + 1$. If $S \cap F \neq \emptyset$ ($\emptyset$ is empty set), then a unique finite logistic regression MLE exists. If however $S \cap F = \emptyset$ then no MLE exists.

For practical purposes, we could understand the theorem as, if there is **no overlap** between values of $\boldsymbol{x}$ **for which** $\boldsymbol{y_i = 0}$ and those for which $\boldsymbol{y_i = 1}$, then the **MLE** for logistic regression does **not exist**.

Infinite Imbalance

- Credit data can be **highly imbalanced** (i.e. one class is extremely rare). We introduce Owen's [1] results about the limited behavior of logistic regression in infinitely imbalanced data set.
- Let Y be the minority class (default) when $y = 1$ and Y be the majority class (non-default) when $y = 0$, thus $n \ll N$.
- We also suppose the conditional distribution of X given $Y = 0$ (majority class) could be estimated as $\boldsymbol{F_0(x)}$.

Infinite Imbalance

Then the main result is (β_0 denotes the intercept, β denotes the slope vectors)

Theorem

(Theorem 8 in [1]) Let $n \geq 1$, and $x_{11}, \cdots, x_{1n} \in R^p$ are fixed. Suppose that F_0 satisfies some tail and overlap conditions. Then the maximizer (β_0, β) of the likelihood function of logistic regression satisfies

$$\lim_{N\to\infty} \frac{\int e^{x^T\beta} x dF_0(x)}{\int e^{x^T\beta} dF_0(x)} = \overline{x}, \quad \text{where } \overline{x} = \frac{\sum_{i=1}^{n}(x_{11}+, \cdots, +x_{1n})}{n}. \tag{2}$$

This theorem could be understood as: when $N \to \infty$ (means going to limit), logistic regression only depends on the minority class data $\{x_{11}, \cdots, x_{1n}\}$ via the minority class mean vector $\overline{x}$.

Infinite Imbalance

Some comments:

- We could replace $\{x_{11}, \cdots, x_{1n}\}$ by one single vector at the mean vector of the minority class and still get the same coefficient estimates of the slope vector β in the limit $N \to \infty$.
- We conduct several simulations which show that β approaches to the limit very fast, when $\frac{N}{n} > 10000$.

Infinite Imbalance

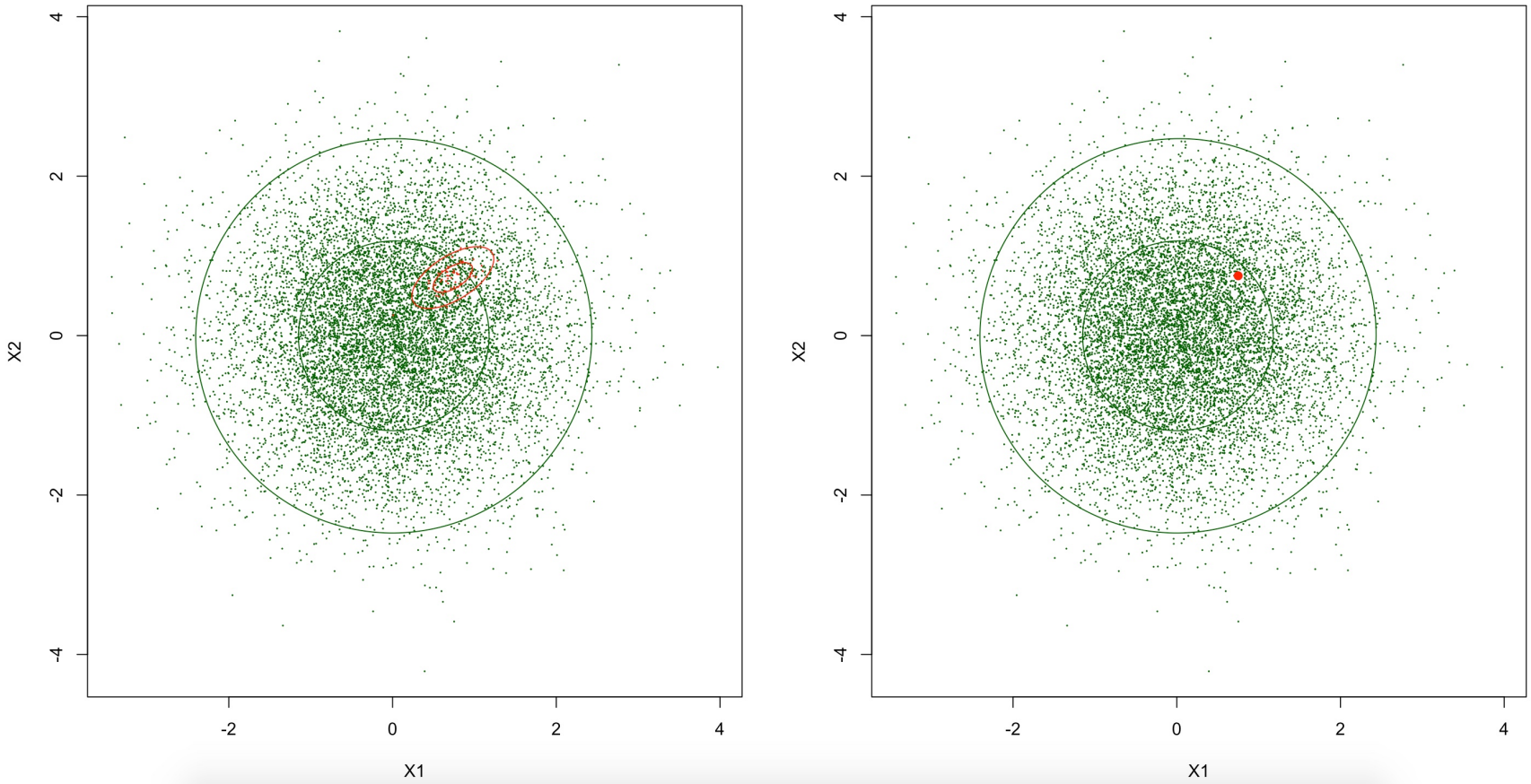


Figure: Illustration of infinitely imbalanced logistic regression.

Results about Penalized Logistic Regression

Penalized logistic regression adds a complexity penalty to log-likelihood function:

$$\textit{Objective Function} = L(\beta; x) - \lambda\left[(1-\alpha)\frac{1}{2}|\beta|_2^2 + \alpha|\beta|_1\right], \quad \text{where } \lambda > 0 \text{ and } 0 \leq \alpha \leq 1. \tag{3}$$

- $L(\beta; x)$ is the log-likelihood function for logistic regression.
- designed for coefficient shrinkage and variable selection.
- alternative model to logistic regression.

We would like to extend the results of existence of MLE and the limited behavior to the infinitely imbalanced penalized logistic regression.

Results about Penalized Logistic Regression

Existence of maximum likelihood estimation in penalized logistic regression:

Theorem

For data as described previously, assume that the $n+N$ by $p+1$ data matrix (include constant vector 1 to accommodate the intercept) has rank $p+1$. If $S \cap F \neq \emptyset$ ($\emptyset$ is empty set), then a unique finite ridge or lasso penalized logistic regression MLE exists.

Thus, the necessity of the condition $S \cap F \neq \emptyset$ is not satisfied, however the sufficiency could be proved.

In fact, if $S \cap F = \emptyset$, when penalty parameter λ **is bigger than a certain value**, the penalized MLE will exist for penalized logistic regression, which means that an appropriate penalty can force penalized MLE existing in linearly separated data sets.

Results about Penalized Logistic Regression

Infinitely Imbalanced Penalized Logistic Regression:
By conducting several simulation processes in penalized logistic regression, we can obtain the following shrinkage law table when $N \to \infty$ (β is slope vector, β_0 is intercept):

Table: Infinitely imbalanced logistic regression shrinkage law.

Fixture	Logistic Regression	Ridge	Lasso
β_0	$-\infty$	$-\infty$	$-\infty$
$N \times e^{\beta_0}$	certain value, k_1	n	n
β	certain value, k_2	0	0

Results about Penalized Logistic Regression

We also observed:

- With penalty parameter λ increasing, the shrinkage speed will increase in penalized logistic regression.
- Since $N \times e^{\beta_0} \to n$ and $\beta \to 0$ in penalized logistic regression when $N \to \infty$ (**no matter what data you use**), which means with N increasing, data will not do its work in the penalized logistic regression model. The outputs probability for minority and majority class will simply approach to their frequency respectively:

$$\Pr(\text{default}|X = x) \xrightarrow{N\to\infty} \frac{n}{n+N},$$
$$\Pr(\text{non-default}|X = x) \xrightarrow{N\to\infty} \frac{N}{n+N}.$$

Consequences of the Theorems

The theorems in the previous slides show challenge and relevant opportunities of deploying logistic regression in the financial industry.

We would like to emphasize **the power of clustering** in logistic regression.

Consequences of the Theorems - Clustering

In highly imbalanced logistic regression, we can get the same β coefficient estimates when we replace all the minority class vectors by their mean vector. The **information loss** in minority class is serious here.

A solution is to cluster minority class data into several well separated clusters and train one logistic regression model per cluster.

By utilizing each cluster's mean vector rather than only minority class mean vector, we reduce the information loss.

Consequences of the Theorems - Clustering

We design a bivariate normal distribution example.

- **Majority class:** (10000 samples) $Y = 0$, $X \sim N(\mu_1, \Sigma_1)$, $\mu_1 = (0, 0)$, the correlation coefficient $\rho_1 = 0$ and the standard deviation in both dimensions are 1.
- **Minority class:** (100 samples) $Y = 1$, 50 points following $X \sim N(\mu_2, \Sigma_2)$ and 50 points following $X \sim N(\mu_3, \Sigma_3)$, $\mu_2 = (0.75, 0.75)$ and $\mu_3 = (-0.75, -0.75)$, the standard deviation in both dimensions are 0.2 and ρ_2, ρ_3 are equal to 0.7.

Consequences of the Theorems - Clustering

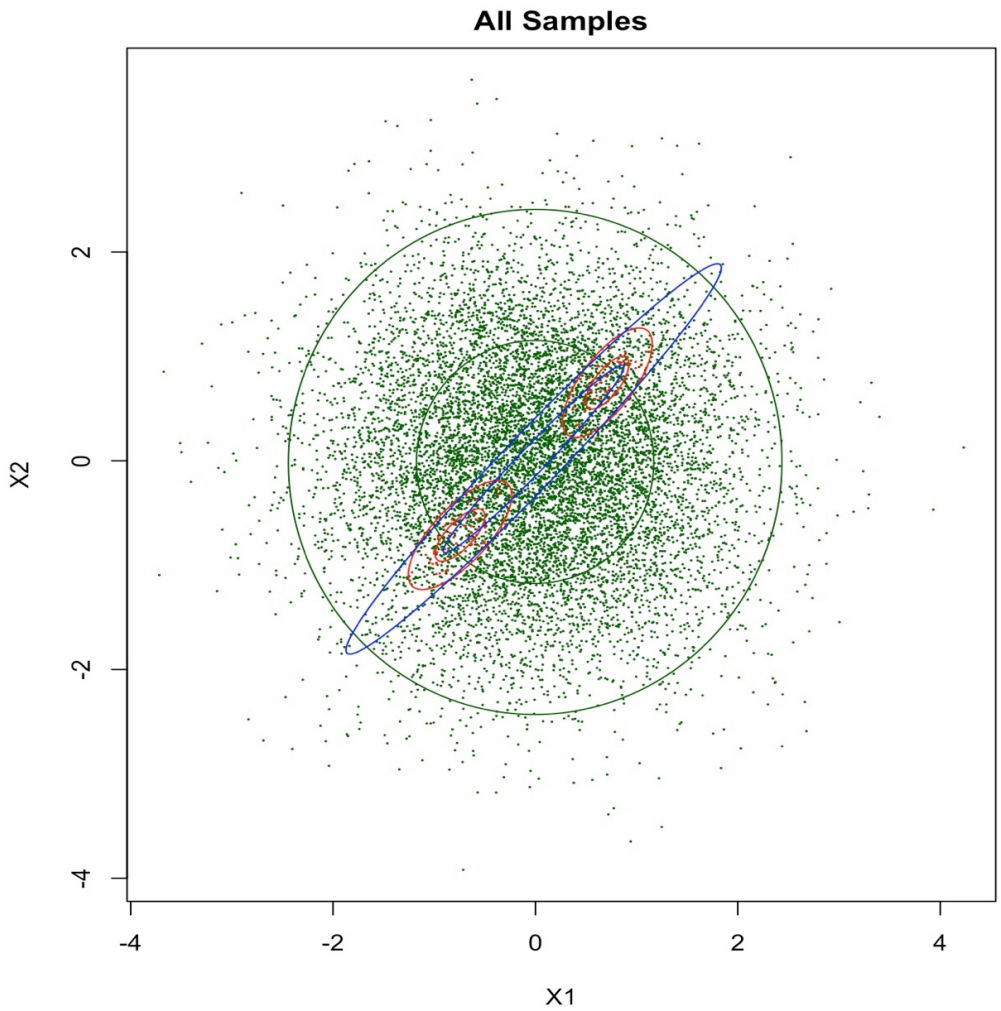


Figure: Scatter plot of sample sets.

Consequences of the Theorems - Clustering

The first model is a **logistic regression model**, the second one is a **multinomial logistic regression model** which has **one majority class** ($c1, Y = 0$) and **two separate minority classes** ($c2$ and $c3$, $Y = 1$) which are separated by K-means.

Table: Coefficient estimates

Coefficients	Estimate	$Pr(>\|z\|)$	Cluster	Coefficients	Estimate	$Pr(>\|t\|)$
Intercept	-4.6052	$< 2 \times 10^{-16}$	$c2$	Intercept	-5.8225	$< 2.2 \times 10^{-16}$
x_1	0.0107	0.914	$c3$	Intercept	-5.7971	$< 2.2 \times 10^{-16}$
x_2	0.0035	0.972	$c2$	$x1$	0.7258	5.260×10^{-7}
			$c3$	$x1$	-0.7012	7.425×10^{-7}
			$c2$	$x2$	0.7216	8.564×10^{-7}
			$c3$	$x2$	-0.7215	5.578×10^{-7}

The prediction AUC of logistic regression is 0.503, which is lower than AUC of multinomial logistic regression 0.684.

Consequences of the Theorems - Freddie Mac Data

We investigated on Freddie Mac mortgage credit data (detail will be presented in another session). Our initial experiments show that a logistic regression model has significant declining performance in the financial crisis period.

Based on the previous research, we propose the following two-step procedure to enhance the performance of logistic regression:

1. **Variable deletion**: We delete the variables which are constant in minority class (because of linear separation).
2. **Clustering**: after variables deletion, we use both K-medians and hierarchical clustering to split minority class into two clusters and then deploy multinomial logistic regression.

Consequences of the Theorems - Freddie Mac Data

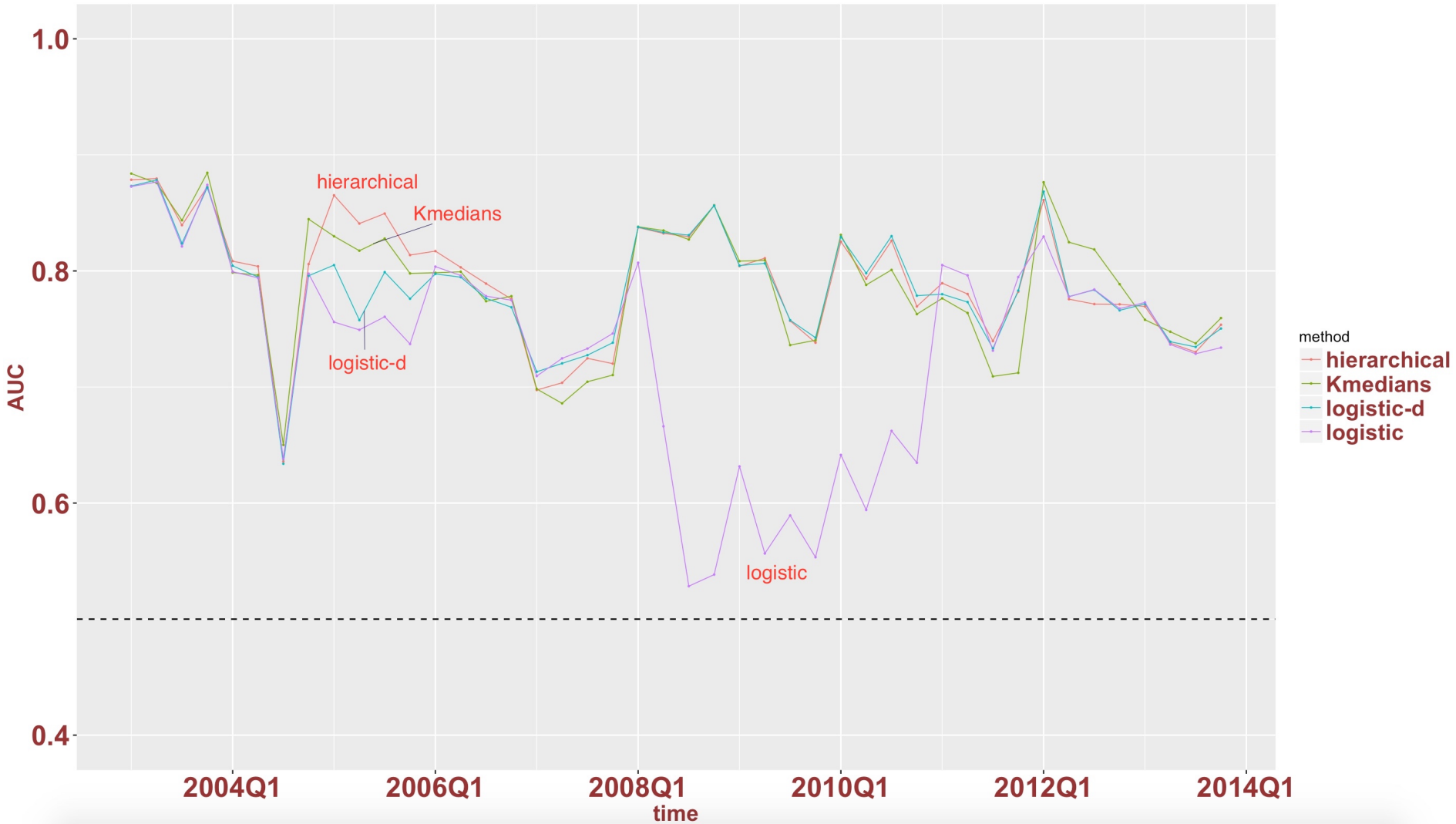


Figure: Forecast AUC from 2003 to 2013.

Conclusion

- Linear separation in data makes the MLE of logistic regression not exist.
- Highly imbalanced logistic regression only uses the minority class mean vector. Clustering is a choice to fix this problem.
- By adding penalty terms, penalized logistic regression can force MLE existing in linear separated data.
- However, highly imbalanced penalized logistic regression approximately does not utilize data.

References I

[1] A. B. Owen.
Infinitely imbalanced logistic regression.
Journal of Machine Learning Research, 8(Apr):761–773, 2007.

[2] M. J. Silvapulle.
On the existence of maximum likelihood estimators for the binomial response models.
Journal of the Royal Statistical Society. Series B (Methodological), pages 310–313, 1981.

[3] L. C. Thomas.
Consumer credit models: pricing, profit and portfolios: pricing, profit and portfolios.
OUP Oxford, 2009.

Thanks for your attention! Any questions?

Appendix: Infinite Imbalance

In order to avoid perfect separation, we use following definition to describe overlap condition infinitely imbalanced logistic regression:

Definition

(Definition 3) The distribution F on R^d has a point $x_\star$ surrounded if

$$\int_{(x-x_\star)^T\omega\geq\epsilon} dF(x) > \delta \tag{4}$$

holds for some $\epsilon > 0$, some $\delta > 0$ and all $\omega \in \Omega$. Here $\Omega = \{\omega \in R^d | \omega\omega^T = 1\}$.

Appendix: Infinite Imbalance

- We also assume that

$$\int e^{x^T\beta}(1+|x|)dF_0(x) < \infty \tag{5}$$

for all $\beta \in R^d$, because the heavy tails of F_0 will degenerate logistic regression.

- Thus, the log-likelihood function simplifies to

$$\begin{aligned} l(\beta_0,\beta) = n\beta_0 - \sum_{i=1}^{n} \log(1+e^{\beta_0+(x_{1i}-\overline{x})^T\beta}) \\ - N\int \log(1+e^{\beta_0+(x-\overline{x})^T\beta})dF_0(x) \end{aligned} \tag{6}$$