GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 6: LOGISTIC REGRESSION

YUCHEN ZHAO / DAT-14

LAST TIME:

- LINEAR REGRESSION
- POLYNOMIAL REGRESSION
- REGULARIZATION

QUESTIONS?

I. LOGISTIC REGRESSION

- *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	???	???
unsupervised	???	???

	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: Why is **logistic regression** *useful?*

A: A large number of commercially valuable classification problems can be addressed with logistic regression, including:

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

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 $f: X \longrightarrow \{0, 1\}$

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

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.

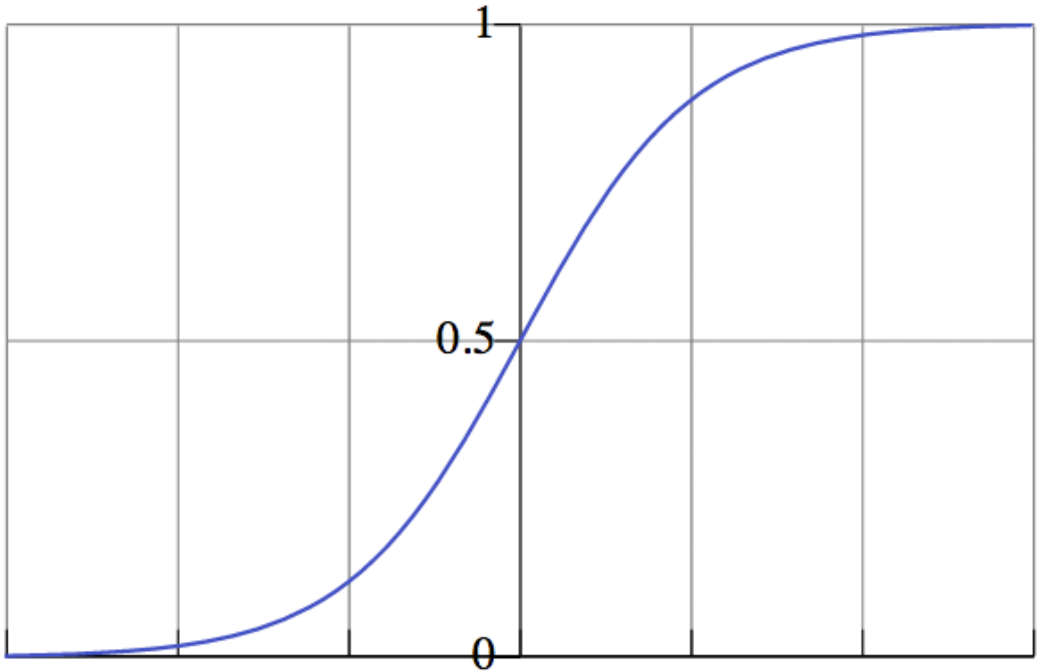
In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.

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.

PROBABILITIES

probability of belonging to class



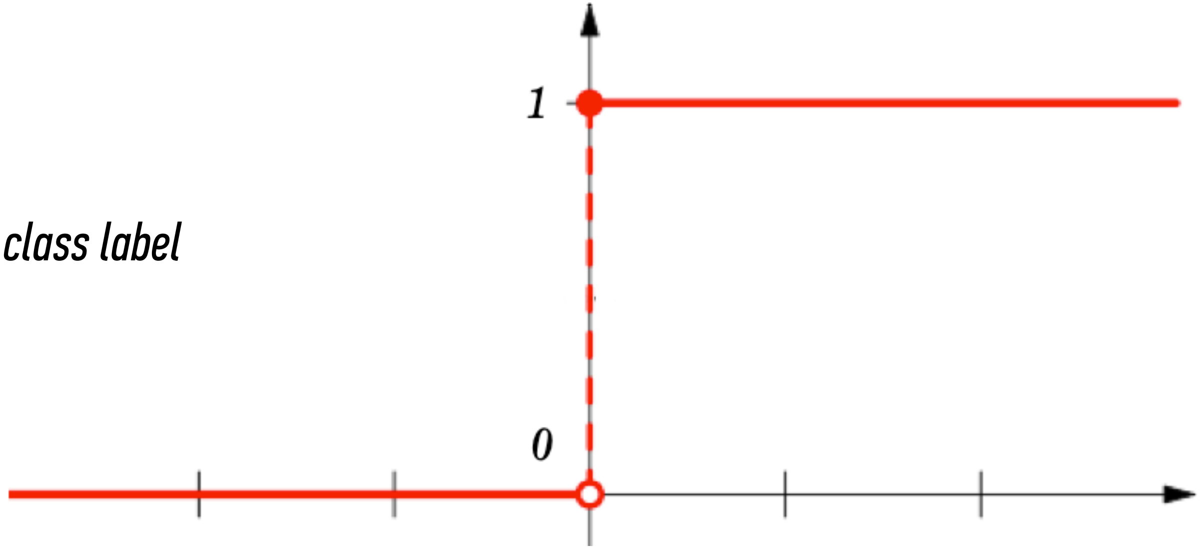
value of independent variable

NOTE

Probability predictions look like this.

CLASS LABELS

class label



value of independent variable

NOTE

Probabilities are "snapped" to class labels (eg by threshholding at 50%).

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

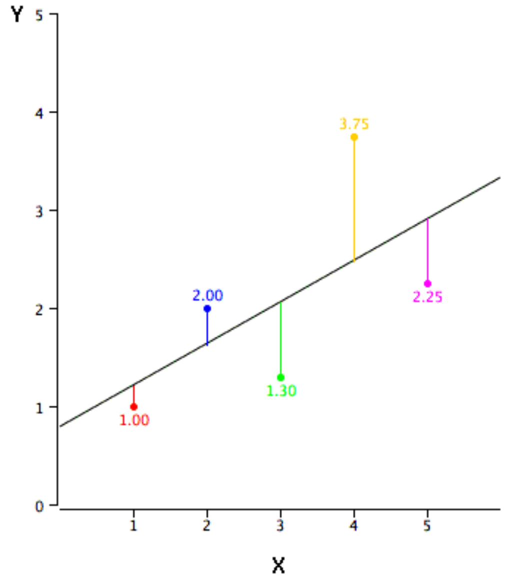
The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

II. OUTCOME VARIABLES

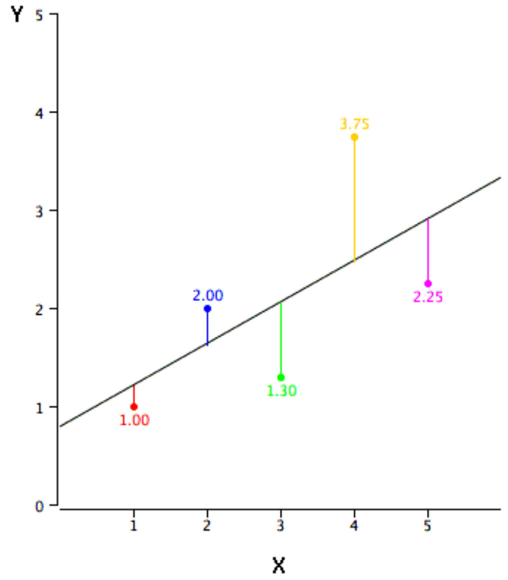
The key variable in any regression problem is the the outcome variable y *given the value of the covariate* x.

$$\mathbf{y} = \alpha + \beta_1 \mathbf{x}_1 + \ \dots \ + \beta_n \mathbf{x}_n + \varepsilon$$



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

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



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 conditional mean of 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 conditional mean of 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?

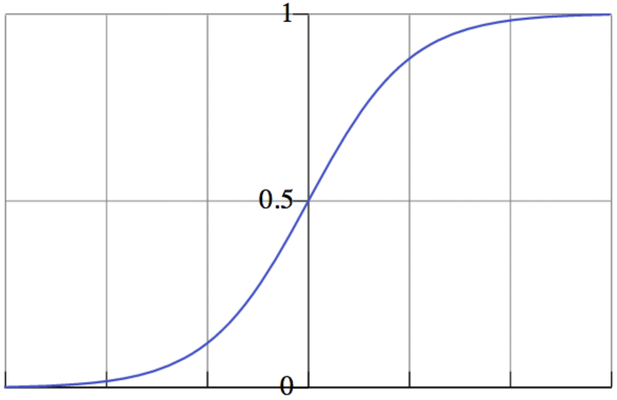
A: By using a transformation called the **logistic function***:*

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

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

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

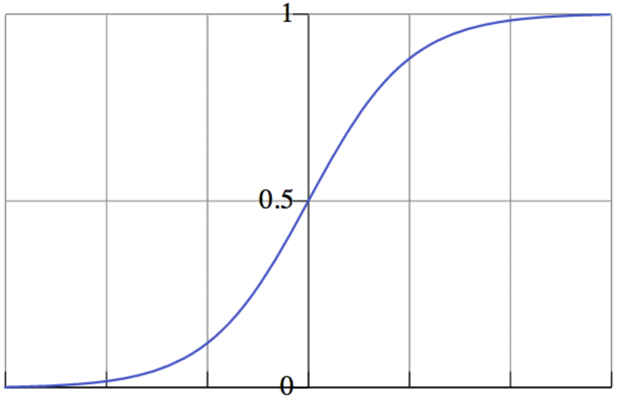
We've already seen what this looks like:



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

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

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!

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

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

NOTE

This name hints at its usefulness in interpreting our results.

We will see why shortly.

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

III. HOW LOGISTIC REGRESSION WORKS

USING LOGISTIC FUNCTION

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

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

$$z = \alpha + \boldsymbol{\beta} \cdot \mathbf{x} = \alpha + \beta_1 x_1 + \cdots + \beta_d x_d$$

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

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

Overall, logistic regression maps a point x in d-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 prediction from a logistic regression model as:

A probability of class membership

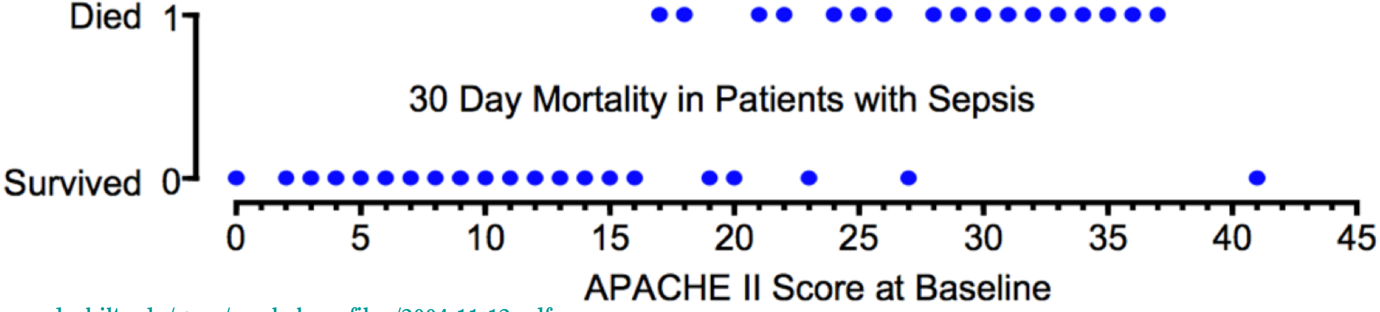
Need to optimize β so the model gives the best possible reproduction of training set labels

(Usually done by numerical approximation of maximum likelihood)

AN APPLICATION

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.

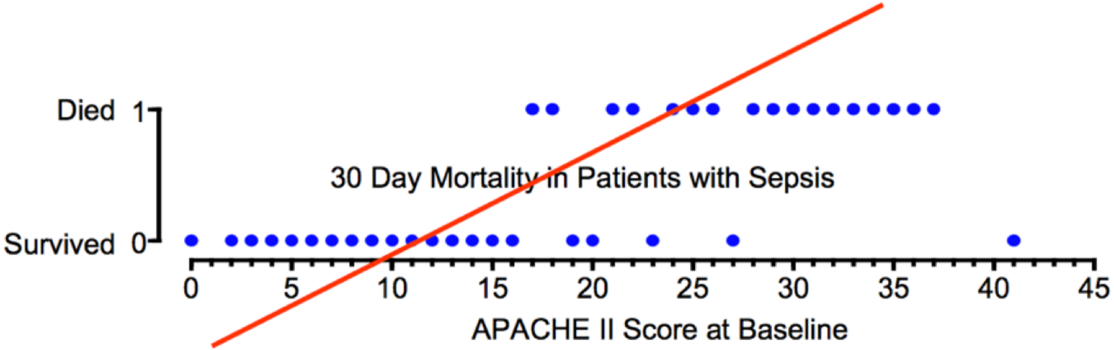
How can we predict death from baseline APACHE II score in these patients?



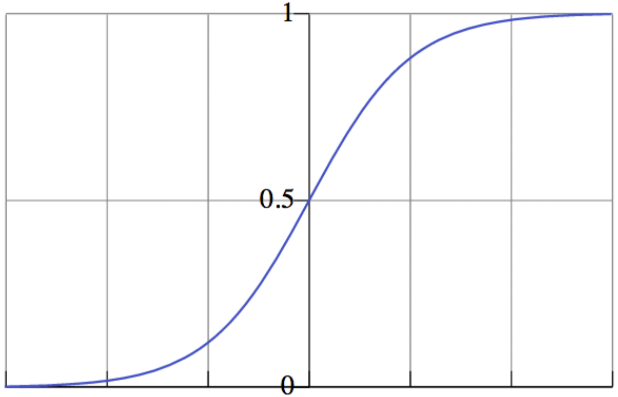
source: http://www.mc.vanderbilt.edu/gcrc/workshop_files/2004-11-12.pdf

Q: How can we predict death probability from baseline APACHE II score in these patients? Will linear regression work?

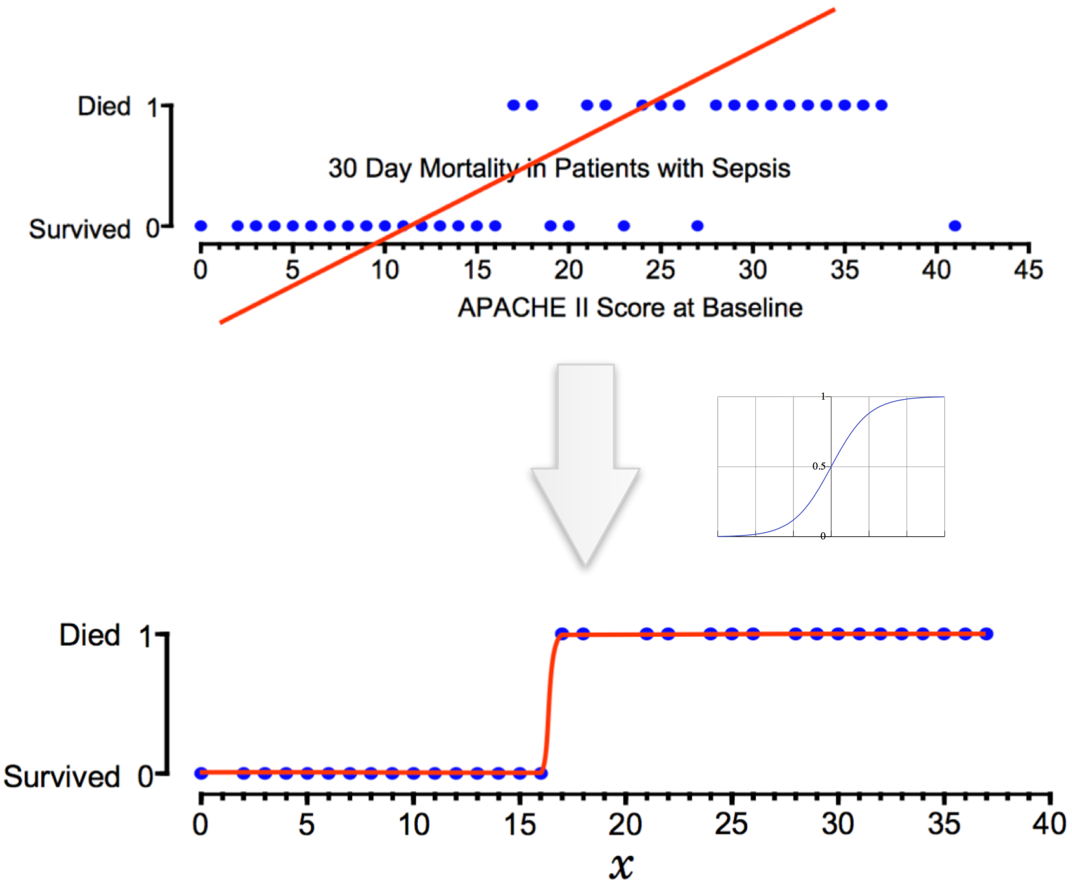
Well, linear regression would not work well here, because it could produce probabilities less than zero or greater than one. Also, one new value could greatly change our model...



source: http://www.mc.vanderbilt.edu/gcrc/workshop_files/2004-11-12.pdf remember logistic function?



AN APPLICATION

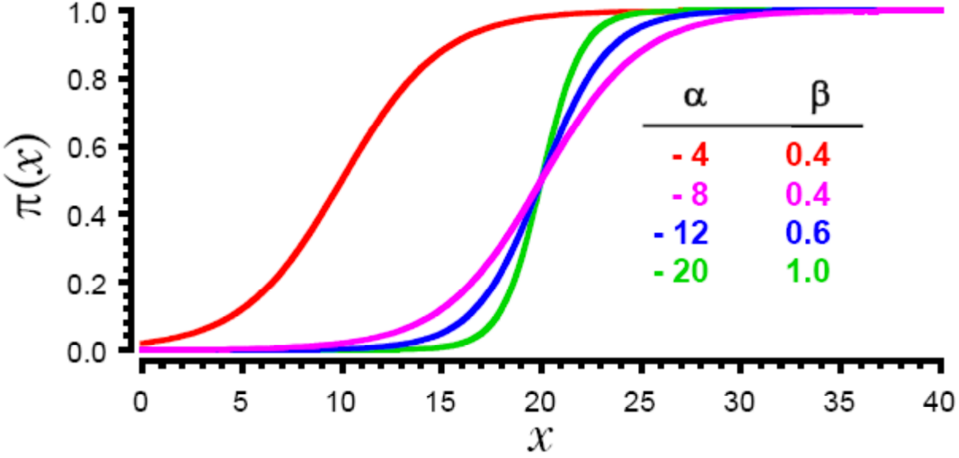


Parameters control shape and location of the curve:

– α controls location of midpoint

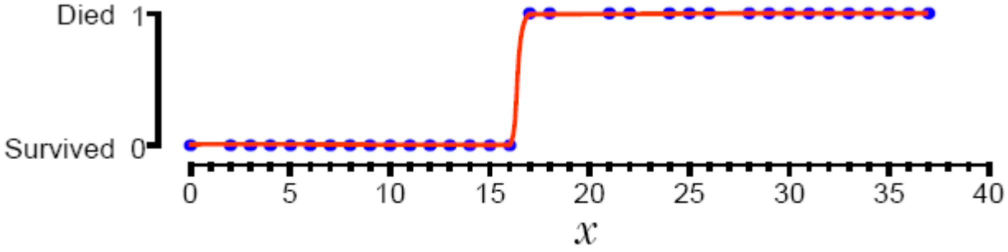
– β controls slope of rise

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

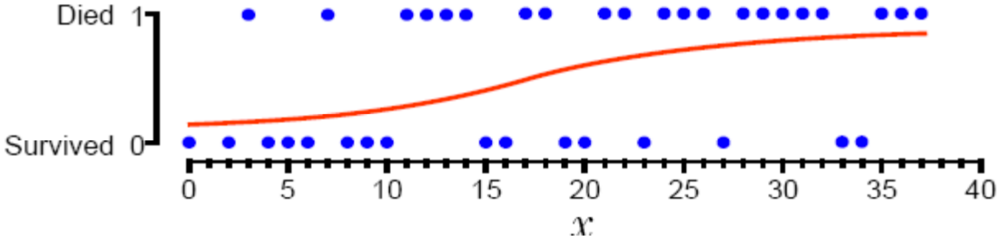


AN APPLICATION

Data that has a sharp survival cut off point between patients who live or die should have a large value of β.



Data with a lengthy transition from survival to death should have a low value of β.



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 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.

Interpreting this change in the logit function requires another definition first.

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:*

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

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

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

Substituting the definition of $\pi(\mathbf{x})$ 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).

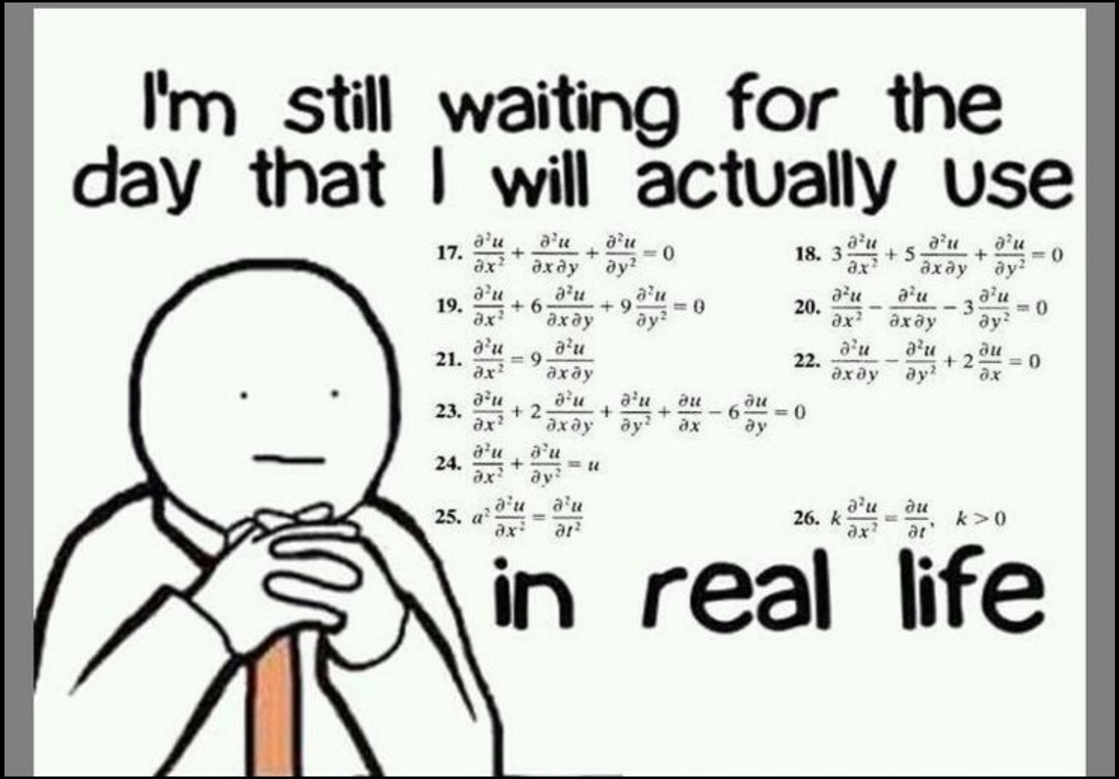
Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote a mobile OS (for example, iOS).

In this case, eg, $\beta = \log(2) = 0.693$ *means an odds ratio of 2*

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

LAB: LOGISTIC REGRESSION

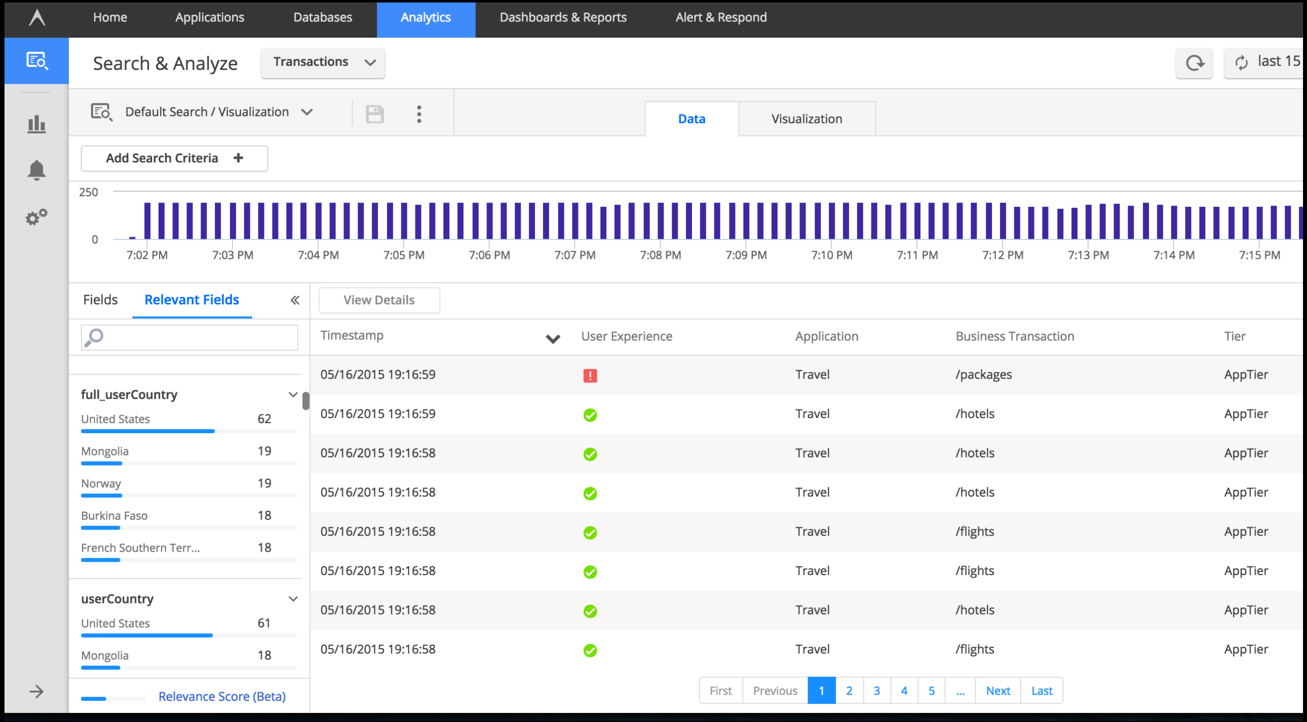
Logistic Regression



Classification Discussion

Logistic Function Application

Normalize Scores



Logistic Regression Application

Predicting Job Changes