GENERAL ASSEMBLY

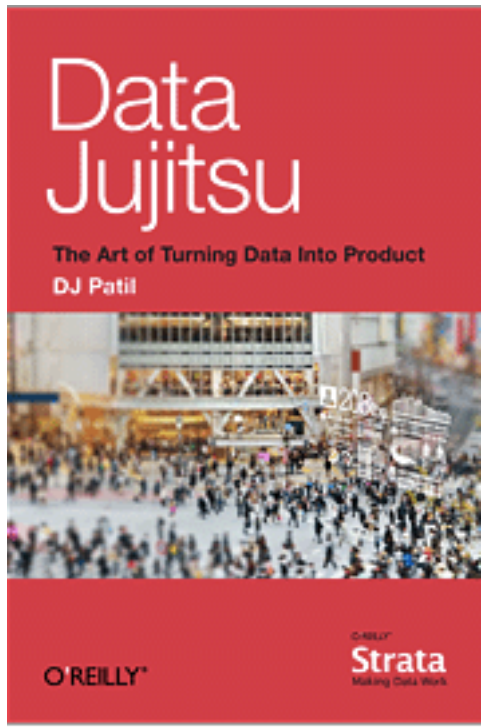
INTRO TO DATA SCIENCE

LECTURE 7: LOGISTIC REGRESSION

Francesco Mosconi
DAT16 SF // August 19th, 2015

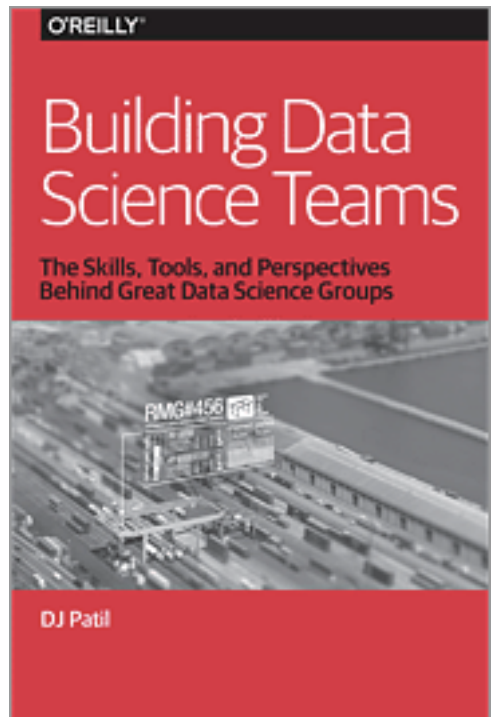
DATA SCIENCE IN THE NEWS

DATA SCIENCE IN THE NEWS



source: http://radar.oreilly.com/2012/07/data-jujitsu.html

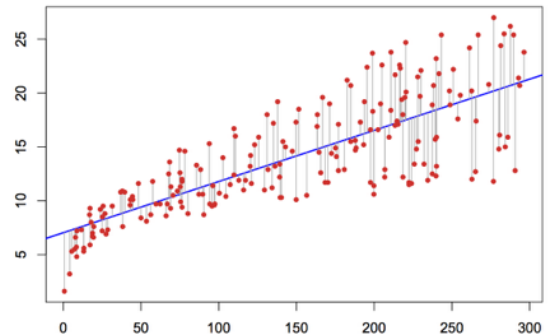
DATA SCIENCE IN THE NEWS



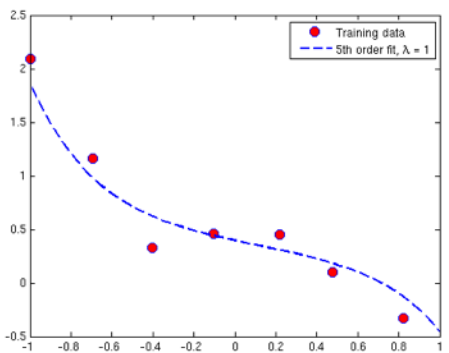
source: http://www.oreilly.com/data/free/building-data-science-teams.csp

RECAP

LAST TIME:



I. LINEAR REGRESSION (INCL. MULTIPLE REGRESSION)
II. POLYNOMIAL REGRESSION
III. REGULARIZATION



LAB:
IV. IMPLEMENTING MULTIPLE REGRESSION & POLYNOMIAL REGRESSION IN PYTHON

QUESTIONS?

QUESTIONS?

WHAT WAS THE MOST INTERESTING THING YOU LEARNT?

WHAT WAS THE HARDEST TO GRASP?

AGENDA

I. LOGISTIC REGRESSION
II. OUTCOME VARIABLES
III. ERROR TERMS
IV. INTERPRETING RESULTS

LAB: IMPLEMENTING LOGISTIC REGRESSION IN PYTHON

KEY OBJECTIVES

- WHAT IS LOGISTIC REGRESSION
- HOW IS LOGISTIC REGRESSION USED
- WHAT ARE THE ADVANTAGES OF USING LOGISTIC REGRESSION
- HOW TO IMPLEMENT LOGISTIC REGRESSION IN PYTHON

I. LOGISTIC REGRESSION

SO, WHERE ARE WE ON OUR 2X2?

	Continuous	*Categorical*
Supervised	*???*	*???*
Unsupervised	*???*	*???*

- *Name is somewhat misleading…*
- *Really a technique for classification, not regression*
- *"Regression" comes from fact that we fit a linear model to the feature space*

	Continuous	***Categorical***
Supervised	*regression*	*classification*
Unsupervised	*dimension reduction*	*clustering*

Q: What is logistic regression*?*

Q: What is logistic regression*?*

A: A generalization of the linear regression model to classification problems.

Q: Where is LR used?

Q: Where is LR used?

A: Many commercially valuable classification problems:

- *Fraud detection (payments, e-commerce)*
- *Churn prediction (marketing)*
- *Medical diagnoses (is the test positive or negative?)*
- *and many, many others...*

It's a binary classification technique

which means….

Two classes: Y = {0,1}

Our goal is to learn to classify correctly two types of examples

- *Class 0 – labeled as 0*
- *Class 1 – labeled as 1*

We would like to learn ***f: X —> {0, 1}***

In linear regression, we used a set of covariates to predict the value of ***a (continuous) outcome variable****.*

In linear regression, we used a set of covariates to predict the value of ***a (continuous) outcome variable*****.**

In logistic regression, we use a set of covariates to predict ***probabilities of (binary) class membership*****.**

In linear regression, we used a set of covariates to predict the value of ***a (continuous) outcome variable.***

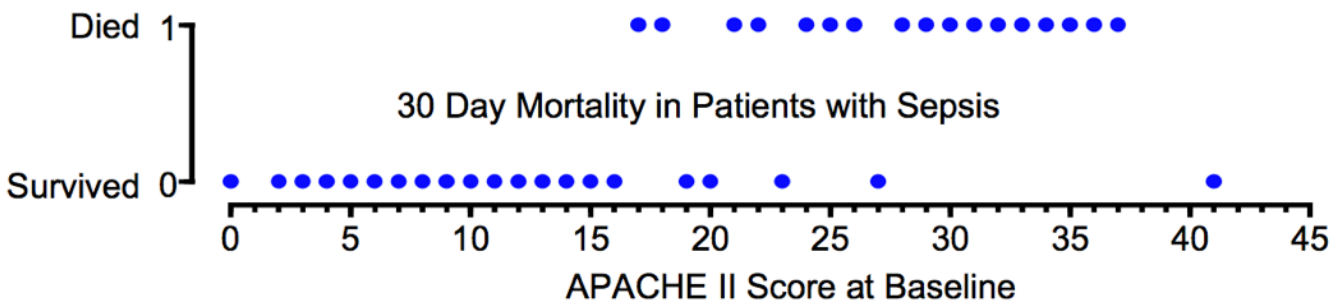
In logistic regression, we use a set of covariates to predict ***probabilities of (binary) class membership***

These probabilities are then mapped to class labels, thus solving the classification problem.

A motivating problem:

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

How can we predict death from baseline APACHE II score in these patients?

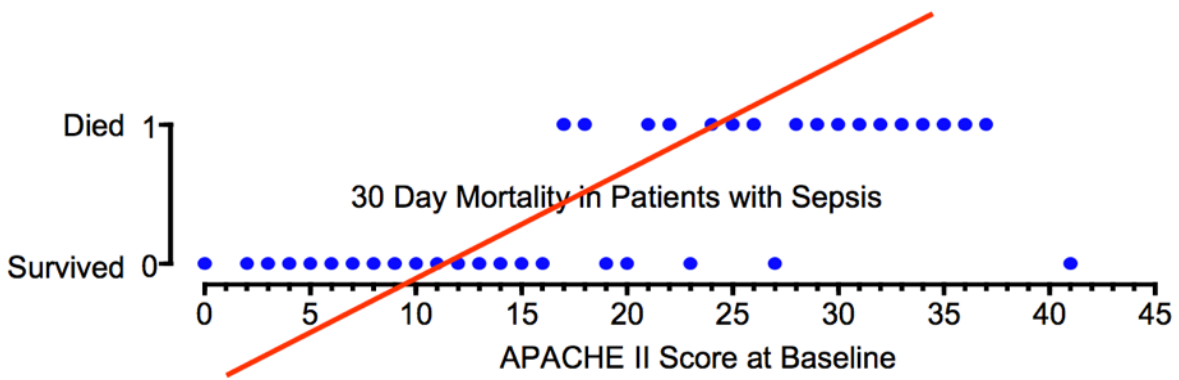


Source: http://www.mc.vanderbilt.edu/gcrc/workshop_files/2004-11-12.pdf

Q: How can we predict death from baseline APACHE II score in these patients?

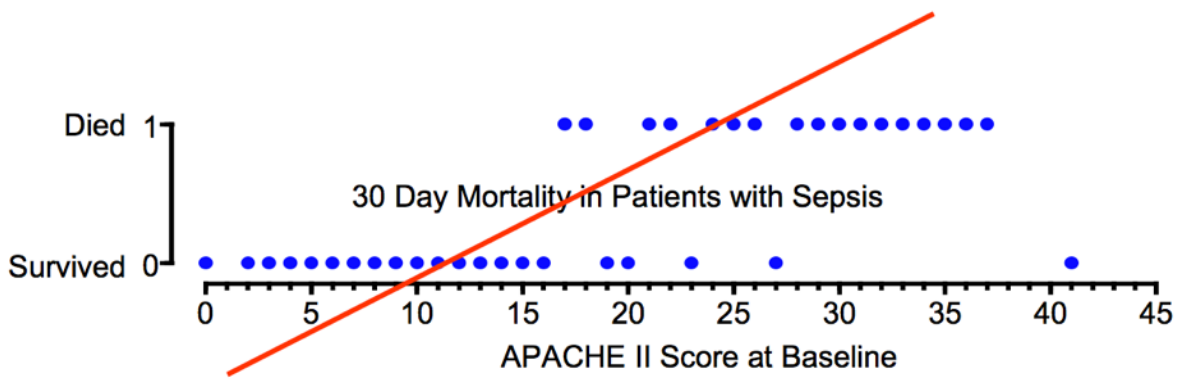
Let p(x) be the probability that a patient with score x will die within 30 days.

Well, linear regression would not work well here, because it could produce probabilities less than zero or greater than one. Also, one new value could greatly change our model…

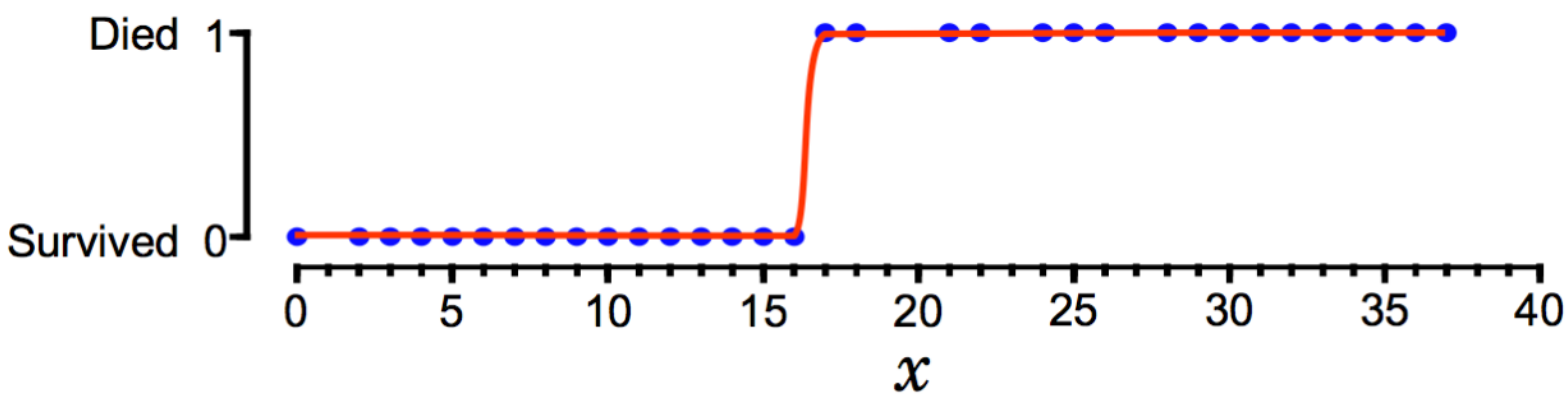


Source: http://www.mc.vanderbilt.edu/gcrc/workshop_files/2004-11-12.pdf

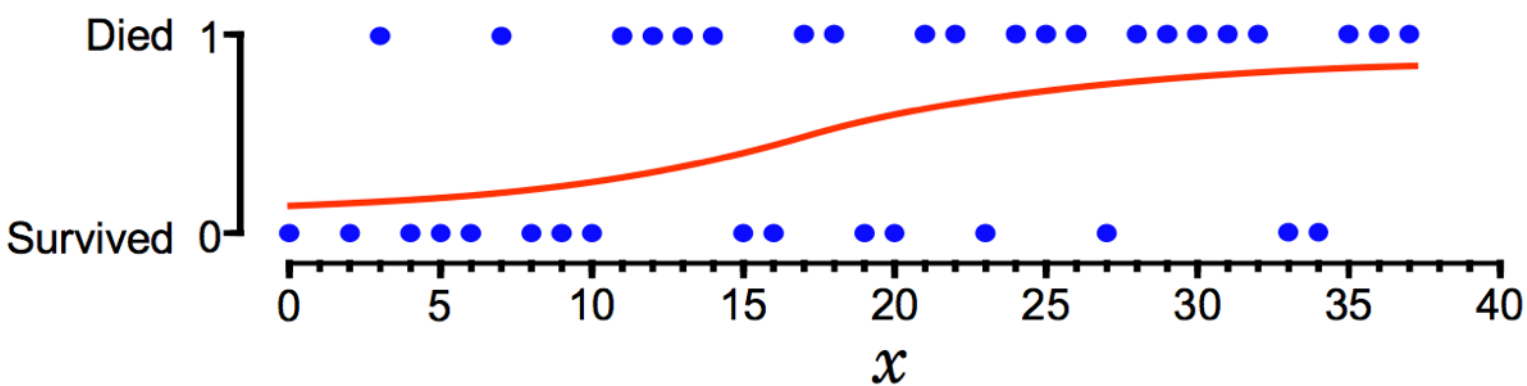
So, what can we do instead of linear regression?

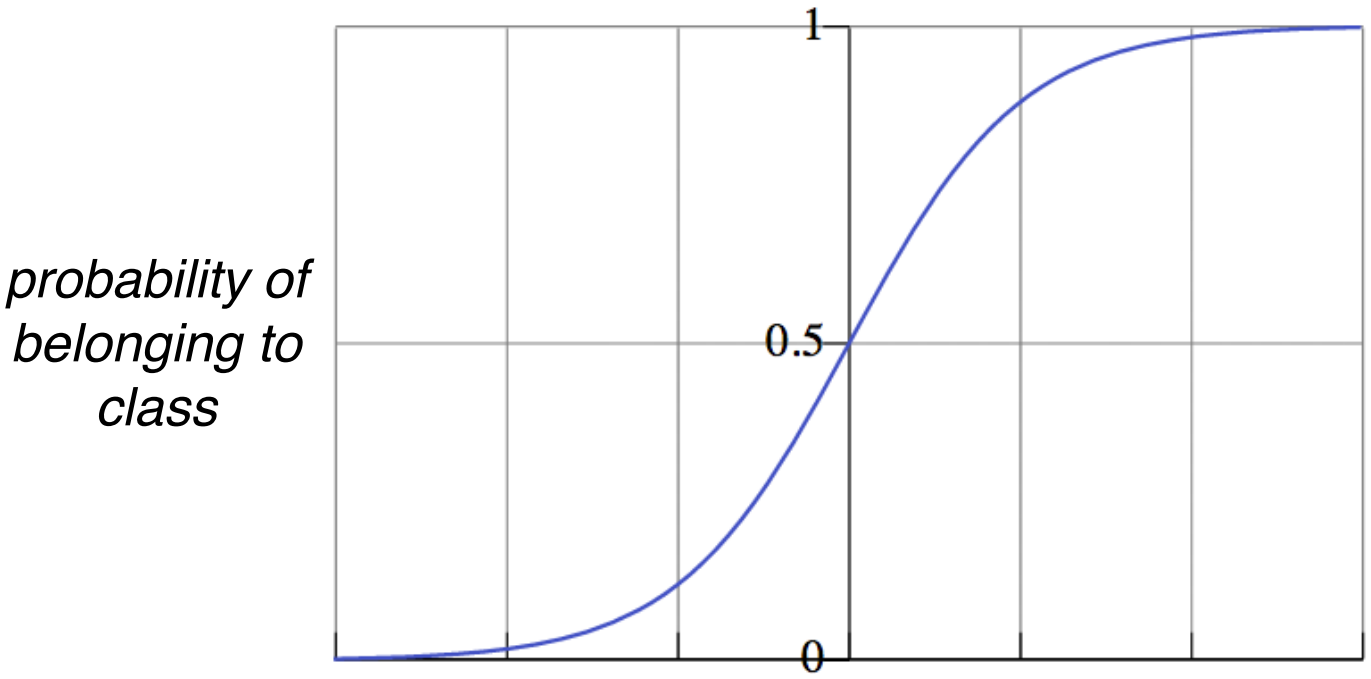


Going back to our example of patient survival given a sepsis test score: Data that has a sharp cut off point between the two classes (living / dying) should have a large value of B_1.



Going back to our example of patient survival given a sepsis test score: Data that has a lengthy transition between the two classes (living / dying) should have a small value of B_1.

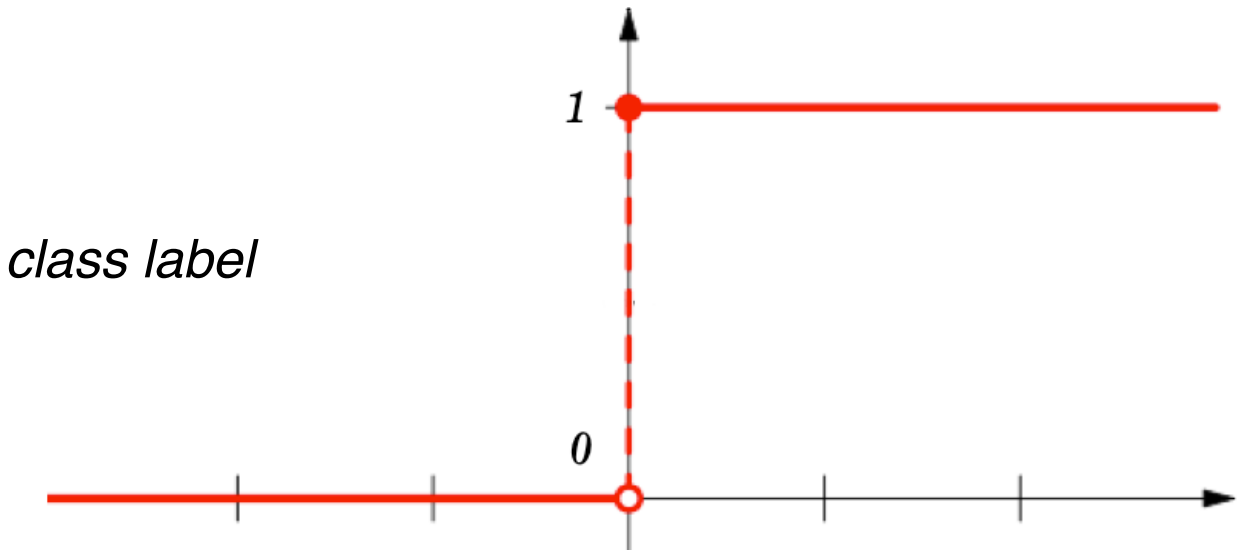




NOTE

Probability predictions look like this.

CLASS LABELS



NOTE

Probabilities are "snapped" to class labels (e.g. by thresholding at 50%).

The logistic regression model is an extension of the linear regression model, with a couple of important differences.

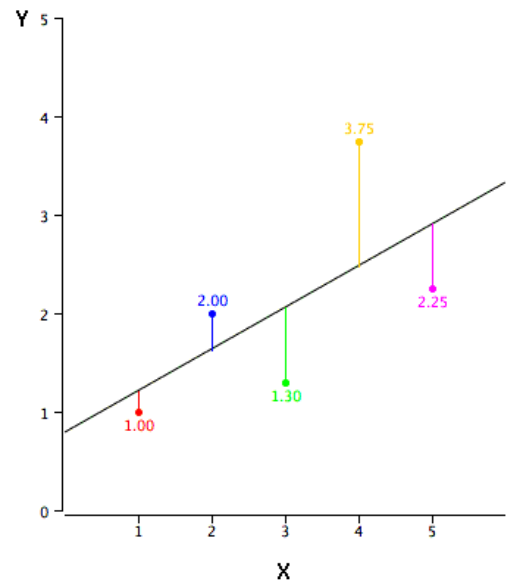
The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

II. OUTCOME VARIABLES

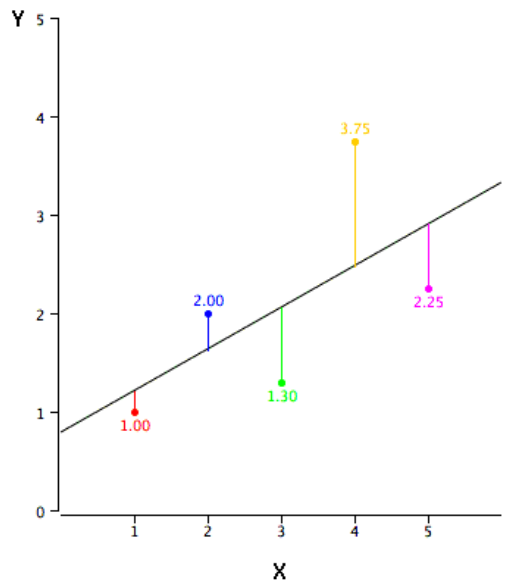
The key variable in any regression problem is the the outcome variable y *given the value of the covariate* x.

$$\mathrm{y} = \alpha + \beta_1 \mathrm{x}_1 + \ldots + \beta_n \mathrm{x}_n + \varepsilon$$



In linear regression, we assume that this outcome value is a linear function taking values in $(-\infty, +\infty)$*:*

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$



In logistic regression, we've seen that the outcome variable takes values only in the unit interval $[0, 1]$.

In logistic regression, we've seen that the outcome variable takes values only in the unit interval $[0, 1]$.

The first step in extending the linear regression model to logistic regression is to map the outcome value in $(-\infty, +\infty)$ *into the unit interval.*

In logistic regression, we've seen that the outcome variable takes values only in the unit interval $[0, 1]$.

The first step in extending the linear regression model to logistic regression is to map the outcome value in $(-\infty, +\infty)$ *into the unit interval.*

Q: How do we do this?

A: By using a transformation called the logistic function*:*

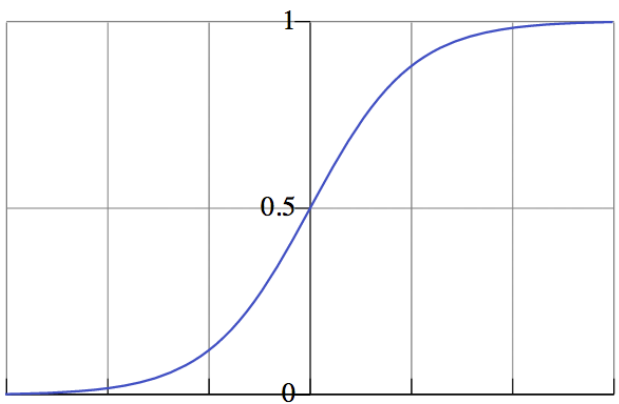
$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

THE LOGISTIC FUNCTION

A: By using a transformation called the logistic function*:*

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

We've already seen what this looks like:

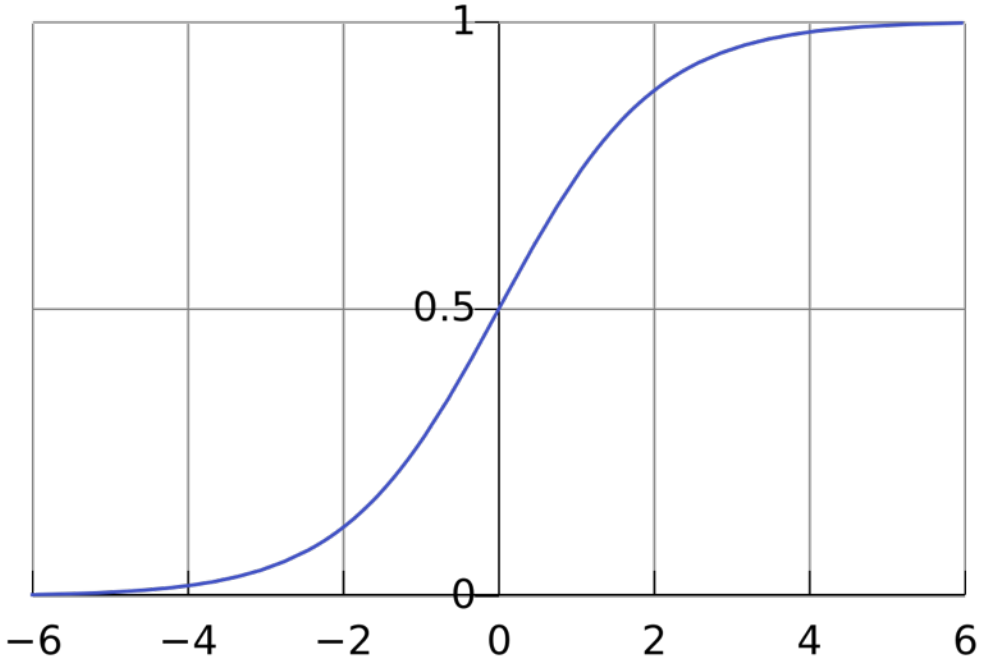


NOTE

For any value of x, y is in the interval [0, 1]

This is a nonlinear transformation!

probability of belonging to class



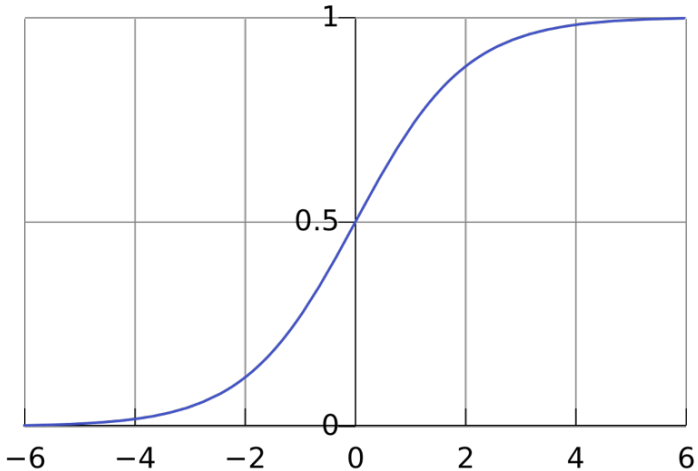
value of independent variable

NOTE

Probability predictions look like this.

This function fits our problem much better:

$$0 \leq h_\theta(x) \leq 1$$



In other words, our classifier will output values between 0 and 1. It asymptotically approaches 0 and 1.

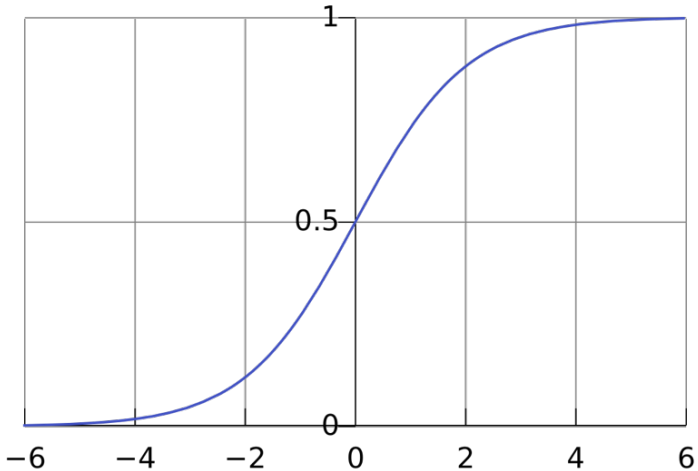
This is called the Sigmoid Function, or the Logistic Function (synonymous)

This function fits our problem much better:

$$0 \leq h_\theta(x) \leq 1$$

In other words, our classifier will output values between 0 and 1. It asymptotically approaches 0 and 1.

This is called the Sigmoid Function, or the Logistic Function (synonymous)

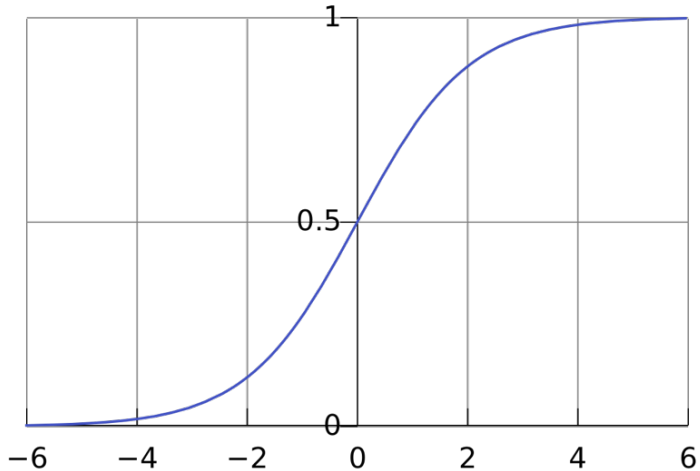


NOTE

This function gives Logistic Regression its name!

The logistic function:

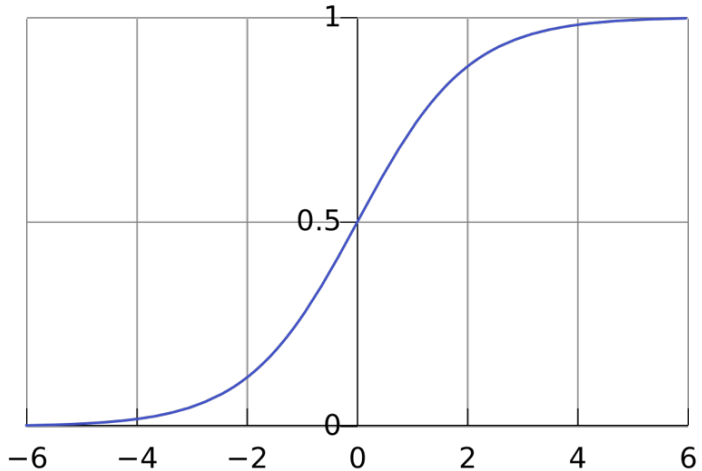
$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$



Notice that f(t) = 0.5 when t = 0
f(t) >= 0.5 when t > 0
f(t) <= 0.5 when t < 0

Suppose we predict class 1 when f(t) >= 0.5 and class 0 when f(t) < 0.5

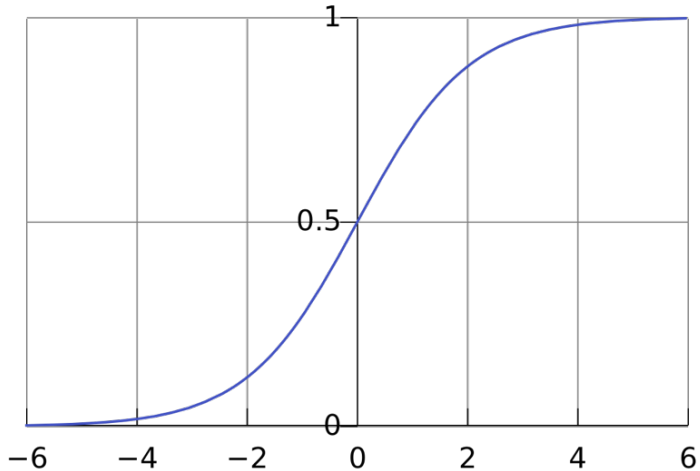
$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$



So, if the t in the logistic function is a linear function of an explanatory variable x, or a linear combination of explanatory variables, the logistic function becomes:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

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

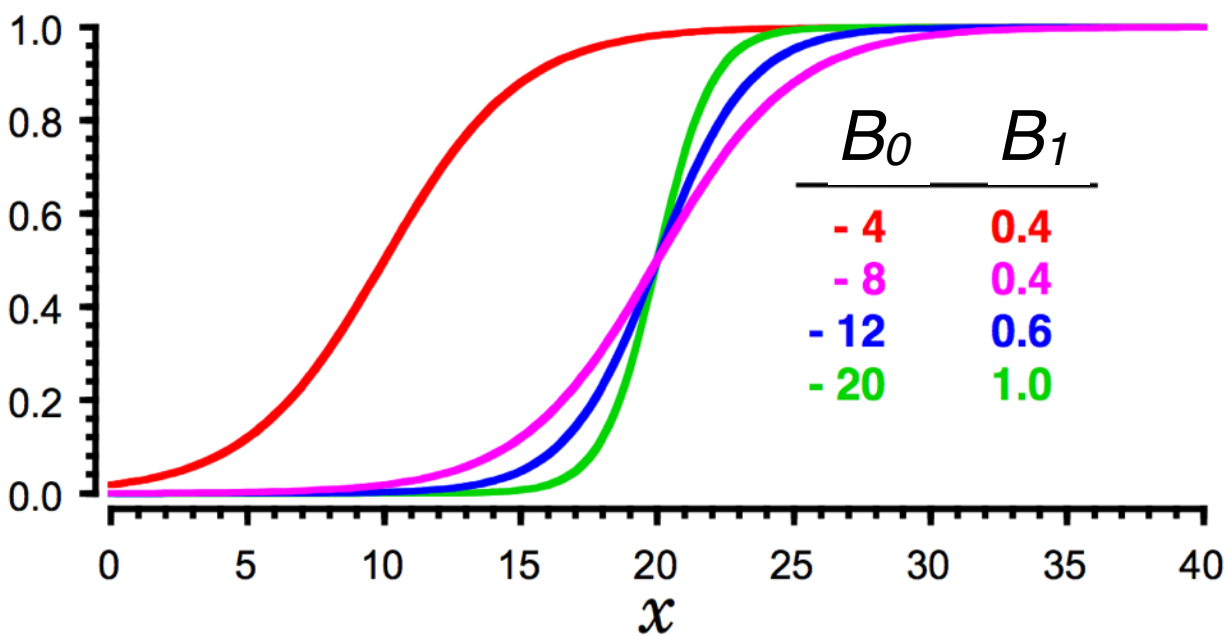


So, if the t in the logistic function is a linear function of an explanatory variable x, or a linear combination of explanatory variables, the logistic function becomes:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Does that exponent look familiar…?

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



When $B_0 + B_1x = 0$, then $F(x) = 0.5$, which is the inflection point on all these curves.

The logit function *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

The logit function *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

The logit function is also called the log-odds function.

THE LOGISTIC FUNCTION

The logit function *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

The logit function is also called the log-odds func

NOTE

This name hints at its usefulness in interpreting our results.

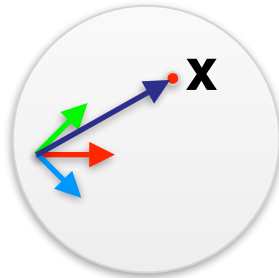
We will see why shortly.

III. HOW LOGISTIC REGRESSION WORKS

USING LOGISTIC FUNCTION

1. Model consists of a vector **β** in n-dimensional feature space

$$\boldsymbol{\beta} = \beta_1 + \beta_2 + \ldots + \beta_n$$



source: http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf

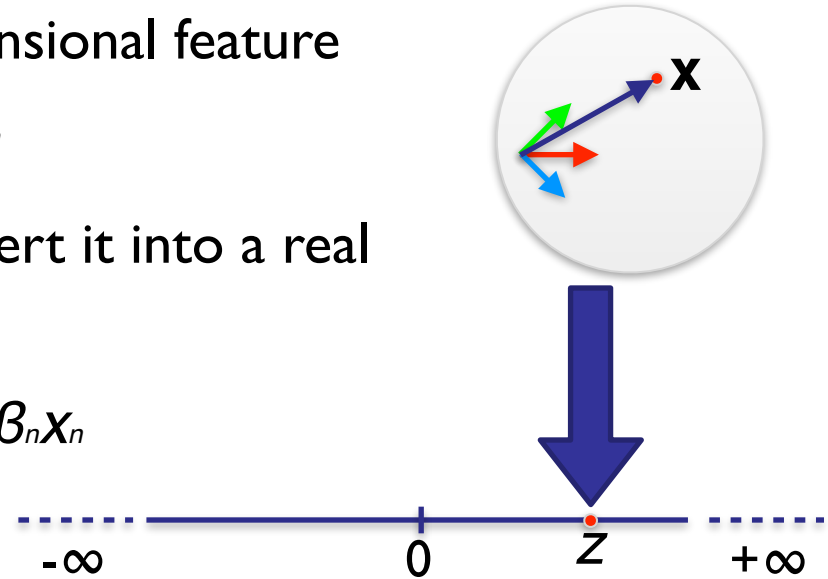
USING LOGISTIC FUNCTION

1. Model consists of a vector **β** in n-dimensional feature space

$$\boldsymbol{\beta} = \beta_1 + \beta_2 + \ldots + \beta_n$$

2. For a point **x**, project it onto β to convert it into a real number z in the range -∞ to +∞

$$z = \alpha + \boldsymbol{\beta}.\mathbf{x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_n x_n$$

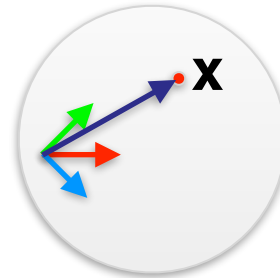


source: http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf

USING LOGISTIC FUNCTION

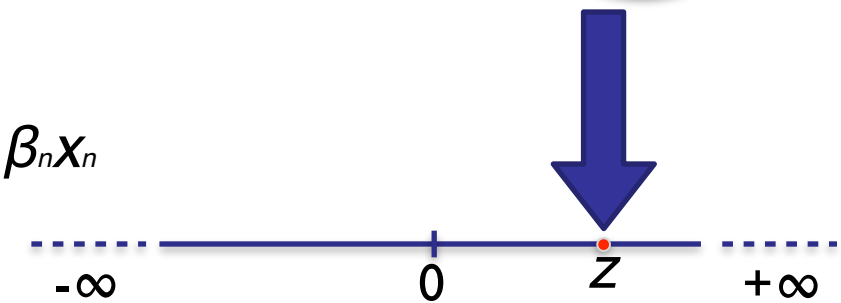
1. Model consists of a vector **β** in n-dimensional feature space

$$\boldsymbol{\beta} = \beta_1 + \beta_2 + \ldots + \beta_n$$



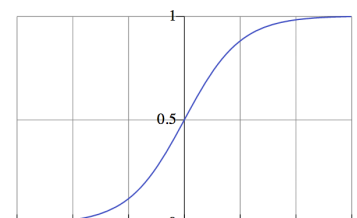
2. For a point **x**, project it onto β to convert it into a real number z in the range -∞ to +∞

$$z = \alpha + \boldsymbol{\beta}.\mathbf{x} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$



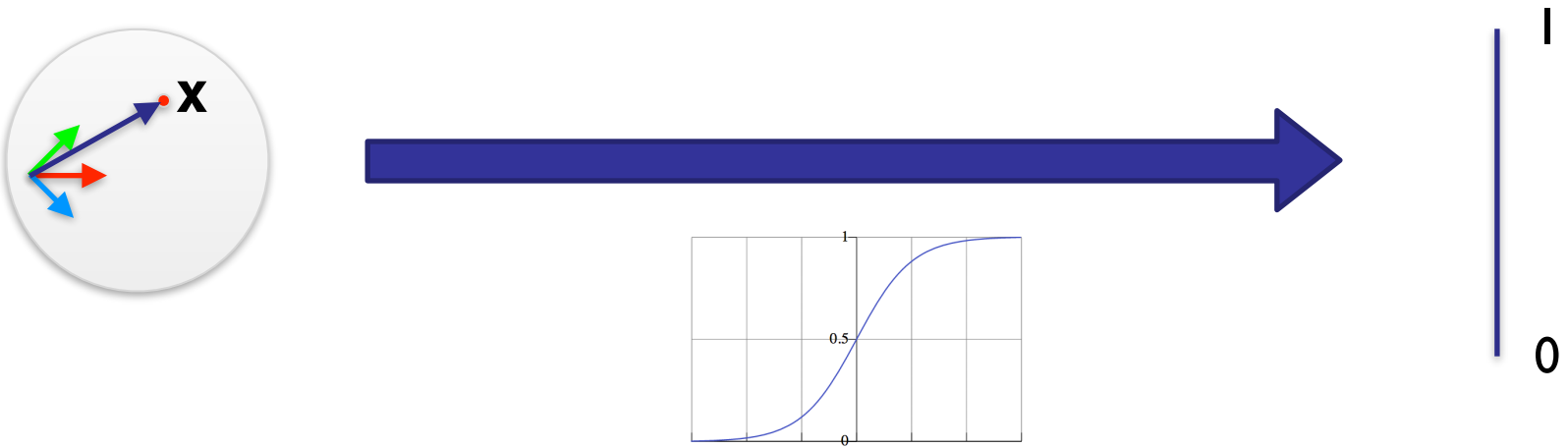
3. Map z to the range 0 to 1 using the logistic function

$$p = 1/(1+e^{-z})$$



source: http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf

USING LOGISTIC FUNCTION



Overall, logistic regression maps a point x in n-dimensional feature space to a value in the range 0 to 1.

source: http://courses.washington.edu/css490/2012.Winter/lecture_slides/05b_logistic_regression.pdf

PREDICTED VALUE

prediction from a logistic regression model as:

A probability of class membership

Need to optimize β so the model gives the best possible reproduction of training set labels

(Usually done by numerical approximation of maximum likelihood)

The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

The second difference is in the error terms.

III. ERROR TERMS

The second difference between linear regression and the logistic regression model is in the error term.

The second difference between linear regression and the logistic regression model is in the error term.

One of the key assumptions of linear regression is that the error terms follow independent Gaussian distributions with zero mean and constant variance:

$$\epsilon \sim N(0, \sigma^2)$$

In logistic regression, the outcome variable can take only two values: 0 or 1.

In logistic regression, the outcome variable can take only two values: 0 or 1.

It's easy to show from this that instead of following a Gaussian distribution, the error term in logistic regression follows a Bernoulli distribution:

$$\epsilon \sim B(0, \pi(1-\pi))$$

In logistic regression, the outcome variable can take only two values: 0 or 1.

It's easy to show from this that instead of following a Gaussian distribution, the error term in logistic regression follows a Bernoulli distribution:

$$\epsilon \sim B(0, \pi(1-\pi))$$

NOTE

This is the same distribution followed by a coin toss.

Think about why this makes sense!

These two key differences define the logistic regression model, and they also lead us to a kind of unification of regression techniques called generalized linear models*.*

These two key differences define the logistic regression model, and they also lead us to a kind of unification of regression techniques called generalized linear models*.*

Briefly, GLMs generalize the distribution of the error term, and allow the conditional mean of the response variable to be related to the linear model by a link function*.*

In the present case, the error term follows a Bernoulli distribution, and the logit is the link function that connects us to the linear predictor.

In the present case, the error term follows a Bernoulli distribution, and the logit is the link function that connects us to the linear predictor.

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

IV. INTERPRETING RESULTS

In linear regression, the parameter β represents the change in the response variable for a unit change in the covariate.

In linear regression, the parameter β represents the change in the response variable for a unit change in the covariate.

In logistic regression, β represents the change in the logit function for a unit change in the covariate.

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

In order to interpret the outputs of a logistic function we must understand the difference between probability and odds.

The odds of an event are given by the ratio of the probability of the event by its complement:

$$Odds = \frac{\pi}{1-\pi}$$

Quiz: You're trying to determine whether a customer will convert or not. The customer conversion rate is 33.33%. what are the odds that a customer will convert?

$$Odds = \frac{\pi}{1-\pi} = \frac{.3333}{.6666} = \frac{1}{2}$$

Quiz: You're trying to determine whether a customer will convert or not. The customer conversion rate is 33.33%. what are the odds that a customer will convert?

$$Odds = \frac{\pi}{1-\pi} = \frac{.3333}{.6666} = \frac{1}{2}$$

NOTE

means for every customer that converts you will have two customers that do not convert

The odds ratio *of a binary event is given by the odds of the event divided by the odds of its complement:*

$$OR = \frac{O(x{=}1)}{O(x{=}0)} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

The odds *of an event are given by the ratio of the probability of the event by its complement:*

$$O(x = 1) = \frac{\pi(1)}{(1-\pi(1))}$$

The odds *of an event are given by the ratio of the probability of the event by its complement:*

$$O(x=1) = \frac{\pi(1)}{(1-\pi(1))}$$

The odds ratio *of a binary event is given by the odds of the event divided by the odds of its complement:*

$$OR = \frac{O(x=1)}{O(x=0)} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

Substituting the definition of $\pi(x)$ into this equation yields (after some algebra),

$$OR = e^{\beta}$$

Substituting the definition of $\pi(x)$ into this equation yields (after some algebra),

$$OR = e^{\beta}$$

This simple relationship between the odds ratio and the parameter β is what makes logistic regression such a powerful tool.

Q: So how do we interpret this?

Q: So how do we interpret this?

A: The odds ratio of a binary event gives the increase in likelihood of an outcome if the event occurs.

$$OR = e^{\beta}$$

INTERPRETING RESULTS – AN EXAMPLE

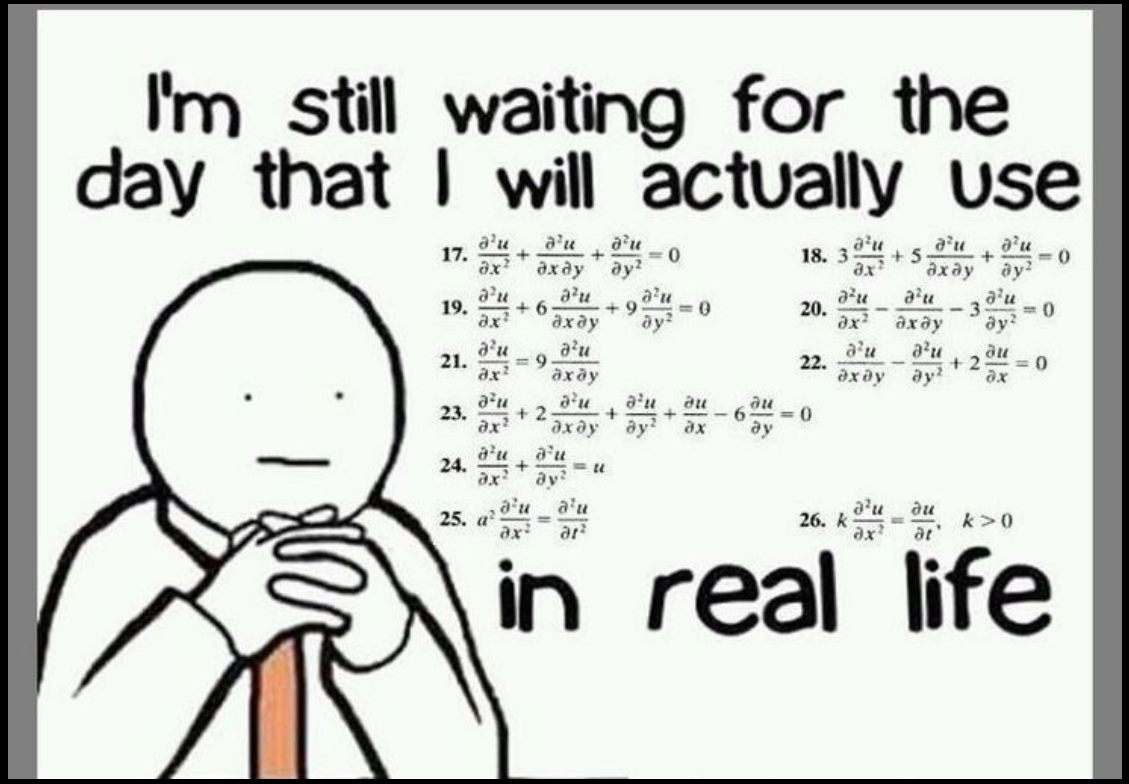
Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote a mobile OS (for example, iOS).

Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote a mobile OS (for example, iOS).

In this case, eg, β = log(2)=0.693 means an odds ratio of 2

indicates that a purchase is twice as likely for an iOS user as for a non-iOS user.

Logistic Regression



LAB: LOGISTIC REGRESSION