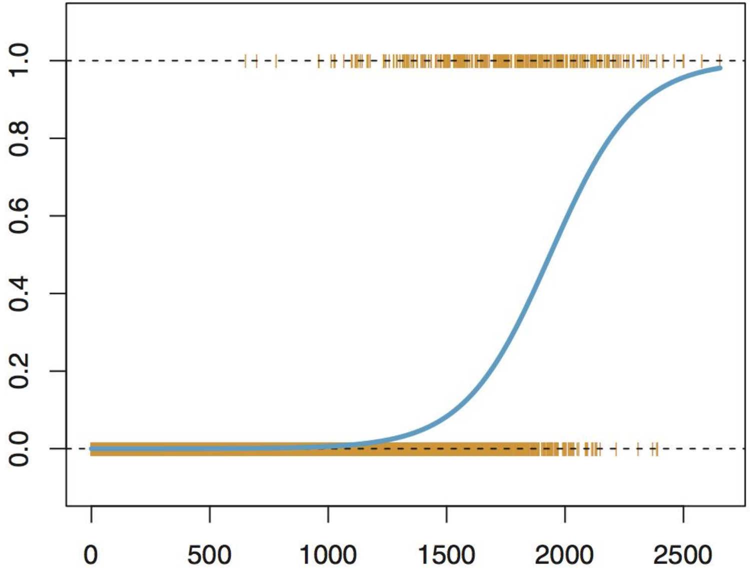
Logistic Regression 2/2

DSI, jf.omhover, Dec 6, 2016



Logistic Regression 2/2

DSI, jf.omhover, Dec 6, 2016

OBJECTIVES (morning)

- **Relate** Regression to Classification in the context of supervised learning
- **Compare** Logistic Regression to Linear Regression
- **Define** and **compute** metrics for evaluating classifiers

OBJECTIVES (afternoon)

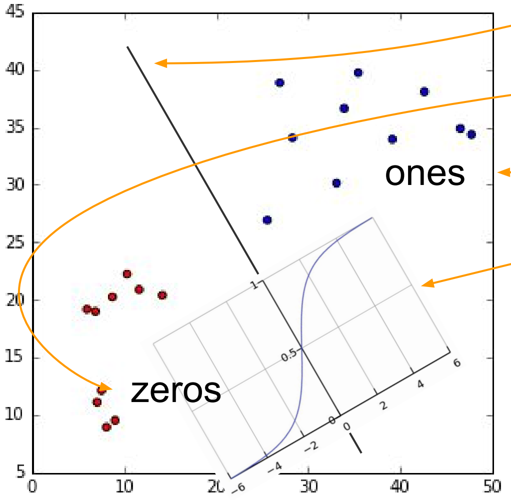
- **Describe** the process for computing parameter values in LogReg
- **Use** the parameters of a LogReg model to **compute** the class of an obverstion

Using LogReg to predict

Let’s suppose we have a LogReg model already...

LogReg in a nutshell

REALITY



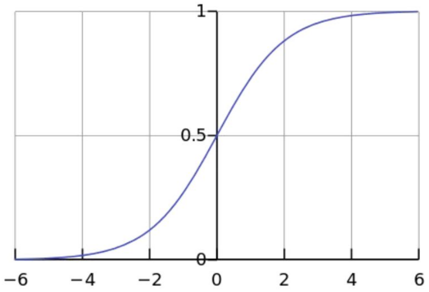
MODEL

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

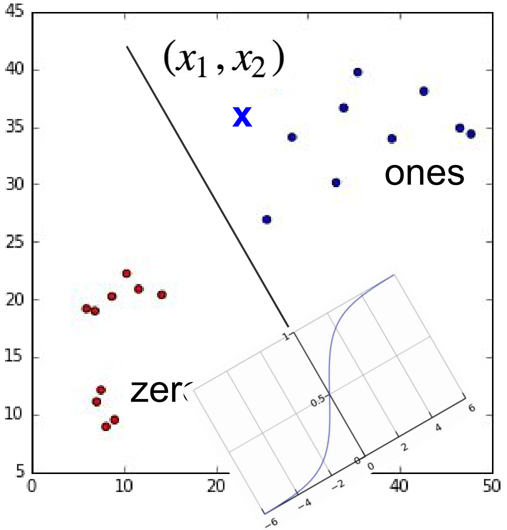
It (badly) translates as :
computes the probability of being in one of the two classes
depending on of the side and distance of the plan

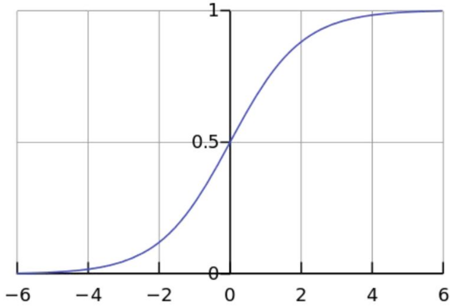
$$h : \mathbb{R} \rightarrow [0, 1]$$

$$h(t) = \frac{1}{1+e^{-t}}$$



Propagation of a change in attributes



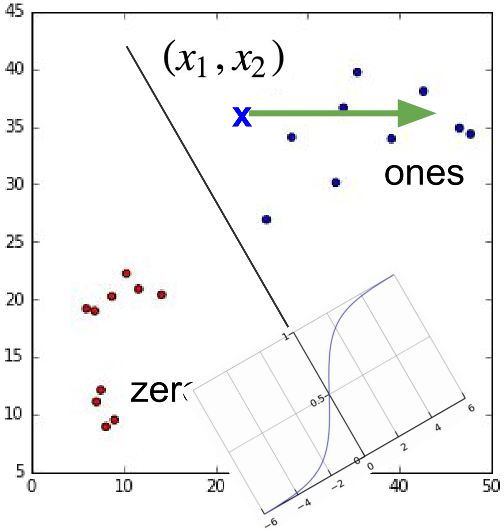


$$h : \mathbb{R} \to [0, 1]$$

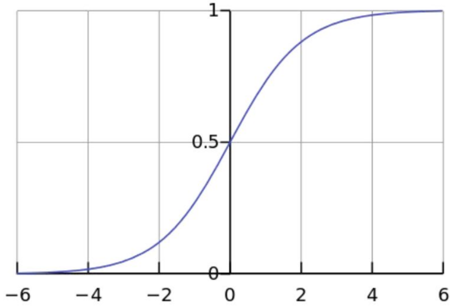
$$h(t) = \frac{1}{1+e^{-t}}$$

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Propagation of a change in attributes



x_1 increase

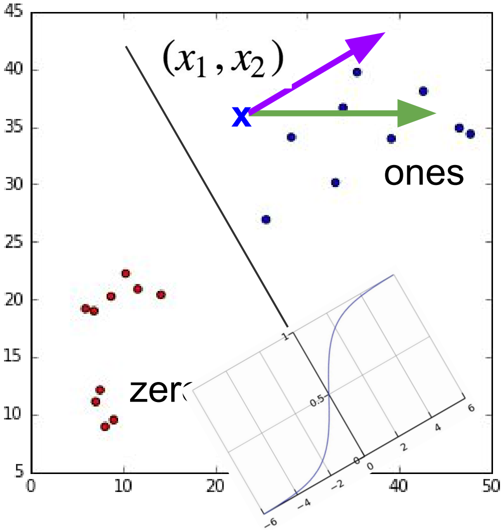


$h : \mathbb{R} \to [0, 1]$

$h(t) = \frac{1}{1+e^{-t}}$

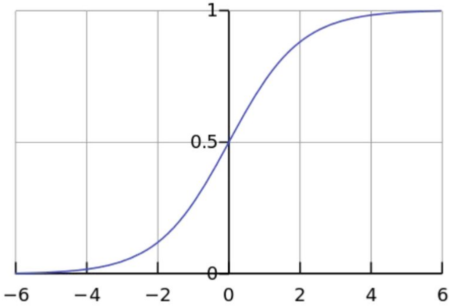
$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Propagation of a change in attributes



x_1 increase

$\beta_1 . x_1$

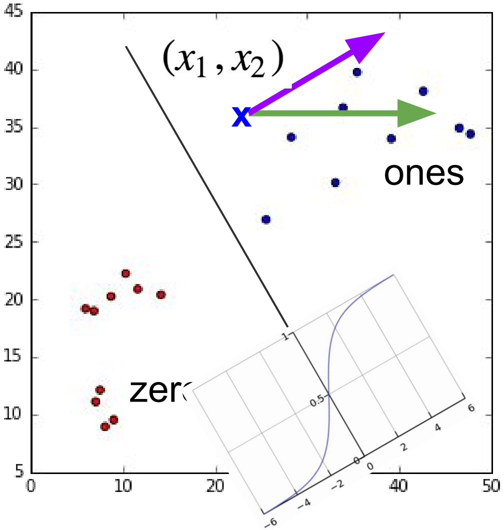


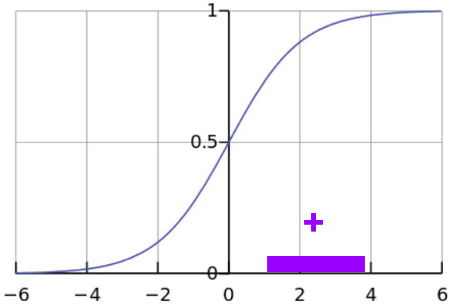
$$h : \mathbb{R} \to [0, 1]$$

$$h(t) = \frac{1}{1+e^{-t}}$$

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Propagation of a change in attributes





$h : \mathbb{R} \to [0, 1]$

$h(t) = \frac{1}{1+e^{-t}}$

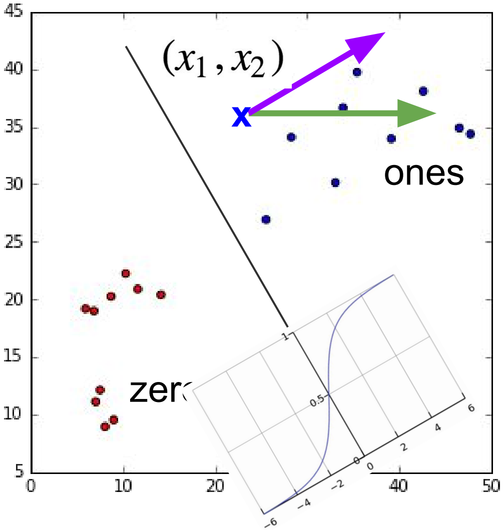
x_1 increase

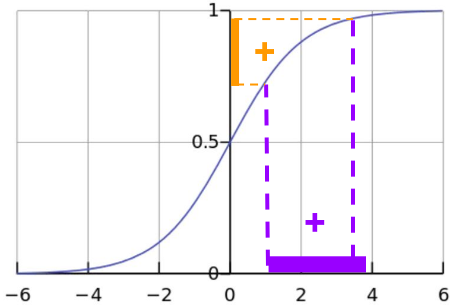
$\beta_1 . x_1$

$+ \; (\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Propagation of a change in attributes





$$h : \mathbb{R} \to [0, 1]$$

$$h(t) = \frac{1}{1+e^{-t}}$$

x_1 increase

$\beta_1 . x_1$

$+ \; (\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$

$+ \; h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Interpreting coefficients

Making sense of the logistic function

Probs, odds, log-odds, odds-ratio

Probabilities range between 0 and 1.

[examples link]

$$p(x)$$

Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted.

For males: **p = 7/10 = .7** **1 - p = 1 - .7 = .3**
For females: **p = 3/10 = .3** **1 - p = 1 - .3 = .7**

Odds are defined as the ratio of the probability of success and the probability of failure.

$$\frac{p(X)}{1-p(X)}$$

odds(male) = .7/.3 = 2.33333
odds(female) = .3/.7 = .42857

Log-odds are the log of odds

$$log(\frac{p(X)}{1-p(X)})$$

Odds-ratio is comparing two properties in terms of odds.

$$\frac{odds(A)}{odds(B)}$$

OR = 2.3333/.42857 = 5.44

Thus, for a male, the odds of being admitted are 5.44 times larger than the odds for a female being admitted.

Probs, odds, log-odds, odds-ratio in LogReg

Probabilities range between 0 and 1.

$$p(x) = \frac{e^{\beta^T.x}}{1+e^{\beta^T.x}}$$

[examples link]

Odds are defined as the ratio of the probability of success and the probability of failure.

$$\frac{p(X)}{1-p(X)} = e^{\beta^T.x}$$

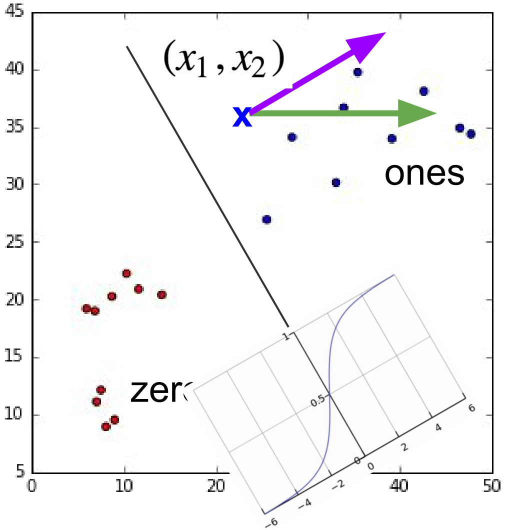
Log-odds are the log of odds

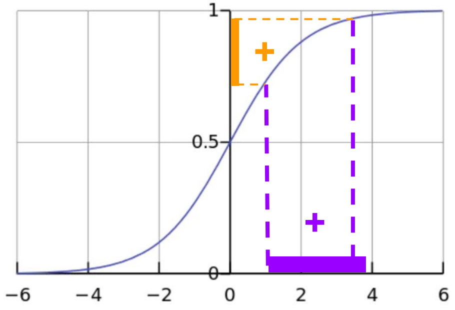
$$log(\frac{p(X)}{1-p(X)}) = \beta^T.x = \beta_0 + \beta_1.x_1 + \ldots + \beta_n.x_n$$

Odds-ratio is comparing two properties in terms of odds.

$$\frac{odds(A)}{odds(B)} \qquad OR = e^{\beta_i}$$

Propagation of a change in attributes





$$h : \mathbb{R} \to [0, 1]$$

$$h(t) = \frac{1}{1+e^{-t}}$$

x_1 increase

$\beta_1 . x_1$

+ $(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$

+ $h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$

$$p(X) = h(\beta_0 + \beta_1 . x_1 + \cdots + +\beta_p . x_p)$$

Estimating a LogReg Model

NOW, machine, it’s your turn to learn...

Notations

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y$$

x_i, y_i

...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

$p(x_i)$

1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

$$p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$$

Likelihood of the LogReg model

x_i, y_i	
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

Likelihood of the LogReg model

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

What is the "likelihood" of our dataset to have be drawn out of that probability ?

Likelihood of the LogReg model

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

$p(x_i)$	
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

What is the "likelihood" of our dataset to have be drawn out of that probability ?

Let's first do that for each observation

$$y_i = 1 \implies p(x_i)$$

$$y_i = 0 \implies 1 - p(x_i)$$

Likelihood of the LogReg model

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \dfrac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

What is the "likelihood" of our dataset to have be drawn out of that probability ?

Let's first do that for each observation

$$y_i = 1 \implies p(x_i)$$

$$y_i = 0 \implies 1 - p(x_i)$$

Let's do that for the whole dataset $L(\beta) = \prod_{i:y_i=1} p(x_i) * \prod_{i:y_i=0} (1 - p(x_i))$

Optimization of the likelihood

x_i, y_i	
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

$p(x_i)$	
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$L(\beta) = \prod_{i:y_i=1} p(x_i) * \prod_{i:y_i=0} (1 - p(x_i))$$

Can we find the maximum of that likelihood ?

Optimization of the likelihood

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$L(\beta) = \prod_{i:y_i=1} p(x_i) * \prod_{i:y_i=0} (1 - p(x_i))$$

Can we find the maximum of that likelihood ?

$$LogL(\beta) = \sum_{i:y_i=1} log(p(x_i)) + \sum_{i:y_i=0} log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . log(p(x_i)) + (1 - y_i). log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . \beta^T x_i - log(1 + e^{\beta^T x_i})$$

Optimization of the likelihood

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$L(\beta) = \prod_{i:y_i=1} p(x_i) * \prod_{i:y_i=0} (1 - p(x_i))$$

Can we find the maximum of that likelihood ?

$$LogL(\beta) = \sum_{i:y_i=1} log(p(x_i)) + \sum_{i:y_i=0} log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . log(p(x_i)) + (1 - y_i). log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . \beta^T x_i - log(1 + e^{\beta^T x_i})$$

that we can differentiate...

$$\frac{\partial LL(\beta)}{\partial \beta} = \sum_i x_i . (y_i - p(x_i; \beta))$$

Optimization of the likelihood

x_i	y_i
...	1
...	0
...	0
...	1
...	1
...	1
...	0
...	1

	$p(x_i)$
1	0.95
0	0.21
1	0.55
0	0.43
1	0.77
0	0.44
0	0.15
1	0.81

Let's suppose we have $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ which gives us $p(x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$L(\beta) = \prod_{i:y_i=1} p(x_i) * \prod_{i:y_i=0} (1 - p(x_i))$$

Can we find the maximum of that likelihood ?

$$LogL(\beta) = \sum_{i:y_i=1} log(p(x_i)) + \sum_{i:y_i=0} log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . log(p(x_i)) + (1 - y_i). log(1 - p(x_i))$$

$$LogL(\beta) = \sum_i y_i . \beta^T x_i - log(1 + e^{\beta^T x_i})$$

that ... ntiate...

spoiler alert

$$\frac{\partial LL(\beta)}{\partial \beta} = \sum_i x_i . (y_i - p(x_i; \beta))$$

Logistic Regression 2/2

DSI, jf.omhover, Dec 6, 2016