Logistic Model

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

Outline

Odds Ratio

Definition

Let π be the probability of an event occurring. The odds ratio $\mathcal{R}$ for the event is the ratio of the probability of the event occurring to the probability of the event not occurring

$$\mathcal{R} = \frac{\pi}{1 - \pi}$$

$$\pi = \frac{\mathcal{R}}{1 + \mathcal{R}}$$

Logit

Definition

The Logit function is the natural log of the odds ratio.

$$Logit(\mathcal{R}) \quad = \quad \log(\mathcal{R}) = \log(\frac{\pi}{1-\pi})$$

Logistic Linear Model

- x measurement vector
- θ parameter vector
- θ_0 parameter scalar
- c^1 event that true class of measurement vector is c^1

$$\begin{aligned}
\log(\mathcal{R}(x;\theta,\theta_0)) &= \log\left(\frac{P(c^1 \mid x;\theta,\theta_0)}{1 - P(c^1 \mid x;\theta,\theta_0)}\right) \\
&= \theta_0 + \theta' x
\end{aligned}$$

Logistic Linear Model

Two classes: c^1 and c^2

$$
\begin{aligned}
\log\left(\frac{P(c^1 \mid x;\theta,\theta_0)}{1-P(c^1 \mid x;\theta,\theta_0)}\right) &= \theta_0 + \theta' x \\
\frac{P(c^1 \mid x;\theta,\theta_0)}{1-P(c^1 \mid x;\theta,\theta_0)}) &= e^{\theta_0+\theta' x} \\
P(c^1 \mid x;\theta,\theta_0) &= [1-P(c^1 \mid x;\theta,\theta_0)]e^{\theta_0+\theta' x} \\
P(c^1 \mid x;\theta,\theta_0) &= \frac{e^{\theta_0+\theta' x}}{1+e^{\theta_0+\theta' x}} \\
P(c^2 \mid x;\theta,\theta_0) &= \frac{1}{1+e^{\theta_0+\theta' x}}
\end{aligned}
$$

Logistic Linear Model

Given the parameter vector θ, θ_0 and a measurement vector x,

$$\frac{e^{\theta_0+\theta' x}}{1+e^{\theta_0+\theta' x}}$$

produces the conditional probability that the true class is c^1 given that the measurement vector is x.

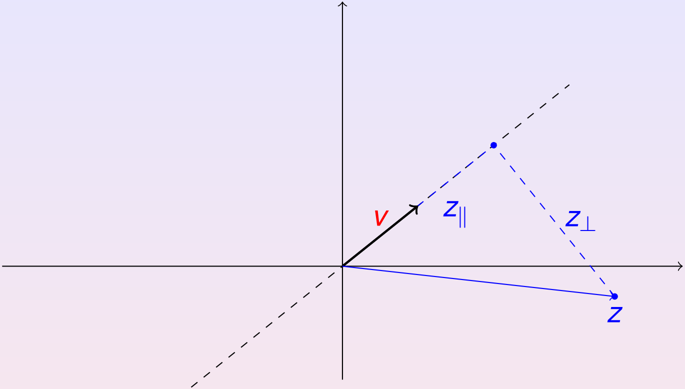
Expected Values

Let y be an indicator variable.

- $y = 1$ indicates class c^1
- $y = 0$ indicates class c^2

$$\begin{aligned} E[y \mid x; \theta, \alpha] &= 1P(y = 1 \mid x; \theta, \alpha) + 0P(y = 0 \mid x; \theta, \alpha) \\ &= P(y = 1 \mid x; \theta, \alpha) \\ &= \frac{e^{\alpha + \theta' x}}{1 + e^{\alpha + \theta' x}} \end{aligned}$$

Projection



If $||v|| = 1$, then $v'z$ is the signed length of the orthogonal projection of z onto v.

$$\begin{aligned} v'z = v'(z_{||} + z_{\perp}) &= v'z_{||} + v'z_{\perp} = v'z_{||} = v'(\pm||z_{||}||v) = \pm||z_{||}|| \\ |v'z| &= ||z_{||}|| \end{aligned}$$

Scaling x

Without loss of generality, we can always scale x so that the θ associated with the scaled x has norm 1.

$$\begin{aligned}\log\left(\frac{P(c^1 \mid x;\theta,\theta_0)}{1-P(c^1 \mid x;\theta,\theta_0)}\right) &= \theta_0+\theta' x \\ &= \theta_0+\frac{\theta'}{||\theta||}(x||\theta||)\end{aligned}$$

For convenience we scale x and normalize θ so that

$$\begin{aligned} x_{new} &= x||\theta|| \\ \theta_{new} &= \frac{\theta}{||\theta||}\end{aligned}$$

Changes in x

$$
\begin{aligned}
\mathcal{R}(x;\theta,\theta_0) &= e^{\theta_0+\theta' x} \\
\mathcal{R}(x+\delta;\theta,\theta_0) &= e^{\theta_0+\theta'(x+\delta)} \\
&= e^{\theta_0+\theta' x} e^{\theta'\delta} \\
&= \mathcal{R}(x) e^{\theta'\delta}
\end{aligned}
$$

Odds ratio is multiplied by $e^{\theta'\delta}$.

Changes in x

Let δ be a change in x. Define $\delta_{||}$ and $\delta_\perp$ so that

- $\delta = \delta_{||} + \delta_\perp$
- $\theta' \delta_\perp = 0$
- $|\theta' \delta_{||}| = ||\delta_{||}||$

$$
\begin{aligned}
\mathcal{R}(x;\theta,\theta_0) &= e^{\theta_0+\theta' x} \\
\mathcal{R}(x+\delta;\theta,\theta_0) &= e^{\theta_0+\theta'(x+\delta)} = e^{\theta_0+\theta'(x+\delta_{||}+\delta_\perp)} \\
&= e^{\theta_0+\theta' x} e^{\theta'(\delta_{||}+\delta_\perp)} \\
&= \mathcal{R}(x;\theta,\theta_0) e^{\theta'\delta_{||}} e^{\theta'\delta_\perp} \\
&= \mathcal{R}(x;\theta,\theta_0) e^{\theta'\delta_{||}}
\end{aligned}
$$

Odds ratio is multiplied by $e^{\theta'\delta_{||}}$.

Changes in x

$$\begin{aligned}
\mathcal{R}(x+\delta;\theta,\theta_0) &= \mathcal{R}(x;\theta,\theta_0)e^{\theta^{'}\delta_{\|}} \\
\log(\mathcal{R}(x+\delta;\theta,\theta_0)) &= \log(\mathcal{R}(x;\theta)) + \log(e^{\theta^{'}\delta_{\|}}) \\
&= \log(\mathcal{R}(x;\theta,\theta_0)) + \theta^{'}\delta_{\|}
\end{aligned}$$

Log of odds ratio increases by $\theta^{'}\delta_{\|}$.

Changes in Odds Ratio

Suppose the new odds ratio is multiplied by λ as a result in the change of x. Then, what happens to the probability of the event?

$$\begin{aligned}
\mathcal{R}_{new} &= \mathcal{R}\lambda \\
\frac{\pi_{new}}{1-\pi_{new}} &= \frac{\pi}{1-\pi}\lambda \\
\pi_{new} &= \frac{\pi\lambda}{1-\pi}(1-\pi_{new}) \\
\pi_{new}\frac{1-\pi+\pi\lambda}{1-\pi} &= \frac{\pi\lambda}{1-\pi} \\
\pi_{new} &= \frac{\pi\lambda}{1-\pi+\pi\lambda}
\end{aligned}$$

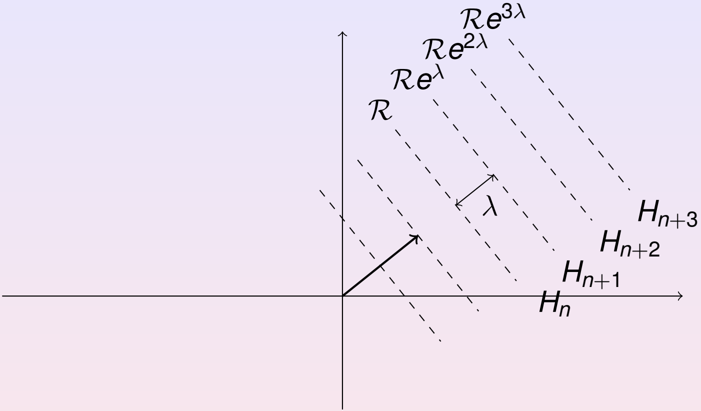
Changes in x

Suppose that $\mathcal{R} = 9$. Then $\pi = \frac{\mathcal{R}}{1+\mathcal{R}} = \frac{9}{1+9} = .9$

Suppose that $e^{\theta'\delta} = 3$.

$$
\begin{aligned}
\pi_{new} &= \frac{.9(3)}{1 - .9 + .9(3)} \\
&= \frac{2.7}{2.8} \\
&= .9642857 \\
\mathcal{R}_{new} &= \frac{.9642857}{1 - .9642857} = \frac{.9642857}{.0357143} = 27
\end{aligned}
$$

Odds Ratio Iso-Contours



$$H_n = \{x \mid \theta' x = n\lambda\}$$

The Isocontours are hyperplanes

Probability Properties

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}$$

$$P(c^1 \mid x) = \frac{e^{\theta_0 + \theta' x}}{1 + e^{\theta_0 + \theta' x}}$$

Fix $x_1, x_2, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N$; Vary x_n

If $\theta_n > 0$ as $x_n \to \infty$, $\theta' x \to \infty$
If $\theta_n < 0$ as $x_n \to \infty$, $\theta' x \to -\infty$

Probability Form

Consider 1D case.

$$
\begin{aligned}
P(c^1 \mid x) &= \frac{e^{\theta_0+\theta x}}{1+e^{\theta_0+\theta x}} \\
&= \frac{e^{\theta(x+\theta_0/\theta)}}{1+e^{\theta(x+\theta_0/\theta)}}
\end{aligned}
$$

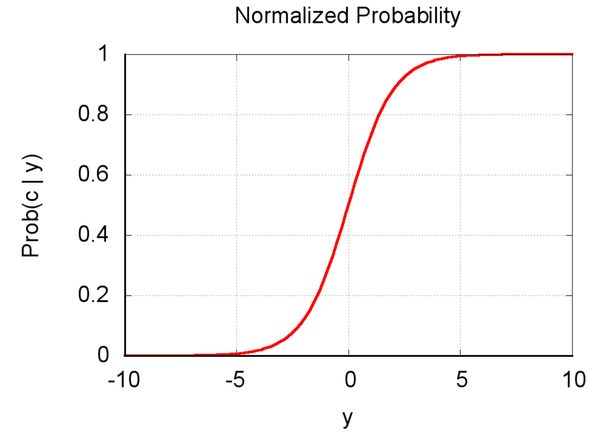
Let

$$y = \theta(x + \theta_0/\theta)$$

$$P(c^1 \mid y) = \frac{e^y}{1+e^y}$$

Normalized Probability

$$P(c^1 \mid y) = \frac{e^y}{1 + e^y}$$



Odd Functions

Fix a point x_0. Look at the values of f at points x_0 plus and minus x: $f(x_0 + x)$ and $f(x_0 - x)$. If the differences $f(x_0 + x) - f(x_0)$ and $f(x_0) - f(x_0 - x)$ are the same for all x, then function f is said to be odd about $(x_0, f(x_0))$.

Definition

A function $f : R \to R$ is called an odd function around $(x_0, f(x_0)$ if and only if for all x,

$$f(x_0 + x) - f(x_0) = f(x_0) - f(x_0 - x)$$

$$P(c^1 \mid y) = \frac{e^y}{1 + e^y}$$

$$\frac{e^y}{1 + e^y} \text{ is odd around } (0, 1/2)$$

Odd Function

$$P(c^1 \mid y) = \frac{e^y}{1+e^y}$$

Compare $f(y-0) - f(0)$ to $f(0) - f(0-y)$

$$\begin{aligned}
f(y-0) - f(0) &= \frac{e^y}{1+e^y} - \frac{1}{2} = \frac{e^y - 1}{2(1+e^y)} \\
f(0) - f(0-y) &= \frac{1}{2} - \frac{e^{-y}}{1+e^{-y}} = \frac{1-e^{-y}}{2(1+e^{-y})} = \frac{e^y - 1}{2(e^y+1)} \\
\Rightarrow f(y-0) - f(0) &= f(0) - f(0-y)
\end{aligned}$$

Normalized Probability Derivative

$$P(c^1 \mid y) = \frac{e^y}{1 + e^y}$$

$$\begin{aligned}
\frac{\partial}{\partial y} \frac{e^y}{1 + e^y} &= \frac{(1 + e^y)e^y - e^y e^y}{(1 + e^y)^2} \\
\frac{\partial}{\partial y} \frac{e^y}{1 + e^y}\Big|_{y=0} &= \frac{2 - 1}{2^2} \\
&= \frac{1}{4}
\end{aligned}$$

Derivative

$$y = \theta(x + \theta_0/\theta)$$

$$\begin{aligned} \frac{\partial}{\partial x}\frac{e^y}{1+e^y} &= \frac{\partial}{\partial y}\frac{e^y}{1+e^y}\frac{\partial y}{\partial x} \\ &= \frac{1}{4}\theta \end{aligned}$$

Translate

It is possible to translate x so that it incorporates the constant term θ_0

$$\theta_0 + \theta^{'} x \quad = \quad \theta^{'} \left(x + \frac{\theta}{||\theta||} \frac{\theta_0}{||\theta||} \right)$$

If

$$x_{new} = x + \frac{\theta}{||\theta||} \frac{\theta_0}{||\theta||}$$

Then

$$\theta_0 + \theta^{'} x = \theta^{'} x_{new}$$

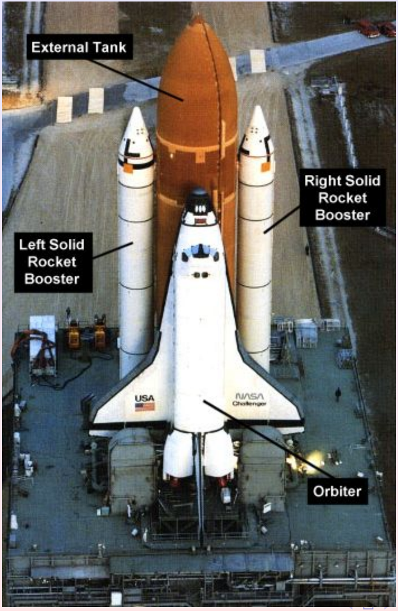
The combination of first translating and then scaling x means that without loss of generality, we can examine the properties of the logistic model assuming that $||\theta|| = 1$ and use the simpler form $\theta^{'} x$ in place of the original form $\theta_0 + \theta^{'} x$.

Derivative

N-Dimensional Case

$$
\begin{aligned}
P(c^1 \mid x) &= \frac{e^{\theta' x}}{1 + e^{\theta' x}} \\
\frac{\partial}{\partial x} P(c^1 \mid x) &= \frac{(1 + e^{\theta' x})\theta e^{\theta' x} - e^{\theta' x}\theta e^{\theta' x}}{(1 + e^{\theta' x})^2} \\
&= \frac{\theta e^{\theta' x}}{(1 + e^{\theta' x})^2} \\
\frac{\partial}{\partial x} P(c^1 \mid x)\Big|_{x=0} &= \frac{1}{4}\theta \\
&= \frac{1}{4}\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}
\end{aligned}
$$

Space Shuttle Challenger



Challenger Space Shuttle: The Cold Snap

- Evening of January 27 through January 28, 1986
- Florida experienced a statewide cold snap
  - The average low is around 50° F
  - The average high is around 72°F
- The January 28 temperature at the launch pad was 31°F

Morton Thiokol

- Cold weather prompted a teleconference between NASA and Morton Thiokol
- The engineers recommended not to launch
- The managers decided to go ahead with the launch

Approval For Challenger Mission

MTI ASSESSMENT OF TEMPERATURE CONCERN ON SRM-25 (51L) LAUNCH

- 0 CALCULATIONS SHOW THAT SRM-25 O-RINGS WILL BE 20° COLDER THAN SRM-15 O-RINGS
- 0 TEMPERATURE DATA NOT CONCLUSIVE ON PREDICTING PRIMARY O-RING BLOW-BY
- 0 ENGINEERING ASSESSMENT IS THAT:
  - 0 COLDER O-RINGS WILL HAVE INCREASED EFFECTIVE DUROMETER ("HARDER")
  - 0 "HARDER" O-RINGS WILL TAKE LONGER TO "SEAT"
    - 0 MORE GAS MAY PASS PRIMARY O-RING BEFORE THE PRIMARY SEAL SEATS (RELATIVE TO SRM-15)
      - 0 DEMONSTRATED SEALING THRESHOLD IS 3 TIMES GREATER THAN 0.038" EROSION EXPERIENCED ON SRM-15
  - 0 IF THE PRIMARY SEAL DOES NOT SEAT, THE SECONDARY SEAL WILL SEAT
    - 0 PRESSURE WILL GET TO SECONDARY SEAL BEFORE THE METAL PARTS ROTATE
      - 0 O-RING PRESSURE LEAK CHECK PLACES SECONDARY SEAL IN OUTBOARD POSITION WHICH MINIMIZES SEALING TIME
- 0 MTI RECOMMENDS STS-51L LAUNCH PROCEED ON 28 JANUARY 1986
  - 0 SRM-25 WILL NOT BE SIGNIFICANTLY DIFFERENT FROM SRM-15

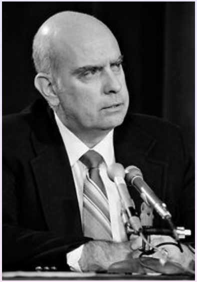
Joe C. Kilminster

JOE C. KILMINSTER, VICE PRESIDENT
SPACE BOOSTER PROGRAMS

Space Shuttle Challenger

January 28, 1986

The Whistle Blower



Roger Boisjoly

- Roger Boisjoly was one of the Morton Thiokol engineers
- Became the outspoken whistleblower
- Six months before he wrote a memo that there would be a failure of the seals if the weather was cold
- Testified to the Presidential Commission
- Gave the Presidential Commission internal Morton Thiokol documents

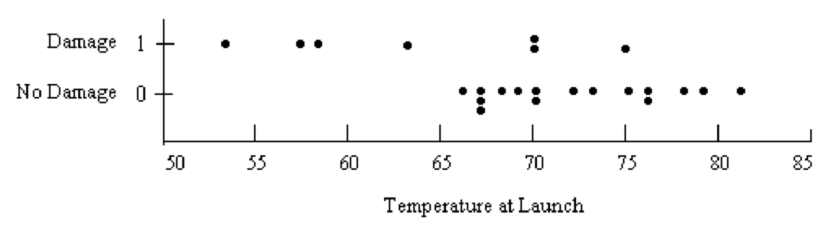
Space Shuttle Challenger Disaster Data

#F Number of O-ring Failures
T Outside Air Temperature

#F	T	#F	T	#F	T
0	66	1	70	0	69
0	68	0	67	0	72
0	73	0	70	1	57
1	63	1	70	0	78
0	67	2	53	0	67
0	75	0	70	0	81
0	76	0	79	2	75
0	76	1	58		

Space Shuttle Challenger Disaster

Damage = 1 or more O-ring failures.



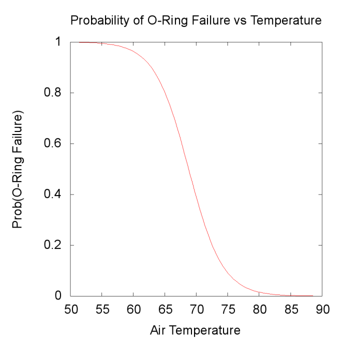
Temperature at the Launch pad of January 28, 1986 was 31 Degrees Farenheit.

Space Shuttle Challenger Disaster

Use the data to estimate the parameters of a logistic regression

$$\log(\mathcal{R}(T)) = 25.386 - .369T$$

$$P(O-RingFailure \mid x) = \frac{e^{25.386-.369T}}{1+e^{25.386-.369T}}$$



Economic Gain Matrix

			ASSIGNED	
			c^1 *Fail*	c^2 *Success*
T	c^1 Fail	$P_T(c^1 \mid d)$	$e(c^1, c^1)$	$e(c^1, c^2)$
R				
U				
E	c^2 Success	$P_T(c^2 \mid d)$	$e(c^2, c^1)$	$c(c^2, c^2)$

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j | d)$$

Economic Gain Matrix

		ASSIGNED *Fail*	ASSIGNED *Success*
TRUE	Fail	1	-100
	Success	1	2

Assign to class c^1 if

$$\begin{aligned}
\log(\mathcal{R}(d)) &\geq \frac{e(c^2,c^2) - e(c^2,c^1)}{e(c^1,c^1) - e(c^1,c^2)} \\
&\geq \frac{2-1}{1-(-100)} = \frac{1}{101}
\end{aligned}$$

Decision Rule

Assign to class Fail if

$$\begin{aligned} \log(\mathcal{R}(T)) &= 25.386 - .369T \geq \frac{1}{101} = .0099 \\ T &\leq 68.77 \end{aligned}$$

On the day of the launch, January 28, 1986, $T = 31^\circ$ Farenheit.

$$\begin{aligned} \log(\mathcal{R}(31)) &= 25.386 - .369 * 31 \\ &= 13.947 \\ P(Fail|31) &= \frac{e^{13.947}}{1 + e^{13.947}} \\ &= \frac{1,140,526}{1,140,527} = .999999 \end{aligned}$$

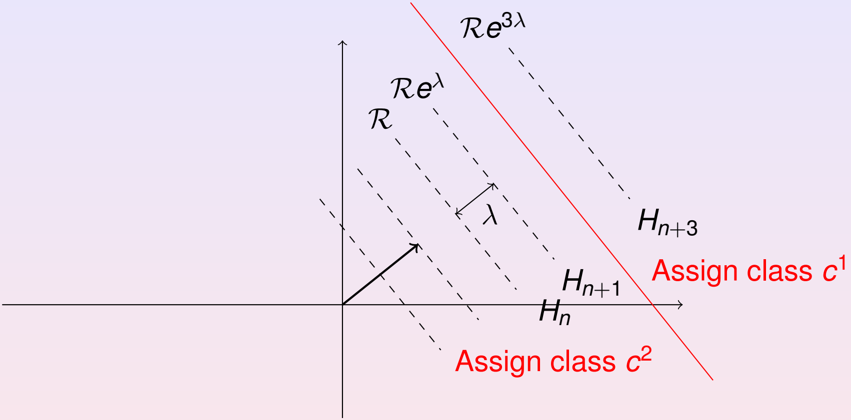
Logit Decision Rule

$$\log(\mathcal{R}(x;\theta_0,\theta)) = \theta_0 + \theta^{'}x$$

Assign class c^1 when

$$\theta_0 + \theta^{'}x \geq \Theta$$

Decision Rule



$$H_n = \{x \mid \theta' x = n\lambda\}$$

K Class Logistic Model

$$
\begin{aligned}
p(c_k|x) &= \frac{e^{\theta_{0k}+\theta_k^{'}x}}{1+\sum_{j=1}^{K-1} e^{\theta_{0j}+\theta_j^{'}x}},\ k=1,\ldots,K-1 \\
P(c_K|x) &= \frac{1}{1+\sum_{j=1}^{K-1} e^{\theta_{0j}+\theta_j^{'}x}} \\
\mathcal{R}_k(x) &= \frac{P(c_k|x)}{P(c_K|x)} \\
log(\mathcal{R}_k(x)) &= \theta_{0k}+\theta_k^{'}x,\ k=1,\ldots,K-1 \\
log(\mathcal{R}_K(x)) &= 1
\end{aligned}
$$

General Two Class Logistic Model

$$\begin{aligned}
\log(\mathcal{R}(x;\beta)) &= g(x;\beta) \\
P(c_1|x) &= \frac{e^{g(x;\beta)}}{1+e^{g(x;\beta)}} \\
P(c_2|x) &= \frac{1}{1+e^{g(x;\beta)}}
\end{aligned}$$

Decision Rule

Assign to class c^1 when

$$g(x;\beta) > \Theta$$

Otherwise assign to class c^2.

Logistic Regression

$$
\begin{aligned}
\log(\mathcal{R}(x;\beta)) &= g(x;\beta) \\
P(c_1|x) &= \frac{e^{g(x;\beta)}}{1+e^{g(x;\beta)}} \\
P(c_2|x) &= \frac{1}{1+e^{g(x;\beta)}}
\end{aligned}
$$

Logistic regression is the name given to the method that solves the estimation problem for β given a training set $<(c_1, x_1), \ldots, (c_N, x_N)>$, where c_n is the true class label associated with measurement vector x_n.

Training Data: Two Class

$$< (c_1, x_1), (c_2, x_2), \ldots, (c_M, x_M) >$$

Class Label c_m

- $c_m = 1$ for class 1
- $c_m = 0$ for class 2

Measurement Vector $x_m \in R^N$

Maximum Likelihood Estimation

Find θ_0 and θ to maximize

$$P(x_1, \ldots, x_M \mid c_1, \ldots, c_M, \theta_0, \theta)$$

Conditional Independence Assumption 1
Given the class labels and the parameters, the measurement vectors are independent.

$$P(x_1, \ldots, x_M \mid c_1, \ldots, c_M, \theta_0, \theta) = \prod_{m=1}^{M} P(x_m \mid c_1, \ldots, c_M, \theta_0, \theta)$$

Conditional Independence Assumption 2
No class labels other than c_m are relevant to measurement x_m

$$P(x_m \mid c_1, \ldots, c_M, \theta_0, \theta) = P(x_m \mid c_m, \theta_0, \theta)$$

Conditional Independence

Definition

x is conditionally independent of c_2 given c_1 if and only if

$$P(x \mid c_1, c_2) = P(x \mid c_1)$$

Theorem

$P(x \mid c_1, c_2) = P(x \mid c_1)$ *if and only if*

$$P(x, c_2 \mid c_1) = P(x \mid c_1) P(c_2 \mid c_1)$$

Conditional Independence

Theorem

$P(x \mid c_1, c_2) = P(x \mid c_1)$ *if and only if*

$$P(x, c_2 \mid c_1) = P(x \mid c_1) P(c_2 \mid c_1)$$

Proof.

$\Rightarrow$ *Suppose* $P(x \mid c_1, c_2) = P(x \mid c_1)$. *Then*

$$\begin{aligned}
P(x, c_2 \mid c_1) &= \frac{P(x, c_1, c_2)}{P(c_1)} \\
&= \frac{P(x \mid c_1, c_2) P(c_1, c_2)}{P(c_1)} \\
&= \frac{P(x \mid c_1) P(c_1, c_2)}{P(c_1)} \\
&= P(x \mid c_1) P(c_2 \mid c_1)
\end{aligned}$$

Conditional Independence

Theorem

$P(x \mid c_1, c_2) = P(x \mid c_1)$ *if and only if*

$$P(x, c_2 \mid c_1) = P(x \mid c_1) P(c_2 \mid c_1)$$

Proof.

$\Leftarrow$ *Suppose* $P(x, c_2 \mid c_1) = P(x \mid c_1) P(c_2 \mid c_1)$. *Then*

$$\begin{aligned} P(x \mid c_1, c_2) &= \frac{P(x, c_1, c_2)}{P(c_1, c_2)} \\ &= \frac{P(x, c_2 \mid c_1) P(c_1)}{P(c_1, c_2)} \\ &= \frac{P(x \mid c_1) P(c_2 \mid c_1) P(c_1)}{P(c_1, c_2)} \\ &= P(x \mid c_1) \end{aligned}$$

Maximum Likelihood Estimation

Use the conditional independences to find θ_0 and θ to maximize

$$\mathcal{L}(\theta_0, \theta) = P(x_1, \ldots, x_M \mid c_1, \ldots, c_M, \theta_0, \theta)$$

Find θ_0 and θ to maximize

$$\begin{aligned}
\mathcal{L}(\theta_0, \theta) &= \prod_{m=1}^{M} P(x_m \mid c_m, \theta_0, \theta) \\
&= \prod_{m=1}^{M} \frac{P(c_m \mid x_m, \theta_0, \theta) P(x_m, \theta_0, \theta)}{P(c_m, \theta_0, \theta)}
\end{aligned}$$

Assume x_m and (θ_0, θ) are independent and c_m and (θ_0, θ) are independent so that

$$\begin{aligned}
P(x_m, \theta_0, \theta) &= P(x_m) P(\theta_0, \theta) \\
P(c_m, \theta_0, \theta) &= P(c_m) P(\theta_0, \theta)
\end{aligned}$$

Maximum Likelihood Estimation

Then

$$
\begin{aligned}
P(x_m \mid c_m, \theta_0, \theta) &= \frac{P(c_m|x_m, \theta_0, \theta)P(x_m, \theta_0, \theta)}{P(c_m, \theta_0, \theta)} \\
&= \frac{P(c_m \mid x_m, \theta_0, \theta)P(\theta_0, \theta)P(x_m)}{P(\theta_0, \theta)P(c_m)} \\
&= \frac{P(c_m \mid x_m, \theta_0, \theta)P(x_m)}{P(c_m)}
\end{aligned}
$$

so that

$$
\prod_{m=1}^{M} P(x_m \mid c_m, \theta_0, \theta) = \prod_{m=1}^{M} \frac{P(c_m \mid x_m, \theta_0, \theta)P(x_m)}{P(c_m)}
$$

Maximum Likelihood Estimation

$$\begin{aligned}\mathcal{L}(\theta_0, \theta) &= \prod_{m=1}^{M} P(x_m \mid c_m, \theta_0, \theta) \\ &= \prod_{m=1}^{M} \frac{P(c_m \mid x_m, \theta_0, \theta) P(x_m)}{P(c_m)}\end{aligned}$$

Since $x_1, \ldots, x_M$ and $c_1, \ldots, c_M$ are given, each $P(x_m)$ and $P(c_m)$ are fixed so that the (θ_0, θ) that maximizes $\mathcal{L}(\theta_0, \theta)$ maximizes

$$\prod_{m=1}^{M} P(c_m \mid x_m, \theta_0, \theta)$$

Maximum Likelihood Estimation

Find the (θ_0, θ) to maximize

$$\begin{aligned}
\prod_{m=1}^{M} P(c_m \mid x_m, \theta_0, \theta) &= \prod_{m=1}^{M} \begin{cases} \frac{e^{\theta_0+\theta' x_m}}{1+e^{\theta_0+\theta' x_m}} & \text{if } c_m = 1 \\ \frac{1}{1+e^{\theta_0+\theta' x_m}} & \text{if } c_m = 0 \end{cases} \\
&= \prod_{m=1}^{M} \frac{e^{c_m(\theta_0+\theta' x_m)}}{1+e^{\theta_0+\theta' x_m}} \\
&= \mathcal{L}^*(\theta_0, \theta)
\end{aligned}$$

Log Likelihood

Since the log function is strictly monotonically increasing, the parameters that maximize $\mathcal{L}^*$ maximize $\log \mathcal{L}^*$.
Find θ_0, θ to maximize

$$\begin{aligned} \log \mathcal{L}^*(\theta_0, \theta) &= \log\left(\prod_{m=1}^{M} \frac{e^{c_m(\theta_0+\theta' x_m)}}{1 + e^{\theta_0+\theta' x_m}}\right) \\ &= \sum_{n=1}^{M} c_m(\theta_0 + \theta' x_m) - \log(1 + e^{\theta_0+\theta' x_m}) \end{aligned}$$

Log Likelihood

Transformation of Variables

$$x_{new} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \quad \theta_{new} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}$$

Then

$$\log \mathcal{L}^*(\theta_0, \theta) = \sum_{m=1}^{M} c_m(\theta'_{new} x_{new\ m}) - \log(1 + e^{\theta'_{new} x_{new\ m}})$$

Log Likelihood

θ maximizes

$$\log \mathcal{L}^*(\theta) \quad = \quad \sum_{n=1}^{M} c_m(\theta' x_m) - \log(1 + e^{\theta' x_m})$$

if and only if

- $\frac{\partial}{\partial \theta} \log \mathcal{L}^*(\theta) = 0$
- $\frac{\partial^2}{\partial \theta \partial \theta} \log \mathcal{L}^*(\theta)$ is negative definite

$$\frac{\partial}{\partial \theta} \log \mathcal{L}^*(\theta) \quad = \quad \sum_{m=1}^{M} c_m x_m - \frac{1}{1 + e^{\theta' x_m}} e^{\theta' x_m} x_m = 0$$

$$\frac{\partial^2}{\partial \theta \partial \theta} \log \mathcal{L}^*(\theta) \quad = \quad - \sum_{m=1}^{M} x_m x_m' \frac{e^{\theta' x_m}}{(1 + e^{\theta' x_m})^2} = \sum_{m=1}^{M} \frac{-x_m x_m'}{2 + e^{\theta' x_m} + e^{-\theta' x_m}}$$

Positive and Negative Definite

Definition

A matrix B is positive definite if and only if for every $x \neq 0$,

$$x^{'}Bx > 0$$

A matrix B is negative definite if and only if for every $x \neq 0$,

$$x^{'}Bx < 0$$

Theorem

B is negative definite if and only if $-B$ is positive definite

Proof.

$$\begin{aligned} x^{'}Bx &= < 0 \\ x^{'}(-B)x &= > 0 \end{aligned}$$

Log Likelihood

Is

$$\frac{\partial^2}{\partial\theta\partial\theta}\log\mathcal{L}^*(\theta) \quad = \quad -\sum_{m=1}^{M}\frac{x_m x_m'}{2+e^{\theta' x_m}+e^{-\theta' x_m}}$$

negative definite?

Is

$$\sum_{m=1}^{M}\frac{x_m x_m'}{2+e^{\theta' x_m}+e^{-\theta' x_m}}$$

positive definite?

Log Likelihood

$$-\frac{\partial^2}{\partial\theta\partial\theta}\log\mathcal{L}^*(\theta) = \sum_{m=1}^{M}\frac{x_m x_m^{'}}{2+e^{\theta' x_m}+e^{-\theta' x_m}}$$

Examine $-\frac{\partial^2}{\partial\theta\partial\theta}\log\mathcal{L}^*(\theta)$ and fix θ.

$$\begin{aligned}-\frac{\partial^2}{\partial\theta\partial\theta}\log\mathcal{L}^*(\theta) &= \sum_{m=1}^{M} x_m x_m^{'} k_m^2 \\ &= \sum_{m=1}^{M}(k_m x_m)(k_m x_m)^{'}\end{aligned}$$

Positive Definite

$$
\begin{aligned}
x' \sum_{m=1}^{M} (k_m x_m)(k_m x_m)' x &= \sum_{m=1}^{M} (x' k_m x_m)(k_m x_m' x) \\
&= \sum_{m=1}^{M} (k_m x_m' x)(k_m x_m' x) \\
&= \sum_{m=1}^{M} (k_m x_m' x)^2 \\
&> 0
\end{aligned}
$$

if and only if $\sum_{m=1}^{M} (k_m x_m)(k_m x_m)'$ is of full rank
if and only if $< x_1, \ldots, x_M >$ spans R^N

Maximum Likelihood Estimation

Find θ so that

$$\log \mathcal{L}^*(\theta) = \sum_{n=1}^{M} c_m(\theta' x_m) - \log(1 + e^{\theta' x_m})$$

is maximized. This happens if and only if

$$\frac{\partial}{\partial \theta} \log \mathcal{L}^*(\theta) = \sum_{m=1}^{M} c_m x_m - \frac{1}{1 + e^{\theta' x_m}} e^{\theta' x_m} x_m = 0$$

and $< x_1, \ldots, x_M >$ spans R^N

Maximum Likelihood Estimation

Find θ so that

$$-\log \mathcal{L}^*(\theta) \quad = \quad \sum_{n=1}^{M} -c_m(\theta' x_m) + \log(1 + e^{\theta' x_m})$$

is minimized. This happens if and only if

$$\begin{aligned} -\frac{\partial}{\partial \theta} \log \mathcal{L}^*(\theta) \quad &= \quad \sum_{m=1}^{M} -c_m x_m + \frac{1}{1 + e^{\theta' x_m}} e^{\theta' x_m} x_m = 0^{N \times 1} \\ &= \quad \sum_{m=1}^{M} -c_m x_m + \frac{1}{1 + e^{-\theta' x_m}} x_m = 0^{N \times 1} \end{aligned}$$

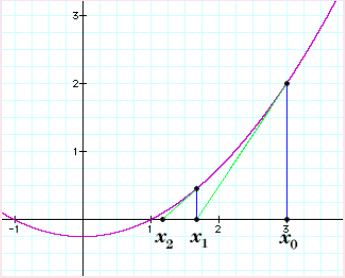
and $< x_1, \ldots, x_M >$ spans R^N

Root Finding

Given a function $f(\theta)$, find θ so that $f(\theta) = 0$
Newton's Method 1D

$$
\begin{aligned}
f(x_{k+1}) &= f(x_k) + (x_{k+1} - x_k) f'(x_k) \\
x_{k+1} &= x_k + \frac{f(x_{k+1}) - f(x_k)}{f'(x_k)}
\end{aligned}
$$

Want $f(x_{k+1}) = 0$. Hence, $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$



Root Finding

N-Dimensional: $x^{N\times 1}$, $f(x)^{N\times 1}$

$$
\begin{aligned}
f(x_{k+1}) &= f(x_k) + \frac{\partial}{\partial z} f(z)\Big|_{z=x_k}(x_{k+1} - x_k) \\
&= f(x_k) + J(x_k)(x_{k+1} - x_k)
\end{aligned}
$$

Set $f(x_{k+1}) = 0$

$$
x_{k+1} = x_k - J^{-1}(x_k) f(x_k)
$$

The way it is actually solved: Set $z = -J^{-1}(x_k)f(x_k)$
Find z to satisfy

$$
-f(x_k) = J(x_k) z
$$

Define

$$
x_{k+1} = z + x_k
$$

Maximum Likelihood: Logistic Regression

Represented as a minimization problem.

$$
\begin{aligned}
-\log \mathcal{L}^*(\theta) &= \sum_{n=1}^{M} -c_m(\theta' x_m) + \log(1 + e^{\theta' x_m}) \\
f(\theta) &= -\frac{\partial}{\partial \theta} \log \mathcal{L}^*(\theta) = \sum_{m=1}^{M} -c_m x_m + \frac{1}{1 + e^{-\theta' x_m}} x_m \\
J(\theta) &= -\frac{\partial^2}{\partial \theta \partial \theta} \log \mathcal{L}^*(\theta) = \sum_{m=1}^{M} \frac{x_m x_m'}{2 + e^{\theta' x_m} + e^{-\theta' x_m}}
\end{aligned}
$$

Find z to satisfy

$$-f(\theta_k) = J(\theta_{k+1}) z$$

Set

$$\theta_{k+1} = z + \theta_k$$

Metabolic Marker Data

- **g** Group index
- x_g Metabolic Marker Value
- n_g Number Patients died
- N_g Total Number Patients
- $p_{obs}(g) = n_g/N_g$ Observed Proportion died

g	x_g	n_g	N_g	$p_{obs}(g)$
1	0.75	7	182	.0385
2	1.25	27	233	.116
3	1.75	44	224	.196
4	2.25	91	236	.386
5	2.75	130	225	.578
6	3.25	168	215	.781
7	3.75	194	221	.878
8	4.25	191	200	.955
9	4.75	260	264	.985

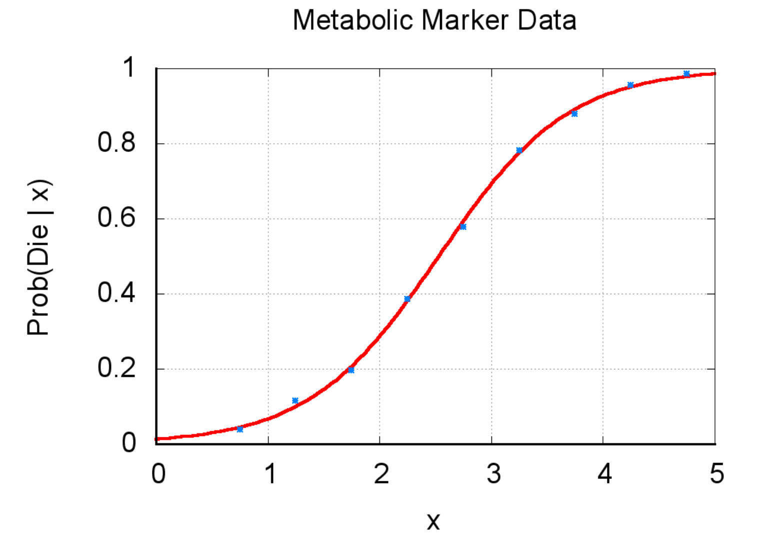
Grouped Data Calculations

$$
\begin{aligned}
x_g & \quad \text{1 D} \\
t_g & = \theta_0 + \theta x_g \\
B & = \text{Bound; } \; e^B \text{has a value}
\end{aligned}
$$

$$
f = -\sum_{g=1}^{G} \begin{pmatrix} n_g \\ x_g n_g \end{pmatrix} + \sum_{\{g \mid t_g > -B\}} \begin{pmatrix} \frac{N_g}{1+e^{-t_g}} \\ \frac{x_g N_g}{1+e^{-t_g}} \end{pmatrix}
$$

$$
J = \sum_{\{g \mid -B < t_g < B\}} \begin{pmatrix} \frac{N_g}{2+e^{t_g}+e^{-t_g}} & \frac{x_g N_g}{2+e^{t_g}+e^{-t_g}} \\ \frac{x_g N_g}{2+e^{t_g}+e^{-t_g}} & \frac{x_g^2 N_g}{2+e^{t_g}+e^{-t_g}} \end{pmatrix}
$$

Metabolic Marker Data



Model Questions

- Does the model fit the data well enough?
- Does the fitted model generalize to unseen data?
- Would the model fitting on chance data produce as good a fit as it did on the real data?

Goodness of Fit

Grouped Data

- G groups
- x_g Vector Value for group g
- K Dimension of vector
- n_g Observed number who died for group g (need $n_g > 5$)
- N_g Total number in group g
- $p_{exp}(g) = 1/(1 + e^{-\theta' x_g})$ Logistic Model probability

$$\begin{aligned}\chi^2_{obs} &= \sum_{g=1}^{G} \frac{(n_g - N_g * p_{exp}(g))^2}{N_g * p_{exp}(g)} \\ p_{value} &= Prob(\chi^2 > \chi^2_{obs} \mid G - K)\end{aligned}$$

If p_{value} is too small, then reject the hypothesis that the model fits the data.

$$\chi^2_{obs} = \sum_{g=1}^{G} \frac{(n_g - N_g * p_{exp}(g))^2}{N_g * p_{exp}(g)}$$

g	x_g	n_g	N_g	$p_{exp}(g)$	$N_g p_{exp}(g)$
1	0.75	7	182	.044	8.04
2	1.25	27	233	.099	22.98
3	1.75	44	224	.206	46.10
4	2.25	91	236	.380	89.73
5	2.75	130	225	.592	133.26
6	3.25	168	215	.775	166.57
7	3.75	194	221	.891	196.83
8	4.25	191	200	.951	190.14
9	4.75	260	264	.979	258.34

$$Prob(\chi^2_7 > \chi^2_{obs} = 1.098) = .993$$

Distribution Free Test

Real Data

x_g	Value for group g
N_g	Number of people in group g
n_g	Number of people in group g in class c[1]

Use x_g, N_g, n_g in logistic regression to estimate the parameters θ_0, θ

$$p_{exp}(g) = \frac{1}{1 + e^{-\theta_0 - \theta x_g}}$$

Use $N_g, n_g, p_{exp}(g)$ to determine goodness of fit statistics X_0^2.

$$X_0^2 = \sum_{g=1}^{G} \frac{(n_g - N_g * p_{exp}(g))^2}{N_g * p_{exp}(g)}$$

MonteCarlo Experiment

Let $r_{11}, \ldots, r_{GM}$ be independent U(0,1) random variables. Define

$$\hat{n}_g = \#\{j \mid r_{gj} \leq n_g/N_g\},\ g = 1, \ldots, G$$

Use $x_g, N_g, \hat{n}_g$ in logistic regression to estimate the parameters θ_0, θ.

$$p_{exp}(g) = \frac{1}{1 + e^{-\theta_0 - \theta x_g}}$$

Use $N_g, p_{exp}(g), \hat{n}_g$ to determine X^2 statistic.
Repeat Z times generating $X_1^2, X_2^2, \ldots X_Z^2$

$$p_{value} = \frac{\#\{z \mid X_z^2 \geq X_0^2\}}{Z}$$

Reject the hypothesis that the model fits the data if p_{value} is too small.

MonteCarlo Experiment

Z trials estimating $\theta_0 : \ \theta_{01}, \theta_{02}, \ldots, \theta_{0Z}$
$\theta : \ \theta_1, \theta_2, \ldots, \theta_Z$

$$
\begin{aligned}
\mu(\theta_0) &= \frac{1}{Z} \sum_{z=1}^{Z} \theta_{0z} \\
\mu(\theta) &= \frac{1}{Z} \sum_{z=1}^{Z} \theta_z \\
\sigma^2(\theta_0) &= \frac{1}{Z-1} \sum_{z=1}^{Z} (\theta_{0z} - \mu(\theta_0))^2 \\
\sigma^2(\theta) &= \frac{1}{Z-1} \sum_{z=1}^{Z} (\theta_z - \mu(\theta))^2
\end{aligned}
$$

Estimating Parameter Confidence Intervals

MonteCarlo Experiment

Z trials estimating $\theta_0 : \; \theta_{01}, \theta_{02}, \ldots, \theta_{0Z}$

$\theta : \; \theta_1, \theta_2, \ldots, \theta_Z$

Order them from smallest to largest.

$$\begin{aligned} \theta_0 : &\quad \theta_{(0,1)} \leq \theta_{(0,2)} \leq \ldots \leq \theta_{(0,Z)} \\ \theta : &\quad \theta_{(1)} \leq \theta_{(2)} \leq \ldots \leq \theta_{(Z)} \end{aligned}$$

$100\frac{Z-2m}{Z}$% central confidence interval for:

θ_0 is $(\theta_{(0,m)}, \theta_{(0,Z-m)})$

θ is $(\theta_{(m)}, \theta_{(Z-m)})$

Cross Validation

- x is measurement vector
- c is class

$$\text{Data Set} \quad < (c_1, x_1), \ldots, (c_N, x_N) >$$

- Partition Data set into K blocks
- Estimate Model parameters from $K - 1$ blocks
- Test goodness of fit on K^{th} block
- Rotate K times
- Aggregate goodness of fit results

Goodness of Model Fit on Chance Data

What does chance data mean?

It cannot mean data that comes from an underlying model with structure because then we certainly expect a fit to be good modulo the degree of noise perturbation.

It must mean data that comes from a model with no structure, meaning no underlying relationship between the class and the measurement vector.

Goodness of Model Fit on Chance Data

Permutation Test

Let $\pi =< \pi_1, \pi_2, \ldots, \pi_M >$ be a random permutation of $< 1, 2, \ldots, M >$

$$\text{Observed data: } < (c_1, x_1), \ldots, (c_M, x_M) >$$
$$\text{Randomly permuted data: } < (c_{\pi_1}, x_1), (c_{\pi_2}, x_2), \ldots, (c_{\pi_M}, x_M) >$$

Perform the model fitting on the observed data and get a goodness of fit X_0^2.
Perform a model fitting on randomly permuted data Z times getting goodness of fits $X_1^2, \ldots, X_Z^2$.

$$p_{value} = \frac{\#\{z \mid X_z^2 > X_0^2\} + \frac{1}{2}\#\{z \mid X_z^2 = X_0^2\}}{Z}$$

If p_{value} is too small reject the hypothesis that the fitted model is statistically significant.