Infinitely Imbalanced Logistic Regression

Art B. Owen

Stanford University
owen@stat.stanford.edu

12th annual winter workshop
January 2010
University of Florida

Imbalanced data

Setting:

- Data are (X, Y) pairs,

Imbalanced data

Setting:

- Data are (X, Y) pairs,
- Predictors $X \in \mathbb{R}^d$

Imbalanced data

Setting:

- Data are (X, Y) pairs,
- Predictors $X \in \mathbb{R}^d$
- Binary response variable $Y \in \{0, 1\}$

Imbalanced data

Setting:

- Data are (X, Y) pairs,
- Predictors $X \in \mathbb{R}^d$
- Binary response variable $Y \in \{0, 1\}$
- Sample has lots of $Y = 0$,

Imbalanced data

Setting:

- Data are (X, Y) pairs,
- Predictors $X \in \mathbb{R}^d$
- Binary response variable $Y \in \{0, 1\}$
- Sample has lots of $Y = 0$, very few $Y = 1$

Imbalanced data

Setting:

- Data are (X, Y) pairs,
- Predictors $X \in \mathbb{R}^d$
- Binary response variable $Y \in \{0, 1\}$
- Sample has lots of $Y = 0$, very few $Y = 1$

Examples, $Y = 1$ for:

- active drug
- ad gets clicked
- rare disease
- war/coup/veto
- citizen seeks elected office
- non-spam in spam bucket

(Why) does imbalance matter?

Irony:

$$\begin{aligned} 500 \text{ 1s and } \quad 500 \text{ 0s} &\implies \text{OK} \\ 500 \text{ 1s and } 500{,}000 \text{ 0s} &\implies \text{trouble} \end{aligned}$$

(Why) does imbalance matter?

Irony:

$$\begin{aligned} 500 \text{ 1s and } \quad 500 \text{ 0s} &\implies \text{OK} \\ 500 \text{ 1s and } 500{,}000 \text{ 0s} &\implies \text{trouble} \end{aligned}$$

Issues:

1. It is hard to beat the rule that predicts $Y = 0$ always
2. Few $Y = 1$ cases constitute a low effective sample size

(Why) does imbalance matter?

Irony:

500 1s and 500 0s $\implies$ OK
500 1s and $500{,}000$ 0s $\implies$ trouble

Issues:

1. It is hard to beat the rule that predicts $Y = 0$ always
2. Few $Y = 1$ cases constitute a low effective sample size

Approaches:

1. So take account of priors and/or loss asymmetry
   (assuming implicit/explicit probability estimates)
2. Effective sample size really is $\#$ of $Y = 1$s

How to deal with imbalanced data:

Coping strategies:

1. Downsample the 0s (adjust prior accordingly)
2. Upsample the 1s:
   - Repeat some (or upweight them)
   - Add synthetic 1s
3. One class prob.: find small ellipsoid holding the x_i for $y_i = 1$

How to deal with imbalanced data:

Coping strategies:

1. Downsample the 0s (adjust prior accordingly)
2. Upsample the 1s:
   - Repeat some (or upweight them)
   - Add synthetic 1s
3. One class prob.: find small ellipsoid holding the x_i for $y_i = 1$

Workshops on imbalanced data:

- AAAI 2000
- ICML 2003

They prefer “imbalanced” to “unbalanced”

Is it even a problem?

Suppose data are

For $y = 1$: $x_{1i}, \quad i = 1, \ldots, n_1 \equiv n$

For $y = 0$: $x_{0i}, \quad i = 1, \ldots, n_0 \equiv N \qquad N \gg n$

Is it even a problem?

Suppose data are

For $y = 1$: $\quad x_{1i}, \quad i = 1, \ldots, n_1 \equiv n$

For $y = 0$: $\quad x_{0i}, \quad i = 1, \ldots, n_0 \equiv N \qquad N \gg n$

Fit logistic regression

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}$$

Is it even a problem?

Suppose data are

For $y = 1$: $\quad x_{1i}, \quad i = 1, \dots, n_1 \equiv n$

For $y = 0$: $\quad x_{0i}, \quad i = 1, \dots, n_0 \equiv N \qquad N \gg n$

Fit logistic regression

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}$$

Let $N \to \infty$ with n fixed

Expect $\hat{\alpha} \to -\infty$ like $-\log(N)$

But $\hat{\beta}$ can have a useful limit

and $\hat{\beta}$ is of most interest

Is it even a problem?

Suppose data are

For $y = 1$: $x_{1i}, \quad i = 1, \dots, n_1 \equiv n$

For $y = 0$: $x_{0i}, \quad i = 1, \dots, n_0 \equiv N \qquad N \gg n$

Fit logistic regression

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + x'\beta}}{1 + e^{\alpha + x'\beta}}$$

Let $N \to \infty$ with n fixed

Expect $\hat{\alpha} \to -\infty$ like $-\log(N)$

But $\hat{\beta}$ can have a useful limit

and $\hat{\beta}$ is of most interest

$N/n \to \infty$ not necessarily so bad (for logistic regression).

Main result

Suppose

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0$$

Let $\alpha(N)$ and $\beta(N)$ be logistic regression estimates

Main result

Suppose

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0$$

Let $\alpha(N)$ and $\beta(N)$ be logistic regression estimates

Under mild conditions

$$Ne^{\alpha(N)} \to A \in \mathbb{R} \quad \text{and} \quad \beta(N) \to \beta \in \mathbb{R}^d$$

Main result

Suppose

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \in \mathbb{R}^d \quad \& \quad x \sim F_0 \quad \text{when} \quad Y = 0$$

Let $\alpha(N)$ and $\beta(N)$ be logistic regression estimates

Under mild conditions

$$N e^{\alpha(N)} \to A \in \mathbb{R} \quad \text{and} \quad \beta(N) \to \beta \in \mathbb{R}^d$$

where β solves

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

Interpretation

We have

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

β is the *exponential tilt* to take $E_{F_0}(X)$ onto $\bar{x}$

Interpretation

We have

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

β is the *exponential tilt* to take $E_{F_0}(X)$ onto $\bar{x}$

For $F_0 = N(\mu_0, \Sigma_0)$

$$\beta = \Sigma_0^{-1}(\bar{x} - \mu_0)$$

Interpretation

We have

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

β is the *exponential tilt* to take $E_{F_0}(X)$ onto $\bar{x}$

For $F_0 = N(\mu_0, \Sigma_0)$

$$\beta = \Sigma_0^{-1}(\bar{x} - \mu_0)$$

Compare

$\beta = \Sigma^{-1}(\mu_1 - \mu_0)$ for

$X \sim N(\mu_j, \Sigma)$ given $Y = j \in \{0, 1\}$

Surprise!

Suppose β solves

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

Then only $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and F_0 matter

Clearly n is the effective sample size

Surprise!

Suppose β solves

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

Then only $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and F_0 matter

Clearly n is the effective sample size

We could:

replace $(x_{1i}, 1)$ for $i = 1, \dots, n$
by just one point $(X, Y) = (\bar{x}, 1)$
and get the same β as $N \to \infty$

Surprise!

Suppose β solves

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

Then only $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and F_0 matter

Clearly n is the effective sample size

We could:

replace $(x_{1i}, 1)$ for $i = 1, \dots, n$
by just one point $(X, Y) = (\bar{x}, 1)$
and get the same β as $N \to \infty$

Upshot:

IILR downsamples the rare case to a single point
Whether logistic works well or badly on given problem
Other classifiers (e.g. CART) would be different

Uses

The predictions are trivial

$$\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d$$

Uses

The predictions are trivial

$$\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d$$

But ratios are informative and simple

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

Uses

The predictions are trivial

$$\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d$$

But ratios are informative and simple

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

For fraud or active learning, obtain Y corresponding to largest

- $e^{x'\beta}$ (best chance to see a 1)

Uses

The predictions are trivial

$$\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d$$

But ratios are informative and simple

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

For fraud or active learning, obtain Y corresponding to largest

- $e^{x'\beta}$ (best chance to see a 1)
- $v\, e^{x'\beta}$ (when case has value v)

Uses

The predictions are trivial

$$\Pr(Y = 1 \mid X = x) \to 0 \quad \text{for all} \quad x \in \mathbb{R}^d$$

But ratios are informative and simple

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

For fraud or active learning, obtain Y corresponding to largest

- $e^{x'\beta}$ (best chance to see a 1)
- $v\, e^{x'\beta}$ (when case has value v)
- $v\, e^{x'\beta}/c$ (and investigative cost c)

Logistic regression

Log likelihood (with $x_i \equiv x_{1i}$)

$$\sum_{i=1}^{n} \left\{ \alpha + x_i'\beta - \log(1 + e^{\alpha + x_i'\beta}) \right\} - \sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i}'\beta}) \right\}$$

Logistic regression

Log likelihood (with $x_i \equiv x_{1i}$)

$$\sum_{i=1}^{n} \left\{ \alpha + x_i'\beta - \log(1 + e^{\alpha + x_i'\beta}) \right\} - \sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i}'\beta}) \right\}$$

For large N

$$\sum_{i=1}^{N} \left\{ \log(1 + e^{\alpha + x_{0i}'\beta}) \right\} \approx N \int \log(1 + e^{\alpha + x'\beta}) \, dF_0(x)$$

Centering data

With foresight, center data at $\bar{x}$

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x - \bar{x})'\beta}}{1 + e^{\alpha + (x - \bar{x})'\beta}}$$

Centering data

With foresight, center data at $\bar{x}$

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x-\bar{x})'\beta}}{1 + e^{\alpha + (x-\bar{x})'\beta}}$$

Centered log likelihood $\ell(\alpha, \beta)$

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

Centering data

With foresight, center data at $\bar{x}$

$$\Pr(Y = 1 \mid X = x) = \frac{e^{\alpha + (x-\bar{x})'\beta}}{1 + e^{\alpha + (x-\bar{x})'\beta}}$$

Centered log likelihood $\ell(\alpha, \beta)$

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

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

Sketch of the proof

Set $\frac{1}{N}\frac{\partial}{\partial\beta}\ell(\alpha,\beta)=0$

$$0=-\frac{1}{N}\sum_{i=1}^{n}\frac{(x_i-\bar{x})\,e^{\alpha+(x_i-\bar{x})'\beta}}{1+e^{\alpha+(x_i-\bar{x})'\beta}}-\int\frac{(x-\bar{x})\,e^{\alpha+(x-\bar{x})'\beta}}{1+e^{\alpha+(x-\bar{x})'\beta}}\,dF_0(x)$$

Sketch of the proof

Set $\frac{1}{N}\frac{\partial}{\partial\beta}\ell(\alpha,\beta) = 0$

$$0 = -\frac{1}{N}\sum_{i=1}^{n}\frac{(x_i-\bar{x})\,e^{\alpha+(x_i-\bar{x})'\beta}}{1+e^{\alpha+(x_i-\bar{x})'\beta}} - \int \frac{(x-\bar{x})\,e^{\alpha+(x-\bar{x})'\beta}}{1+e^{\alpha+(x-\bar{x})'\beta}}\,dF_0(x)$$

$N \to \infty$, so ignore the first sum:

$$0 = \int \frac{(x-\bar{x})\,e^{\alpha+(x-\bar{x})'\beta}}{1+e^{\alpha+(x-\bar{x})'\beta}}\,dF_0(x)$$

Sketch of the proof

Set $\frac{1}{N}\frac{\partial}{\partial\beta}\ell(\alpha,\beta) = 0$

$$0 = -\frac{1}{N}\sum_{i=1}^{n}\frac{(x_i - \bar{x})\, e^{\alpha+(x_i-\bar{x})'\beta}}{1+e^{\alpha+(x_i-\bar{x})'\beta}} - \int \frac{(x-\bar{x})\, e^{\alpha+(x-\bar{x})'\beta}}{1+e^{\alpha+(x-\bar{x})'\beta}}\, dF_0(x)$$

$N \to \infty$, so ignore the first sum:

$$0 = \int \frac{(x-\bar{x})\, e^{\alpha+(x-\bar{x})'\beta}}{1+e^{\alpha+(x-\bar{x})'\beta}}\, dF_0(x)$$

If $\alpha \to -\infty$, denominator $\to 1$, and so β solves:

$$\int (x-\bar{x})\, e^{\alpha+(x-\bar{x})'\beta}\, dF_0(x) = 0 \quad \square$$

Example: $F_0 = N(0,1)$, $\bar{x} = 1$, $n = 1$, $N \to \infty$

Common values:

$x_{0i} \sim N(0,1)$

Rare value

$n = 1$

$x_{11} = 1$

Example: $F_0 = N(0,1)$, $\bar{x} = 1$, $n = 1$, $N \to \infty$

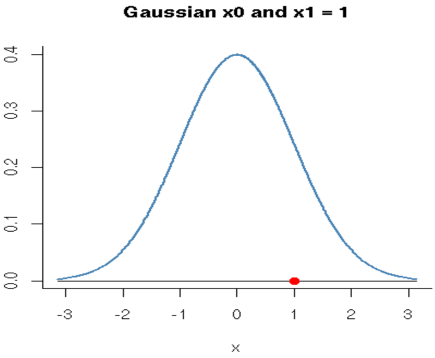
Common values:

$x_{0i} \sim N(0,1)$

Rare value

$n = 1$

$x_{11} = 1$



Example: $F_0 = N(0,1)$, $\bar{x} = 1$, $n = 1$, $N \to \infty$

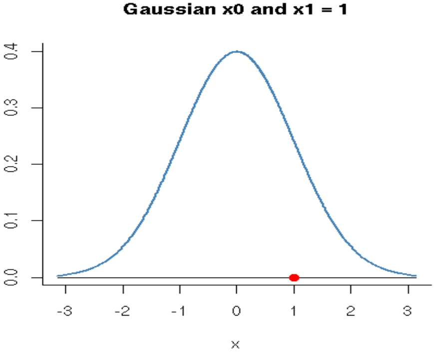
Common values:

$x_{0i} \sim N(0,1)$

Rare value

$n = 1$

$x_{11} = 1$



We should see $\beta \to \Sigma_0^{-1}(\bar{x} - \mu_0) = 1^{-1}(1 - 0) = 1$

Example: $F_0 = N(0,1)$, $\bar{x} = 1$, $n = 1$, $N \to \infty$

For $Y = 0$ and $i = 1, \ldots, N$ take

$$x_{0i} = \Phi^{-1}\left(\frac{i - 1/2}{N}\right)$$

We should see $\beta \to \Sigma_0^{-1}(\bar{x} - \mu_0) = 1^{-1}(1 - 0) = 1$

Logistic regression results

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
∞			1

Next: two counterexamples

We will need conditions for the exponential tilting to work.
One counterexample has a Cauchy distribution.
The other a uniform.

Example: now $F_0 =$ Cauchy

$$f_0(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

$$x_{0i} = F_0^{-1}\left(\frac{i-1/2}{N}\right), \quad i = 1, \ldots, N$$

$$x_{1i} = 1, \quad i = 1 \quad \text{only}$$

Example: now $F_0 =$ Cauchy

$$f_0(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

$$x_{0i} = F_0^{-1}\left(\frac{i-1/2}{N}\right), \quad i = 1,\ldots,N$$

$$x_{1i} = 1, \quad i = 1 \quad \text{only}$$

Logistic regression results

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

Example: now F_0 = Cauchy

$$f_0(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

$$x_{0i} = F_0^{-1}\Big(\frac{i-1/2}{N}\Big), \quad i = 1, \ldots, N$$

$$x_{1i} = 1, \quad i = 1 \quad \text{only}$$

Logistic regression results

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

$\beta(N) \to 0$ Cauchy has no mean to tilt onto $\bar{x}$!

Example: now $F_0 = U[0, 1]$ and $n_1 = 2$

Common values:

$x_{0i} \sim U(0, 1)$

Rare values:

$n = 2$

$x_{11} = 0.5$

$x_{12} = 2.0$

Example: now $F_0 = U[0,1]$ and $n_1 = 2$

Common values:

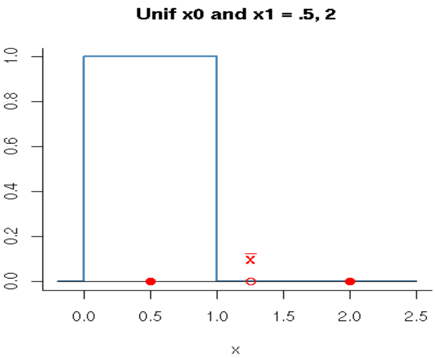
$x_{0i} \sim U(0,1)$

Rare values:

$n = 2$

$x_{11} = 0.5$

$x_{12} = 2.0$



Example: now $F_0 = U[0,1]$ and $n_1 = 2$

Common values:

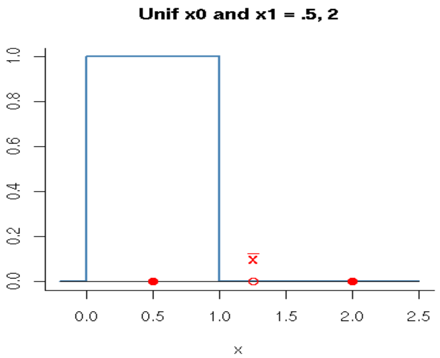
$x_{0i} \sim U(0,1)$

Rare values:

$n = 2$

$x_{11} = 0.5$

$x_{12} = 2.0$



We can't tilt $U(0,1)$ to have mean $\bar{x} = 1.25$

Example: now $F_0 = U[0,1]$ and $n_1 = 2$

$$x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N$$

$$x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}$$

Example: now $F_0 = U[0,1]$ and $n_1 = 2$

$$x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N$$

$$x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}$$

Logistic regression results

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

Example: now $F_0 = U[0,1]$ and $n_1 = 2$

$$x_{0i} = \frac{i - 1/2}{N}, \quad i = 1, \ldots, N$$

$$x_{11} = \frac{1}{2}, \quad x_{12} = 2 \quad \text{only}$$

Logistic regression results

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

$\beta(N) \to \infty$ also $\bar{x} = \frac{5}{4} \notin [0,1]$ (can't tilt mean so far)

We need conditions:

Tail of F_0 not too heavy

$\int \|x\| e^{x'\beta} \, dF_0(x) < \infty$

to fix problem from Cauchy example

tail weight not an issue in finite samples

We need conditions:

Tail of F_0 not too heavy

$\int \|x\| e^{x'\beta}\, dF_0(x) < \infty$

to fix problem from Cauchy example

tail weight not an issue in finite samples

Overlap between F_0 and $\bar{x}$

to fix problem from $U(0,1)$ example

overlap is an issue in finite samples

but we need stronger overlap condition

Overlap conditions

F has $x^* \in \mathbb{R}^d$ surrounded if

- For all unit vectors $\theta \in \mathbb{R}^d$
- $\Pr\big((x - x^*)'\theta > \epsilon \mid x \sim F_0\big) > \delta$
- for some $\epsilon > 0$ and $\delta > 0$

Overlap conditions

F has $x^* \in \mathbb{R}^d$ surrounded if

- For all unit vectors $\theta \in \mathbb{R}^d$
- $\Pr\big((x - x^*)'\theta > \epsilon \mid x \sim F_0\big) > \delta$
- for some $\epsilon > 0$ and $\delta > 0$

For $N \to \infty$ we need:

- F_0 to have $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$ surrounded

Overlap conditions

F has $x^* \in \mathbb{R}^d$ surrounded if

- For all unit vectors $\theta \in \mathbb{R}^d$
- $\Pr\big((x - x^*)'\theta > \epsilon \mid x \sim F_0\big) > \delta$
- for some $\epsilon > 0$ and $\delta > 0$

For $N \to \infty$ we need:

- F_0 to have $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$ surrounded

For finite samples, Silvapulle (1981, JRSS-B)

- If model has intercept and x's are full rank
- We need some x_0 surrounded by both $\hat{F}_1$ and $\hat{F}_0$

Theorem

Let $n \geq 1$ *and* $x_1, \ldots, x_n \in \mathbb{R}^d$ *be fixed. Suppose that*

1. F_0 *surrounds* $\bar{x} = \sum_{i=1}^{n} x_i/n$
2. $\int \|x\| e^{x'\beta} \, dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d$

Theorem

Let $n \geq 1$ *and* $x_1, \ldots, x_n \in \mathbb{R}^d$ *be fixed. Suppose that*

1. F_0 *surrounds* $\bar{x} = \sum_{i=1}^{n} x_i/n$
2. $\int \|x\| e^{x'\beta}\, dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d$

Then the maximizer $(\hat{\alpha}, \hat{\beta})$ *of* ℓ *satisfies*

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

Theorem

Let $n \geq 1$ *and* $x_1, \ldots, x_n \in \mathbb{R}^d$ *be fixed. Suppose that*

1. F_0 *surrounds* $\bar{x} = \sum_{i=1}^n x_i / n$
2. $\int \|x\| e^{x'\beta} \, dF_0(x) < \infty \quad \forall \beta \in \mathbb{R}^d$

Then the maximizer $(\hat{\alpha}, \hat{\beta})$ *of* ℓ *satisfies*

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

Steps

1. show $\alpha(N)$ and $\beta(N)$ exist for each N
2. show $Ne^{\hat{\alpha}(N)}$ is bounded
3. show $\|\hat{\beta}\|$ is bounded
4. then take partial derivatives as before

Computation

Given an approximation to F_0:

Solve	$0 = \int (x - \bar{x}) e^{x'\beta} \, dF_0(x)$	d equations
Same as	$0 = g(\beta) \equiv \int (x - \bar{x}) e^{(x-\bar{x})'\beta} \, dF_0(x)$	
I.E. Minimize	$f(\beta) = \int e^{(x-\bar{x})'\beta} \, dF_0(x)$	
Hessian is	$H(\beta) = \int (x - \bar{x})(x - \bar{x})' e^{(x-\bar{x})'\beta} \, dF_0(x)$	convex

Newton step

$$\beta \leftarrow \beta - H^{-1} g$$

Cost per iteration: $O(d^3)$ vs $O(Nd^2)$ or $O(nd^2)$.

Mixture of Gaussians

$$F_0 = \sum_{k=1}^{K} \lambda_k N(\mu_k, \Sigma_k) \qquad \lambda_k > 0 \quad \sum_k \lambda_k = 1$$

Tilt a Gaussian, get a Gaussian:

$$e^{(x-\bar{x})'\beta}\, e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} = e^{(\mu-\bar{x})'\beta}\, e^{-\frac{1}{2}(x-\mu-\Sigma\beta)'\Sigma^{-1}(x-\mu-\Sigma\beta)}$$

Newton step is

$$\beta \leftarrow \beta - H^{-1}g$$

$$g = \sum_{k=1}^{K} \lambda_k e^{(\mu_k-\bar{x})'\beta}\Bigl(\widetilde{\mu}_k - \bar{x}\Bigr), \qquad \widetilde{\mu}_k = \mu_k + \Sigma_k\beta$$

$$H = \sum_{k=1}^{K} \lambda_k e^{(\mu_k-\bar{x})'\beta}\Bigl(\Sigma_k + (\bar{x}-\widetilde{\mu}_k)(\bar{x}-\widetilde{\mu}_k)'\Bigr)$$

Drug discovery example

Zhu, Su, Chipman

Technometrics, 2005

$Y = 1$ for active drug

$Y = 0$ for inactive drug

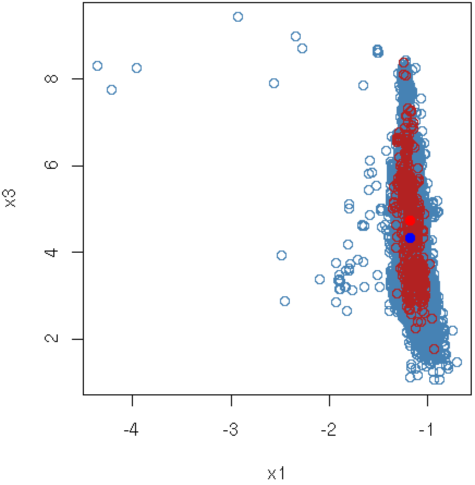
$d = 6$ features

$29{,}821$ chemicals

only 608 active $\approx 2\%$

x_1 x_3 strongest

Group means plotted



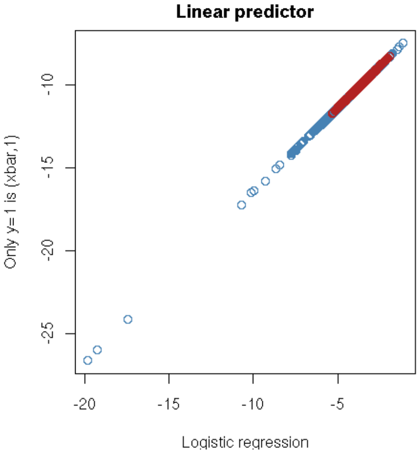
Drug discovery example ctd

Fits

Plain logistic
(608 ones), vs
1 one at $\bar{x}_1$

Upshot

Same ordering, ROC
precision-recall
etc.

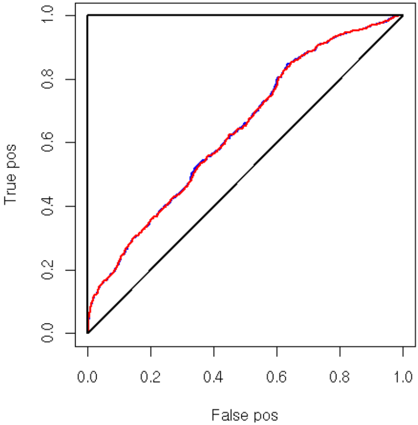


Drug discovery example ctd

ROC curves

Plain logistic

1 one at $\bar{x}_1$



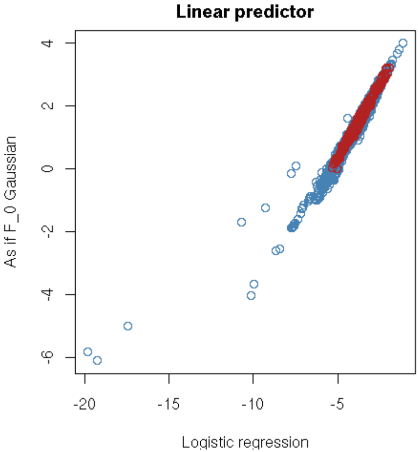
Drug discovery example ctd

Fits

Plain logistic, vs,
Pretend F_0 is Gaussian
And use $\bar{x}_1$

Upshot

Slight difference
For easy 0s
Mixture model might improve



The drug data was not a typical example

Drug data had

- very bad separation
- Poor ROC
- $\bar{x}$ very surrounded

The drug data was not a typical example

Drug data had

very bad separation

Poor ROC

$\bar{x}$ very surrounded

Artificial version

$x_{1i} \leftarrow x_{1i} + \delta$

$\delta = (s/10, \dots, s/10)$

$s = 0, \dots, 10$

Original ROCs in blue

Lumped in red

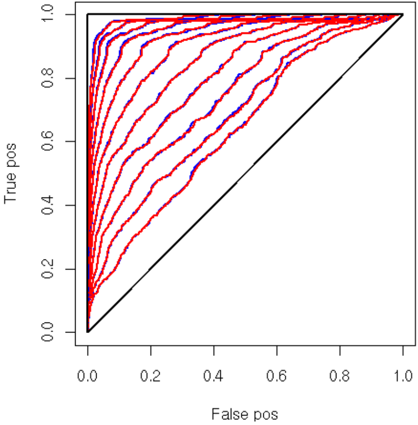
The drug data was not a typical example

Drug data had

very bad separation
Poor ROC
$\bar{x}$ very surrounded

Artificial version

$x_{1i} \leftarrow x_{1i} + \delta$
$\delta = (s/10, \ldots, s/10)$
$s = 0, \ldots, 10$
Original ROCs in blue
Lumped in red



The drug data was not a typical example

Drug data had

very bad separation

Poor ROC

$\bar{x}$ very surrounded

Artificial version

$x_{1i} \leftarrow x_{1i} + \delta$

$\delta = (s/10, \ldots, s/10)$

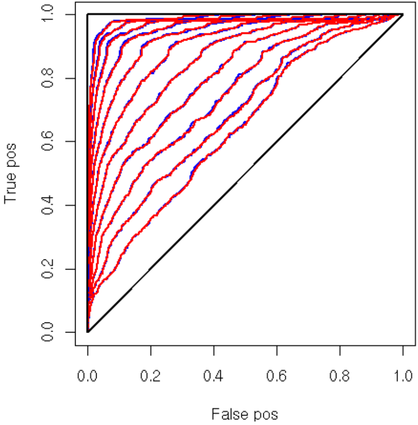
$s = 0, \ldots, 10$

Original ROCs in blue

Lumped in red

Upshot

Still only uses $\bar{x}$



Thoughts for fraud detection

Non fraud data, $Y = 0$

Change slowly over time
Large sample size
So build a rich model for F_0
Update rarely

Thoughts for fraud detection

Non fraud data, $Y = 0$

Change slowly over time
Large sample size
So build a rich model for F_0
Update rarely

Fraud data, $Y = 1$

May change rapidly in response to detection
May have different flavors
Clusters appear, disappear, move, change size
Rapidly refit model using per cluster $\bar{x}$

Acknowledgments

- Paul Louisell for comments
- NSF for funds
- Host: University of Florida
- Organizers: Agresti, Young, Daniels, Casella
- Travel help: Robyn Crawford