Infinitely Imbalanced Logistic Regression

Art B. Owen   OWEN@STAT.STANFORD.EDU
Department of Statistics
Stanford University
Stanford CA, 94305, USA

Editor: Yi Lin

Abstract

In binary classification problems it is common for the two classes to be imbalanced: one case is very rare compared to the other. In this paper we consider the infinitely imbalanced case where one class has a finite sample size and the other class's sample size grows without bound. For logistic regression, the infinitely imbalanced case often has a useful solution. Under mild conditions, the intercept diverges as expected, but the rest of the coefficient vector approaches a non trivial and useful limit. That limit can be expressed in terms of exponential tilting and is the minimum of a convex objective function. The limiting form of logistic regression suggests a computational shortcut for fraud detection problems.

Keywords: classification, drug discovery, fraud detection, rare events, unbalanced data

1. Introduction

In many applications of logistic regression one of the two classes is extremely rare. In political science, the occurrence of wars, coups, vetos and the decisions of citizens to run for office have been modelled as rare events; see King and Zeng (2001). Bolton and Hand (2002) consider fraud detection, and Zhu et al. (2005) look at drug discovery. In other examples the rare event might correspond to people with a rare disease, customer conversions at an e-commerce web site, or false positives among a set of emails marked as spam.

We will let $Y \in \{0, 1\}$ denote a random response with the observed value of Y being $y = 1$ in the rare case and $y = 0$ in the common case. We will suppose that the number of observations with $y = 0$ is so large that we have a satisfactory representation of the distribution of predictors in that setting. Then we explore the limit as the number of $y = 0$ cases tends to infinity while the number of observed cases with $y = 1$ remains fixed. It is no surprise that the intercept term in the logistic regression typically tends to $-\infty$ in this limit. The other coefficients can however tend to a useful limit.

The main result (Theorem 8 below) is that under reasonable conditions, the intercept term tends to $-\infty$ like $-\log(N)$ plus a constant, while the limiting logistic regression coefficient $\beta = \beta(N)$ satisfies

$$\bar{x} = \frac{\int e^{x'\beta} x \, dF_0(x)}{\int e^{x'\beta} \, dF_0(x)} \tag{1}$$

where F_0 is the distribution of X given $Y = 0$ and $\bar{x}$ is the average of the sample x_i values for which $y = 1$. The limiting solution is the exponential tilt required to bring the population mean of X given $Y = 0$ onto the sample mean of X given $Y = 1$.

When F_0 is the $N(\mu_0, \Sigma_0)$ distribution for finite nonsingular Σ_0 then

$$\lim_{N \to \infty} \beta(N) = \Sigma_0^{-1}(\bar{x} - \mu_0). \tag{2}$$

Equation (2) reminds one of a well known derivation for logistic regression. If the conditional distribution of X given that $Y = y$ is $N(\mu_y, \Sigma)$ then the coefficient of X in logistic regression is $\Sigma^{-1}(\mu_1 - \mu_0)$. Equation (2) however holds without assuming that the covariance of X is the same for $Y = 0$ and $Y = 1$, or even that X is Gaussian given that $Y = 1$.

The outline of the paper is as follows. Section 2 gives three numerical examples that illustrate the limiting behavior of β. One is a positive result in which we see β approaching the value computed from (2). The other two examples are negative results where (1) does not hold. Each negative case illustrates the failure of an assumption of Theorem 8. In one case there is no nontrivial estimate of β at any $N > 0$ while in the other β diverges as $N \to \infty$. Section 3 formally introduces the notation of this paper. It outlines the results of Silvapulle (1981) who completely characterizes the conditions under which unique logistic regression estimates exist in the finite sample case. The infinite sample case differs importantly and requires further conditions. A stronger overlap condition is needed between the two X distributions. Also, the distribution of X given $Y = 0$ must not have tails that are too heavy, an issue that cannot arise in finite samples from $\mathbb{R}^d$. Section 4 proves the results in this paper. A surprising consequence of Equation (1) is that the x values when $y = 1$ only appear through their average $\bar{x}$. Section 5 shows that for the drug discovery example of Zhu et al. (2005), we can replace all data points with $y = 1$ by a single one at $(\bar{x}, 1)$ with minimal effect on the estimated coefficient, apart from the intercept term. Section 6 discusses how these results can be used in deciding which unlabelled data points to label, and it shows how the infinitely imbalanced setting may lead to computational savings.

We conclude this introduction by relating the present work to the literature on imbalanced data. The English word "unbalanced" seems to be more popular, at least on web pages, than is "imbalanced". But the latter term has been adopted for this special setting in two recent workshops: AAAI 2000 and ICML 2003, respectively Japkowicz (2000) and Chawla et al. (2003). An extensive survey of the area is given by Chawla et al. (2004). In that literature much attention is paid to undersampling methods in which some of the available cases with $Y = 0$ are either randomly or strategically removed to alleviate the imbalance. Another approach is oversampling in which additional, possibly synthetic, cases are generated with $Y = 1$. It is also clear that prediction accuracy will be very good for a trivial method that always predicts $y = 0$ and so one needs to take care about misclassification cost ratios and prior probability ratios.

2. Numerical Examples

For illustration, suppose that when $Y = 0$ that $X \sim N(0, 1)$ and that we have one single observation with $y = 1$ and it has $x = 1$. To study this case we use logistic regression on $(x_i, y_i) = (\Phi^{-1}((i - 1/2)/N), 0)$ for $i = 1, \ldots, N$ and $(x_{N+1}, y_{N+1}) = (1, 1)$. Here Φ is the cumulative distribution function (CDF) of the $N(0, 1)$ distribution. As N increases the problem becomes more imbalanced and the N points used produce an ever better approximation to the normal distribution. Taking stratified X_i reduces inessential variation in the computation making the convergence pattern

N	α	Ne^{α}	β
10	-3.19	0.4126	1.5746
100	-5.15	0.5787	1.0706
1,000	-7.42	0.6019	1.0108
10,000	-9.71	0.6058	1.0017
100,000	-12.01	0.6064	1.0003

Table 1: Logistic regression intercept α and coefficient β for imbalanced data described in the text. There are N observations with $Y = 0$ and stratified $X \sim N(0,1)$ and one observation with $Y = 1$ and $X = 1$.

clearer. Some resulting values are shown in Table 1. From this table it seems clear that as $N \to \infty$, the intercept term is diverging like $-\log(N)$ while the coefficient of X is approaching the value 1 that we would get from Equation (2). Theorem 8 below shows that such is indeed the limiting behavior.

Next we repeat the computation replacing Φ by the CDF of the standard Cauchy distribution with density $1/(\pi(1+x^2))$. The results are shown in Table 2. Here it is clear that $\beta \to 0$ as $N \to \infty$ and α appears to behave like a constant minus $\log(N)$. It is not surprising that $\beta \to 0$ in this limit. The Cauchy distribution has tails far heavier than the logistic distribution. If $\beta \neq 0$ then the log likelihood (4) that we introduce in Section 3 is $-\infty$. The likelihood is maximized at $\beta = 0$ and $\alpha = -\log(N+1)$. We get slightly different values in Table 2 because the uniform distribution over N Cauchy quantiles that we use has lighter tails than the actual Cauchy distribution it approaches. The heavy tails of the Cauchy distribution make it fail a condition of Theorem 8. The finite sample setting does not need a tail condition on the distribution of X, beyond an assumption that all observed values are finite.

In the next example we use the $U(0,1)$ distribution for X given $Y = 0$. This time we use $n = 2$ points with $y = 1$. One has $x = 1/2$ and the other has $x = 2$. The results are shown in Table 3. Once again the value β does not appear to be converging to a limit. It cannot be due to heavy tails, because the $U(0,1)$ distribution has bounded support. On further thought, we see that $\bar{x} = 5/4$. There is no possible way for an exponential tilt like (1) to reweight the $U(0,1)$ distribution to have mean $5/4$. This example also fails one of the conditions of Theorem 8. We need the point $\bar{x}$ to be surrounded by the distribution of X given $Y = 0$ as defined in Section 3. Such a requirement is stronger than

N	α	Ne^{α}	β	Ne^{β}
10	-2.36	0.94100	0.1222260	1.2222
100	-4.60	0.99524	0.0097523	0.9752
1,000	-6.90	0.99953	0.0009537	0.9536
10,000	-9.21	0.99995	0.0000952	0.9515
100,000	-11.51	0.99999	0.0000095	0.9513

Table 2: Logistic regression intercept α and coefficient β for imbalanced data described in the text. There are N observations with $Y = 0$ and stratified X from the standard Cauchy distribution, and one observation with $Y = 1$ and $X = 1$.

763

N	α	Ne^{α}	β	e^{β}/N
10	-3.82	0.2184	2.85	1.74
100	-7.13	0.0804	4.19	0.66
1,000	-10.71	0.0223	5.82	0.34
10,000	-14.52	0.0050	7.62	0.20
100,000	-18.49	0.0009	9.54	0.14

Table 3: Logistic regression intercept α and coefficient β for imbalanced data described in the text. There are N observations with $Y = 0$ and stratified $X \sim U(0,1)$ and two observations with $Y = 1$, one with $X = 1/2$, the other with $X = 2$.

what is needed in the finite sample setting. Empirically e^{α} and e^{β} both appear to follow a power law in N but we do not investigate this further, focusing instead on the case where β approaches a non-trivial limit.

3. Notation

The data are (x,y) pairs where $x \in \mathbb{R}^d$ and $y \in \{0,1\}$. There are n observations with $y = 1$ and N with $y = 0$. The difference in case serves to remind us that $n \ll N$. The values of x when $y = 1$ are $x_{11}, \ldots, x_{1n}$. The values of x when $y = 0$ are $x_{01}, \ldots, x_{0N}$. Singly subscripted values x_i represent x_{1i}. Sometimes we use n_1 for n and n_0 for N.

The logistic regression model is $\Pr(Y = 1 \mid X = x) = e^{\alpha + x'\beta}/(1 + e^{\alpha + x'\beta})$ for $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$. The log-likelihood in logistic regression is

$$\sum_{i=1}^{n} \left\{ \alpha + x'_{1i}\beta - \log(1 + e^{\alpha + x'_{1i}\beta}) \right\} - \sum_{i=1}^{N} \log(1 + e^{\alpha + x'_{0i}\beta}). \tag{3}$$

We suppose that a good approximation can be found for the conditional distribution of X given that $Y = 0$, as seems reasonable when N is very large. For continuously distributed X we might then replace the second sum in (3) by $N \int \log(1 + \exp(1 + e^{\alpha + x'\beta})) f_0(x) \, dx$ where f_0 is a probability density function. Because some or all of the components of X might be discrete we work instead with a distribution function F_0 for X given that $Y = 0$.

With a bit of foresight we also center the logistic regression around the average $\bar{x} = \sum_{i=1}^{n} x_i / n$ of the predictor values for cases with $Y = 1$. Then the log likelihood we work with simplifies to

$$\ell(\alpha, \beta) = n\alpha - \sum_{i=1}^{n} \log(1 + e^{\alpha + (x_i - \bar{x})'\beta}) - N \int \log(1 + e^{\alpha + (x - \bar{x})'\beta}) \, dF_0(x) \tag{4}$$

where the $n\alpha$ term arises as $\sum_{i=1}^{n} \alpha + (x_i - \bar{x})'\beta$.

When the centered log likelihood ℓ has an MLE $(\hat{\alpha}_0, \hat{\beta})$ we can recover the MLE of the uncentered log likelihood easily: $\hat{\beta}$ remains unchanged while $\hat{\alpha}$ in the uncentered version is $\hat{\alpha}_0 - \bar{x}'\hat{\beta}$. The numerical examples in Section 2 used uncentered logistic regression.

Here we study the maximizer $(\hat{\alpha}, \hat{\beta})$ of (4) in the limit as $N \to \infty$ with n and $x_1, \ldots, x_n$ held fixed. It is reasonable to suppose that $\hat{\alpha} \to -\infty$ in this limit. Indeed we anticipate e^{α} should be $O(1/N)$ since the proportion of observations with $y = 1$ in the data is $n/(N+n) \doteq n/N$. What is interesting and important is that $\hat{\beta}$ does not necessarily diverge.

3.1 Silvapulle's Results

It is well known that the MLE in the usual logistic regression setting can fail to be finite when the x values where $y = 1$ are linearly separable from those where $y = 0$.

The existence and uniqueness of MLE's for linear logistic regression has been completely characterized by Silvapulle (1981). He works in terms of binary regression through the origin. To employ an intercept, one uses the usual device of adjoining a predictor component that is always equal to 1.

For $y = 1$ let $z_{1i} = (1, x'_{1i})'$ for $i = 1, \ldots, n_1$ and for $y = 0$ let $z_{0i} = (1, x'_{0i})'$ for $i = 1, \ldots, n_0$. Let $\theta = (\alpha, \beta')'$. Then the logistic regression model has $\Pr(Y = 1 \mid X = x) = \exp(z'\theta)/(1 + \exp(z'\theta))$ where of course $z = z(x) = (1, x')'$. Silvapulle (1981) employs two convex cones:

$$C_j = \left\{ \sum_{i=1}^{n_j} k_{ji} z_{ji} \mid k_{ji} > 0 \right\}, \quad j \in \{0, 1\}.$$

Theorem 1 *For data as described above, assume that the $n_0 + n_1$ by $d + 1$ matrix with rows taken from z_{ji} for $j = 0, 1$ and $i = 1, \ldots, n_j$ has rank $d + 1$. If $C_0 \cap C_1 \neq \emptyset$ then a unique finite logistic regression MLE $\hat{\theta} = (\hat{\alpha}, \hat{\beta}')$ exists. If however $C_0 \cap C_1 = \emptyset$ then no MLE exists.*

Proof: This result follows from clause (iii) of the Theorem on page 311 of Silvapulle (1981). $\square$

Silvapulle (1981) has more general results. Theorem 1 also holds when the logistic CDF $G(t) = \exp(t)/(1 + \exp(t))$ is replaced by the standard normal one (for probit analysis) or by the $U(0, 1)$ CDF. Any CDF G for which both $-\log G(t)$ and $-\log(1 - G(t))$ are convex, and for which $G(t)$ is strictly increasing when $0 < G(t) < 1$ obeys the same theorem. The CDF G cannot be the Cauchy CDF, because the Cauchy CDF fails the convexity conditions.

The cone intersections may seem unnatural. A more readily interpretable condition is that the relative interior (as explained below) of the convex hull of the x's for $y = 0$ intersects that for $y = 1$. That is $H_0 \cap H_1 \neq \emptyset$ where

$$H_j = \left\{ \sum_{i=1}^{n_j} \lambda_{ji} x_{ji} \mid \lambda_{ji} > 0, \sum_{i=1}^{n_j} \lambda_{ji} = 1 \right\}.$$

When the x_{ji} span $\mathbb{R}^d$ then H_j is the interior of the convex hull of x_{ji}. When x_{ji} lie in a lower dimensional affine subspace of $\mathbb{R}^d$ then the interior of their convex hull is the empty set. However the interior with respect to that subspace, called the relative interior, and denoted H_j above is not empty. In the extreme where $x_{ji} = x_{j1}$ for $i = 1, \ldots, n_j$, then the desired relative interior of the convex hull of $x_{j1}, \ldots, x_{jn_j}$ is simply $\{x_{j1}\}$.

Lemma 2 *In the notation above $H_0 \cap H_1 \neq \emptyset$ if and only if $C_0 \cap C_1 \neq \emptyset$.*

Proof: Suppose that $x_0 \in H_0 \cap H_1$. Then $z_0 = (1, x'_0)' \in C_0 \cap C_1$. Conversely suppose that $z_0 \in C_0 \cap C_1$. Then we may write

$$z_0 = \sum_{i=1}^{n_0} k_{0i} \begin{pmatrix} 1 \\ x_{0i} \end{pmatrix} = \sum_{i=1}^{n_1} k_{1i} \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix},$$

where each $k_{ji} > 0$. From the first component of z_0 we find a common positive value for $\sum_{i=1}^{n_0} k_{0i}$ and $\sum_{i=1}^{n_1} k_{1i}$. Let K denote that value, and put $\lambda_{ji} = k_{ji}/K$ for $j = 0, 1$ and $i = 1, \ldots, n_j$. Then $x_0 = \sum_{i=1}^{n_0} \lambda_{0i} x_{0i} = \sum_{i=1}^{n_1} \lambda_{1i} x_{1i} \in H_0 \cap H_1$. $\square$

3.2 Overlap Conditions

In light of Silvapulle's results we expect that we will need to assume some overlap between the data $x_1, \ldots, x_n$ from the 1s and the distribution F_0 of X given $Y = 0$ in order to get a nontrivial result. The setting here with $N \to \infty$ is different and requires a stronger, but still very weak, overlap condition. In describing this condition, we let $\Omega = \{\omega \in \mathbb{R}^d \mid \omega'\omega = 1\}$ be the unit sphere in $\mathbb{R}^d$.

Definition 3 *The distribution F on $\mathbb{R}^d$ has the point x_* surrounded if*

$$\int_{(x-x_*)'\omega > \varepsilon} dF(x) > \delta \tag{5}$$

holds for some $\varepsilon > 0$, some $\delta > 0$ and all $\omega \in \Omega$.

We will make use of two simple immediate consequences of (5). If F has the point x_* surrounded, then there exist η and γ satisfying

$$\inf_{\omega \in \Omega} \int_{(x-x_*)'\omega \geq 0} dF(x) \geq \eta > 0 \tag{6}$$

and

$$\inf_{\omega \in \Omega} \int [(x-x_*)'\omega]_+ \, dF(x) \geq \gamma > 0 \tag{7}$$

where $Z_+ = \max(Z, 0)$ is the positive part of Z. For example we can take $\eta = \delta$ in (6) and $\gamma = \varepsilon\delta$ in (7). Notice that F cannot surround any point if F concentrates in a low dimensional affine subset of $\mathbb{R}^d$. This implies that having at least one point surrounded by F_0 will be enough to avoid rank deficiency.

In Theorem 1 it follows from Lemma 2 that we only need there to be some point x_* that is surrounded by both $\hat{F}_0$ and $\hat{F}_1$ where $\hat{F}_j$ is the empirical distribution of $x_{j1}, \ldots, x_{jn_j}$. If such x exists, we get a unique finite MLE. (Recall that Theorem 1 assumes full rank for the predictors.)

In the infinitely imbalanced setting we expect that F_0 will ordinarily surround every single one of $x_1, \ldots, x_n$. We do not need F_0 to surround them all but it is not enough to just have some point x_* exist that is surrounded by both F_0 and $\hat{F}_1$. We need to assume that F_0 surrounds $\bar{x}$. We do not need to assume that $\hat{F}_1$ surrounds $\bar{x}$, a condition that fails when the x_i are confined to an affine subset of $\mathbb{R}^d$ as they necessarily are for $n < d$.

There is an interesting case in which F_0 can fail to surround $\bar{x}$. The predictor X may contain a component that is itself an imbalanced binary variable, and that component might never take the value 1 in the $y = 1$ sample. Then $\bar{x}$ is right on the boundary of the support of F_0 and we cannot be sure of a finite β in either the finite sample case or the infinitely imbalanced case.

3.3 Technical Lemmas

The first technical lemma below is used to get some bounds. The second one establishes existence of a finite MLE when $N < \infty$.

Lemma 4 *For $\alpha, z \in \mathbb{R}$,*

$$\begin{aligned}
e^{\alpha+z} \geq \log(1 + e^{\alpha+z}) &\geq [\log(1 + e^\alpha) + ze^\alpha/(1 + e^\alpha)]_+ \\
&\geq [ze^\alpha/(1 + e^\alpha)]_+ = z_+ \, e^\alpha/(1 + e^\alpha).
\end{aligned} \tag{8}$$

Proof: For the leftmost inequality, apply $x \geq \log(1+x)$ to $x = e^{\alpha+z}$. For the others, the function $h(z) = \log(1 + e^{\alpha+z})$ is convex and positive. Therefore $h(z) \geq [h(0) + zh'(0)]_+ \geq [zh'(0)]_+ = z_+ h'(0)$. $\square$

Lemma 5 *Let $n \geq 1$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be given, and assume that the distribution F_0 surrounds $\bar{x} = \sum_{i=1}^n x_i/n$ and that $0 < N < \infty$. Then the log likelihood $\ell(\alpha, \beta)$ given by (4) has a unique finite maximizer $(\hat{\alpha}, \hat{\beta})$.*

Proof: The log likelihood ℓ is strictly concave in (α, β). It either has a unique finite maximizer or it grows forever along some ray $\{(\lambda\alpha_0, \lambda\beta_0) \mid 0 \leq \lambda < \infty\} \subset \mathbb{R}^{d+1}$. By following such a ray back to where it intersects a small cylinder around the origin we may assume that either $0 \leq |\alpha_0| < \varepsilon/2$ and $\beta_0'\beta_0 = 1$, where ε is the constant in Definition 3, or that $0 < |\alpha_0| < \varepsilon/2$ and $\beta_0 = 0$. We will show that $\partial\ell(\lambda\alpha_0, \lambda\beta_0)/\partial\lambda$ is always strictly negative, ruling out infinite growth and thus establishing a unique finite maximizer.

For $\beta_0 = 0$ and $\alpha_0 > 0$ we find $\lim_{\lambda\to\infty} \partial\ell(\lambda\alpha_0, \lambda\beta_0)/\partial\lambda = -N\alpha_0 < 0$. For $\beta_0 = 0$ and $\alpha_0 < 0$ we find $\lim_{\lambda\to\infty} \partial\ell(\lambda\alpha_0, \lambda\beta_0)/\partial\lambda = n\alpha_0 < 0$.

Now suppose $\beta_0'\beta_0 = 1$ and $|\alpha_0| < \varepsilon/2$. Using $n\alpha_0 = \sum_{i=1}^n \alpha_0 + (x_i - \bar{x})'\beta_0$, we find

$$\lim_{\lambda\to\infty} \frac{\partial}{\partial\lambda} \ell(\lambda\alpha_0, \lambda\beta_0) = \sum_{i:\alpha_0+(x_i-\bar{x})'\beta_0<0} \alpha_0 + (x_i - \bar{x})'\beta_0 \\ - N \int_{\alpha_0+(x-\bar{x})'\beta_0>0} \alpha_0 + (x - \bar{x})'\beta_0 \, dF_0(x). \tag{9}$$

The sum in (9) is either 0 or is negative and the integral is either 0 or is positive. For the integral to be 0 we must have $(x - \bar{x})'\beta_0 \leq -\alpha_0$ with probability one for $x \sim F_0$. But this is impossible because F_0 has $\bar{x}$ surrounded. $\square$

4. Main Results

Lemma 6 below shows that, as anticipated, $e^{\hat{\alpha}}$ is typically $O(1/N)$ as $N \to \infty$. Specifically, we find a bound $B = 2n/\eta < \infty$ for which $\limsup_{N\to\infty} Ne^{\hat{\alpha}} < B$.

Lemma 6 *Under the conditions of Lemma 5, let $\hat{\alpha}$ and $\hat{\beta}$ maximize ℓ of (4). Let η satisfy (6). Then for $N \geq 2n/\eta$ we have $e^{\hat{\alpha}} \leq 2n/(N\eta)$.*

Proof: Let β be any point in $\mathbb{R}^d$. Write $e^\alpha = A/N$ for $0 < A < \infty$. Then

$$\frac{\partial}{\partial\alpha} \ell = n - \sum_{i=1}^n \frac{AN^{-1}e^{(x_i-\bar{x})'\beta}}{1 + AN^{-1}e^{(x_i-\bar{x})'\beta}} - N \int \frac{AN^{-1}e^{(x-\bar{x})'\beta}}{1 + AN^{-1}e^{(x-\bar{x})'\beta}} \, dF_0(x)$$

$$\leq n - A \int_{(x-\bar{x})'\beta\geq 0} \frac{e^{(x-\bar{x})'\beta}}{1 + AN^{-1}e^{(x-\bar{x})'\beta}} \, dF_0(x)$$

$$\leq n - \frac{A\eta}{1 + A/N}.$$

Now suppose that $N \geq 2n/\eta$ and that $e^{\alpha} > 2n/(N\eta)$, that is $A > 2n/\eta$. Then $\partial \ell / \partial \alpha < 0$. Because ℓ is concave this negative partial derivative implies that

$$\arg\max_{\alpha} \ell(\alpha, \beta) < \log(2n/\eta) - \log(N). \tag{10}$$

Because β was arbitrary (10) holds for all $\beta \in \mathbb{R}^d$. Lemma 5 implies that $\hat{\beta}$ is finite, and so (10) applies for $\beta = \hat{\beta}$. $\square$

Lemma 7 *Under the conditions of Lemma 5, let $\hat{\alpha}$ and $\hat{\beta}$ maximize ℓ of (4). Then $\limsup_{N\to\infty} \|\hat{\beta}\| < \infty$.*

Proof: Let $e^{\alpha} = A/N$ for $A > 0$ and let $\beta \in \mathbb{R}^d$. Pick γ to satisfy (7). Then

$$\ell(\alpha, 0) - \ell(\alpha, \beta)$$

$$= -(n+N)\log(1+e^{\alpha}) + \sum_{i=1}^{n} \log(1 + e^{\alpha + (x_i - \bar{x})'\beta})$$

$$+ N \int \log(1 + e^{\alpha + (x - \bar{x})'\beta}) \, dF_0(x)$$

$$> -(n+N)e^{\alpha} + N\frac{e^{\alpha}}{1+e^{\alpha}} \int_{(x-\bar{x})'\beta \geq 0} (x - \bar{x})'\beta \, dF_0(x)$$

$$\geq -(n+N)\frac{A}{N} + A\frac{\|\beta\|\gamma}{1+A/N}$$

after applying two inequalities from (8) and making some simplifications. If follows that $\ell(\alpha, \beta) < \ell(\alpha, 0)$ whenever $\|\beta\| \geq \gamma^{-1}(1+A/N)(1+n/N)$. For large enough N we have $\|\hat{\beta}\| \leq 2/\gamma$, using Lemma 6 to control A. $\square$

As illustrated in Section 2, infinitely imbalanced logistic regression will be degenerate if F_0 has tails that are too heavy. We assume that

$$\int e^{x'\beta}(1 + \|x\|) dF_0(x) < \infty, \quad \forall \beta \in \mathbb{R}^d. \tag{11}$$

Condition (11) is satisfied by distributions with bounded support and by light tailed distributions such as the multivariate normal distribution.

Theorem 8 *Let $n \geq 1$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ be fixed and suppose that F_0 satisfies the tail condition (11) and surrounds $\bar{x} = \sum_{i=1}^{n} x_i/n$ as described at (5). Then the maximizer $(\hat{\alpha}, \hat{\beta})$ of ℓ given by (4) satisfies*

$$\lim_{N\to\infty} \frac{\int e^{x'\hat{\beta}} x \, dF_0(x)}{\int e^{x'\hat{\beta}} dF_0(x)} = \bar{x}.$$

Proof: Setting $\partial \ell / \partial \beta = 0$, dividing by $Ne^{\hat{\alpha} - \bar{x}'\hat{\beta}}$ and rearranging terms, gives

$$\int \frac{(x - \bar{x})e^{x'\hat{\beta}}}{1 + e^{\hat{\alpha} + (x - \bar{x})'\hat{\beta}}} dF_0(x) = -\frac{1}{N} \sum_{i=1}^{n} \frac{e^{x_i'\hat{\beta}}(x_i - \bar{x})}{1 + e^{\hat{\alpha} + (x_i - \bar{x})'\hat{\beta}}}. \tag{12}$$

Method	α	β_1	β_2	β_3	β_4	β_5	β_6
Original	-3.707	4.629	4.807	0.398	0.594	0.170	0.130
Single $y = 1$	-10.116	4.623	4.984	0.397	0.595	0.193	0.182
$x_{1j} = \bar{x}$	-3.701	4.765	5.136	0.410	0.614	0.204	0.190
SE	0.041	0.696	0.851	0.040	0.130	0.299	0.413

Table 4: This table shows logistic regression coefficients for the chemical compound data set described in the text. The top row shows ordinary logistic regression coefficients. The second row shows the coefficients when the cases with $y = 1$ are deleted and replaced by a single point $(\bar{x}, 1)$. The third row shows the coefficients when all 608 cases with $y = 1$ are replaced by $(\bar{x}, 1)$. The fourth row shows standard errors for the ordinary logistic regression coefficients in the top row.

As $N \to \infty$ the right side of (12) vanishes because $\|\hat{\beta}\|$ is bounded as $N \to \infty$ by Lemma 7. Therefore the MLEs satisfy

$$\lim_{N \to \infty} \frac{\int x e^{x'\hat{\beta}}[1 + e^{\hat{\alpha}+(x-\bar{x})'\hat{\beta}}]^{-1} dF_0(x)}{\int e^{x'\hat{\beta}}[1 + e^{\hat{\alpha}+(x-\bar{x})'\hat{\beta}}]^{-1} dF_0(x)} = \bar{x}. \tag{13}$$

The denominator of (13) is at most $\int e^{x'\hat{\beta}} dF_0(x)$ and is at least

$$\int e^{x'\hat{\beta}}(1 - e^{\hat{\alpha}+(x-\bar{x})'\hat{\beta}}) dF_0(x) \to \int e^{x'\hat{\beta}} dF_0(x)$$

as $N \to \infty$ because $\alpha \to -\infty$ and $\int e^{2x'\hat{\beta}} dF_0(x) < \infty$ by the tail condition (11). Therefore the denominator of (13) has the same limit as $\int e^{x'\hat{\beta}} dF_0(x)$ as $N \to \infty$. Similarly the numerator has the same limit as $\int e^{x'\hat{\beta}} x \, dF_0(x)$. The limit for the denominator is finite and nonzero, and so the result follows.
$\square$

5. Illustration

It is perhaps surprising that in the $N \to \infty$ limit, the logistic regression depends on $x_1, \ldots, x_n$ only through $\bar{x}$. The precise configuration of those n points in $\mathbb{R}^d$ becomes unimportant. We could rotate them about $\bar{x}$, or replace each of them by $\bar{x}$, or even replace them by one single point at $\bar{x}$ with $Y = 1$ and still get the same $\hat{\beta}$ in the $N \to \infty$ limit.

To investigate whether this effect can hold in finite data sets, we look at an example from Zhu et al. (2005). They study a data set with 29,812 chemical compounds on which 6 predictor variables were measured. Compounds were rated as active ($Y = 1$) or inactive ($Y = 0$) and only 608 of the compounds were active.

Table 4 shows the logistic regression coefficients for this data, as well as what happens to them when we replace the 608 data points (x, y) with $y = 1$ by a single point at $(\bar{x}, 1)$, or by 608 points equal to $(\bar{x}, 1)$. In a centered logistic regression the point $(\bar{x}, 1)$ becomes $(\bar{x} - \bar{x}, 1) = (0, \ldots, 0, 1) \in \mathbb{R}^{d+1}$.

The intercept changes a lot when we reduce the rare cases from 608 to 1 but otherwise the coefficients do not change importantly. Interestingly the single point version has a β vector closer

to the original logistic regression than has the version with 608 points at $(\bar{x}, 1)$. The differences in β are quite small compared to the sampling uncertainty. We would reach the same conclusions about which predictors are most important in all three cases.

The linear predictor $\hat{\alpha} + \hat{\beta}'(x - \bar{x})$ was computed using the coefficients from each of these models (taking care to use the original x_i's not the versions set to $\bar{x}$.) The correlation between the linear predictor from logistic regression to that fit with all $x_i = \bar{x}$ is 0.999881. The correlation between the linear predictor from logistic regression to that fit with just one $(\bar{x}, 1)$ data point is still higher, at 0.999888. The two altered linear predictors have correlation 0.999998. Not surprisingly any two of these linear predictors plot as a virtual straight line. There will be no important differences in ROC curves, precision and recall curves or other performance measures among these three fits.

6. Discussion

This paper has focussed on establishing the limit of $\hat{\beta}$ as $N \to \infty$. This section presents some context and motivation. Section 6.1 shows these findings lead to greater understanding of how logistic regression works or fails and how to improve it. Section 6.2 shows how even after passing to the limit the resulting model makes some useful predictions. Section 6.3 illustrates the special case of F_0 that is Gaussian or a mixture of Gaussians. Section 6.4 describes how using infinitely imbalanced logistic regression may lead to cost savings in fraud detection settings.

6.1 Insight Into Logistic Regression

In the infinitely imbalanced limit, logistic regression only uses the $y = 1$ data points through their average feature vector $\bar{x}$. This limiting behavior is a property of logistic regression, not of any particular data set. It holds equally well in those problems for which logistic regression works badly as it does in problems where the Bayes rule is a logistic regression.

In the illustrative example we got almost the same logistic regression after replacing all the rare cases by a single point at $\bar{x}$. We would not expect this property for learning methods in general. For example classification trees such as those fit by CART (Breiman et al., 1984) will ordinarily change a lot if all of the $Y = 1$ cases are replaced by one or more points $(\bar{x}, 1)$.

Logistic regression only has d parameters apart from the intercept, so it is clear that it cannot be as flexible as some other machine learning methods. But knowing that those parameters are very strongly tied to the d components of $\bar{x}$ gives us insight into how logistic regression works on imbalanced problems. It is reasonable to expect better results from logistic regression when the x_{1i} are in a single tight cluster near $\bar{x}$ than when there are outliers, or when the x_{1i} points are in two well separated clusters in different directions from the bulk of F_0.

The insight also suggests things to do. For example when we detect outliers among the x_{1i}, shrinking them towards $\bar{x}$, or removing them should improve performance. When we detect sharp clusters among x_{1i} then we might fit one logistic regression per cluster, separating that cluster from the x_{0i}'s, and predict for new points by pooling the cluster specific results. Even an $O(n^2)$ clustering algorithm may be inexpensive in the $N \gg n$ setting.

6.2 Nontrivial Limiting Predictions

In the infinitely imbalanced limit with $N \to \infty$ we often find that $\hat{\beta}$ converges to a finite limit while $\hat{\alpha} \to -\infty$. This limit gives $\Pr(Y = 1 \mid X = x) \to 0$ for all x and so it gives trivial probabilities for

prediction purposes. But we are often interested in probability ratios with nontrivial limits such as:

$$\frac{\Pr(\widetilde{Y} = 1 \mid X = \widetilde{x})}{\Pr(Y = 1 \mid X = x)} \to e^{(\widetilde{x} - x)'\beta}.$$

For example if we are presented with a number of cases of potential fraud to investigate and have limited resources then we can rank them by $x'\beta$ and investigate as many of the most likely ones as time or other costs allow.

Because this rank is derived from a probability ratio we can also take into account the monetary or other measured value of the cases. If the values of uncovering fraud in the two cases are v and $\widetilde{v}$, respectively, then we might prefer to investigate the former when $ve^{x'\beta} > \widetilde{v}e^{\widetilde{x}'\beta}$. If the costs of investigation are c and $\widetilde{c}$ then we might prefer the former when $ve^{x'\beta}/c > \widetilde{v}e^{\widetilde{x}'\beta}/\widetilde{c}$.

In active learning problems one must choose which data to gather. There are several kinds of active learning, as described in Tong (2001). The interventional setting is very similar to statistical experimental design. For example, Cohn et al. (1996) describe how to select training data for feedforward neural networks. In the selective setting, the investigator has a mix of labelled cases (both x and y known) and unlabelled cases (x known but y unknown), and must choose which of the unlabelled cases to get a label for. For example the label y might indicate whether a human expert says that a document with feature vector x is on a specific topic. In a rare event setting, finding the cases most likely to have $y = 1$ is a reasonable proxy for finding the most informative cases, and one could then allocate a large part of the labelling budget to cases with high values of $x'\beta$.

6.3 Gaussian Mixtures F_0

When F_0 is a nonsingular Gaussian distribution then as remarked in the introduction, $\beta \to \Sigma_0^{-1}(\bar{x} - \mu_0)$. The effective sample size of an imbalanced data set is often considered to be simply the number of rare outcomes. The formula for β depends on the data only through $\bar{x}$, which as an average of n observations clearly has effective sample size of n. In the limit where $N \to \infty$ first and then $n \to \infty$ we get $\beta \to \Sigma_0^{-1}(\mu_1 - \mu_0)$ where $\mu_j = E(X \mid Y = j)$. A confidence ellipsoid for μ_1 translates directly into one for β.

Gaussian mixture models are a flexible and widely used method for approximating distributions. They have the further advantage for the present problem that exponential tilts of Gaussian mixtures are also Gaussian mixtures. The result is a convenient expression to be solved for β.

Suppose that

$$F_0 = \sum_{k=1}^{K} \lambda_k N(\mu_k, \Sigma_k)$$

where $\lambda_k > 0$ and $\sum_{k=1}^{K} \lambda_k = 1$. If at least one of the Σ_k has full rank then F_0 will surround the point $\bar{x}$. Then the limiting β is defined through

$$\bar{x} = \frac{\sum_{k=1}^{K} \lambda_k (\mu_k + \Sigma_k \beta) e^{\beta'\mu_k + \beta'\Sigma_k\beta/2}}{\sum_{k=1}^{K} \lambda_k e^{\beta'\mu_k + \beta'\Sigma_k\beta/2}},$$

so that β is the solution to

$$0 = \sum_{k=1}^{K} \lambda_k (\mu_k + \Sigma_k \beta - \bar{x}) e^{\beta'\mu_k + \beta'\Sigma_k\beta/2}. \tag{14}$$

Solving Equation (14) for β is cast as a convex optimization in Section 6.4.

6.4 Computational Costs

The exponential tilting solution to (1) is the value β for which $\int (x - \bar{x})e^{x'\beta} dF_0(x) = 0$. That solution is more conveniently expressed as the root of

$$g(\beta) \equiv \int (x - \bar{x})e^{(x-\bar{x})'\beta} dF_0(x) = 0. \tag{15}$$

Equation (15) is the gradient with respect to β of

$$f(\beta) = \int e^{(x-\bar{x})'\beta} dF_0(x),$$

which has Hessian

$$H(\beta) = \int (x - \bar{x})(x - \bar{x})' e^{(x-\bar{x})'\beta} dF_0(x).$$

The tilting problem (1) can be solved by finding the root of (15) which is in turn equivalent to the minimization of the convex function f.

When F_0 is modeled as a mixture F_0 of Gaussians the objective function, gradient, and Hessian needed for optimization have a simple form. They are,

$$f(\beta) = \sum_{k=1}^{K} \lambda_k e^{\beta'(\mu_k - \bar{x}) + \beta' \Sigma_k \beta/2},$$

$$g(\beta) = \sum_{k=1}^{K} \lambda_k (\widetilde{\mu}_k(\beta) - \bar{x}) e^{\beta'(\mu_k - \bar{x}) + \beta' \Sigma_k \beta/2}, \quad \text{where,}$$

$$\widetilde{\mu}_k(\beta) \equiv \mu_k + \Sigma_k \beta, \quad \text{and,}$$

$$H(\beta) = \sum_{k=1}^{K} \lambda_k \left(\Sigma_k + (\widetilde{\mu}_k(\beta) - \bar{x})(\widetilde{\mu}_k(\beta) - \bar{x})' \right) e^{\beta'(\mu_k - \bar{x}) + \beta' \Sigma_k \beta/2}.$$

The cost of solving (14) or (15) by an algorithm based on Newton's method takes $O(d^3)$ computation per iteration. By contrast, each step in iteratively reweighted least squares fitting of logistic regression takes $O((n+N)d^2)$ work. Even if one downsamples the data set, perhaps keeping only $N = 5n$ randomly chosen examples from the $Y = 0$ cases, the work of an iteration is $O(nd^2)$. The one time cost to fit a mixture of Gaussians includes costs of order Nd^2 to form covariance matrix estimates, or $O(nd^2)$ if one has downsampled. But after the first iteration there can be substantial computational savings for solving (15) instead of doing logistic regression, when n/d is large.

When there is one common class and there are numerous rare classes, such as types of fraud or different targets against which a drug might be active, then the cost of approximating F_0 can be shared over the set of uncommon classes.

In fraud detection problems we might expect that the distribution F_0 for legitimate data points is slowly changing while the patterns in the fraudulent points change rapidly in response to improved detection. In such a setting we get a computational saving by fitting an approximation to F_0 once, or at long time intervals, and then computing many different $\beta(\infty)$ vectors. These vectors can be for different known types of fraud, for fraud over shorter time intervals, or even individual fraud cases.

Acknowledgments

References

R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification And Regression Trees*. Wadsworth, Belmont, CA, 1984.

N.V. Chawla, N. Japkowicz, and A. Kolcz. *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets*. 2003.

N.V. Chawla, N. Japkowicz, and A. Kolcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

D.A. Cohn, Z. Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

N. Japkowicz. *Learning from Imbalanced Data Sets: Papers from the AAAI Workshop*. AAAI, 2000. Technical Report WS-00-05.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163, 2001.

M.J. Silvapulle. On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statistical Society, Series B*, 43:310–313, 1981.

S. Tong. *Active learning: Theory and applications*. PhD thesis, Stanford University, 2001. URL http://ai.stanford.edu/~stong/research.html/tong_thesis.pdf.

M. Zhu, W. Su, and H. A. Chipman. LAGO: A computationally efficient approach for statistical detection. *Technometrics*, 48:193–205, 2005.