GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
SESSION 9: LOGISTIC REGRESSION

Rob Hall
DAT13 SF // April 6, 2015

LAST TIME:

- FINAL PROJECT ELEVATOR PITCHES
- CLUSTERING WITH K-MEANS

QUESTIONS?

AGENDA

I. OVERVIEW
II. BASIC FORM
III. INTERPRETATION
IV. LAB
V. Q&A

I. OVERVIEW

	continuous	*categorical*
supervised	???	???
unsupervised	???	???

Q: Where does logistic regression belong in this diagram?

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

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

*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...

Q: ***What*** *is logistic regression?*

Q: ***What*** *is logistic regression?*

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

In linear regression, we used a set of input variables to predict the value of a continuous response variable.

In linear regression, we used a set of input variables to predict the value of a continuous response variable.

In logistic regression, we use a set of input variables to predict probabilities of class membership.

In linear regression, we used a set of input variables to predict the value of a continuous response variable.

In logistic regression, we use a set of input variables to predict probabilities of class membership.

NOTE

Class membership is not always binary, however, that is what we will focus on for this class.

In linear regression, we used a set of input variables to predict the value of a continuous response variable.

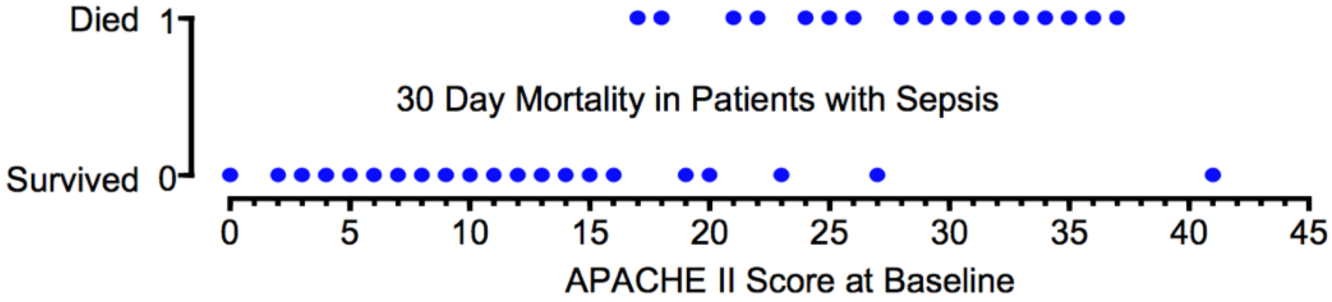
In logistic regression, we use a set of input variables to predict probabilities of class membership.

These probabilities can then mapped to class labels, thus predicting the class for each observation.

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?

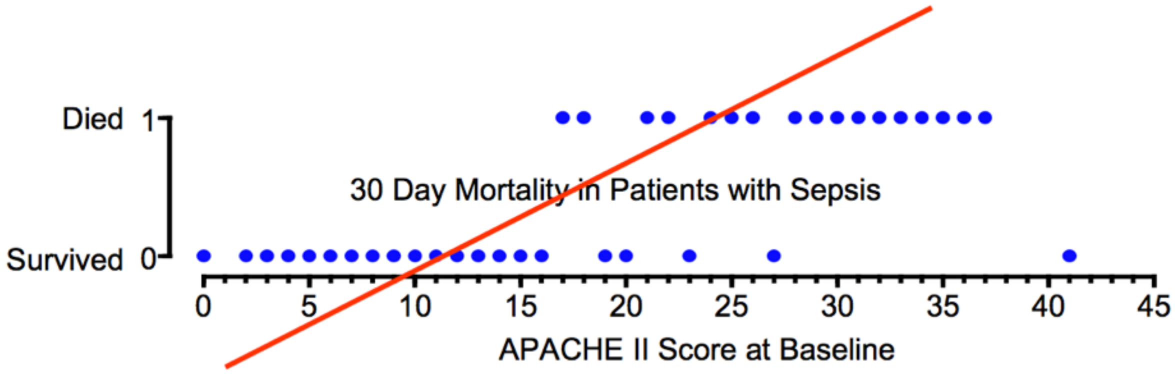


Source: http://www.mc.vanderbilt.edu/gcrc/workshop_files/2004-11-12.pdf

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

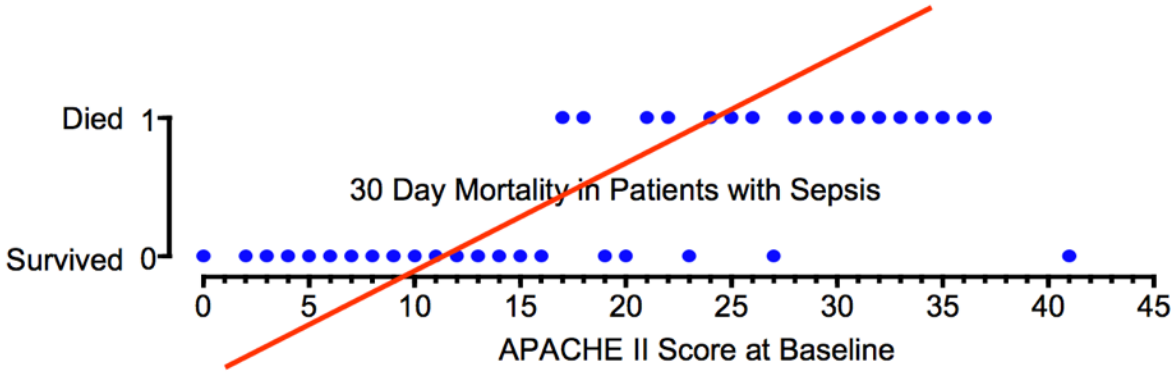
Let p(x) be the probability that a patient with score x will die within 30 days.

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?



II. BASIC FORM

When performing linear regression, we use the following function:

$$y = \beta_0 + \beta_1 x$$

When performing logistic regression, we use the following form:

$$\pi = \Pr(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

When performing linear regression, we use the following function:

$$y = \beta_0 + \beta_1 x$$

When performing logistic regression, we use the following form:

$$\pi = \Pr(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Probability of y = 1, given x

Quiz: *Create a plot of the logistic function.*

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

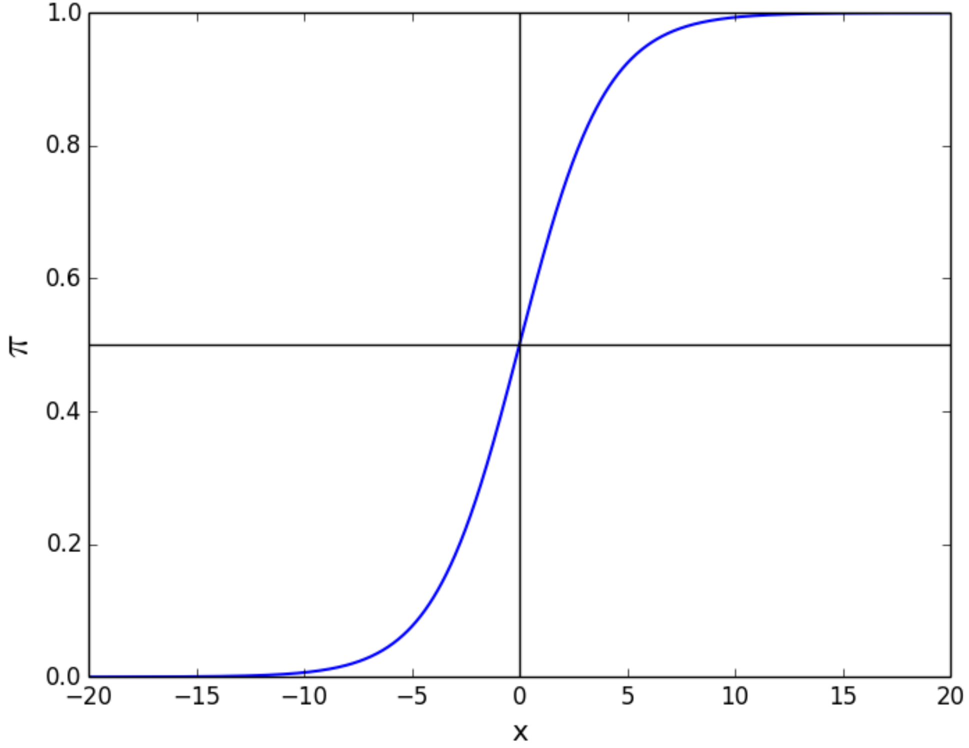
Quiz: *Create a plot of the logistic function.*

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

How would you describe the shape of the function?

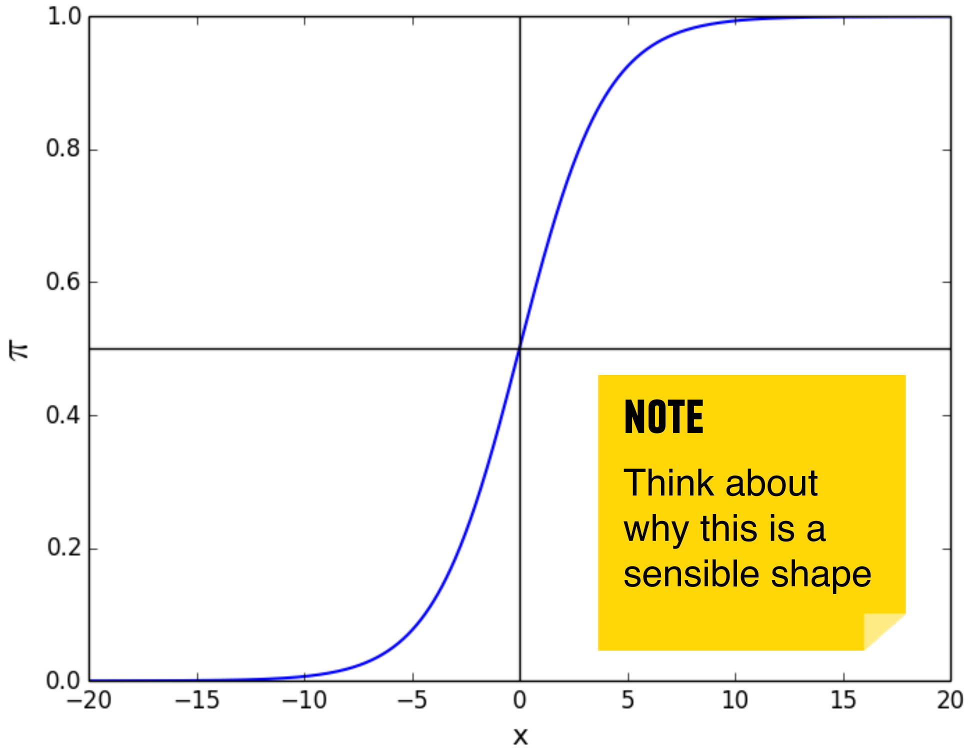
The logistic function takes on an "S" shape, where y is bounded by [0, 1]

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



The logistic function takes on an "S" shape, where y is bounded by [0, 1]

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

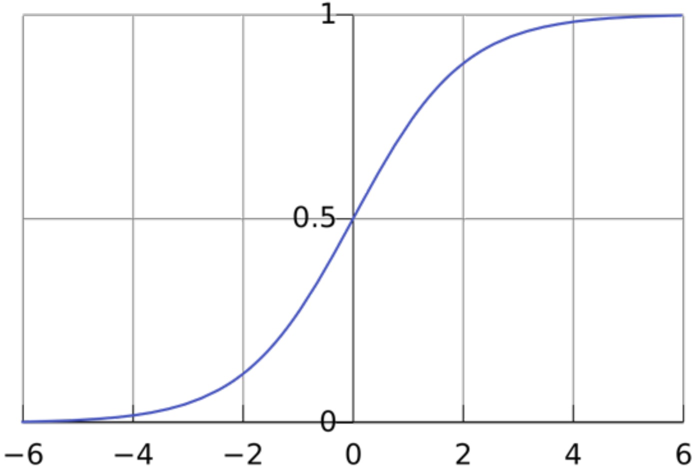


NOTE

Think about why this is a sensible shape

This function fits our problem much better:

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

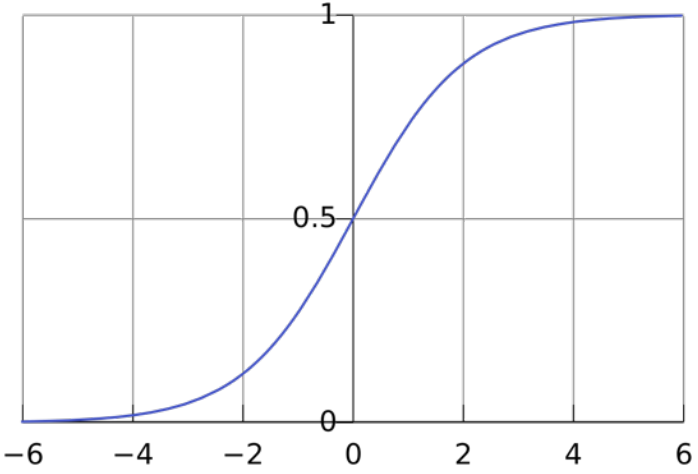


In other words, our classifier will output values between 0 and 1. It asymptotically approaches 0 and 1.

This is called the Sigmoid Function, or the <u>Logistic</u> Function (synonymous)

This function fits our problem much better:

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



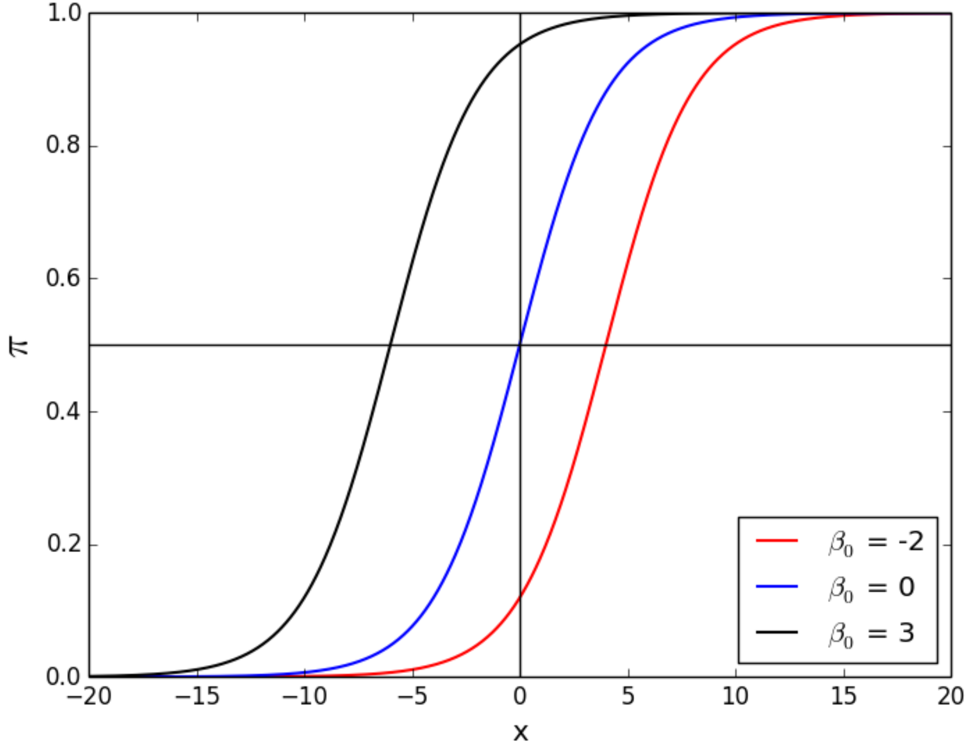
In other words, our classifier will output values between 0 and 1. It asymptotically approaches 0 and 1.

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

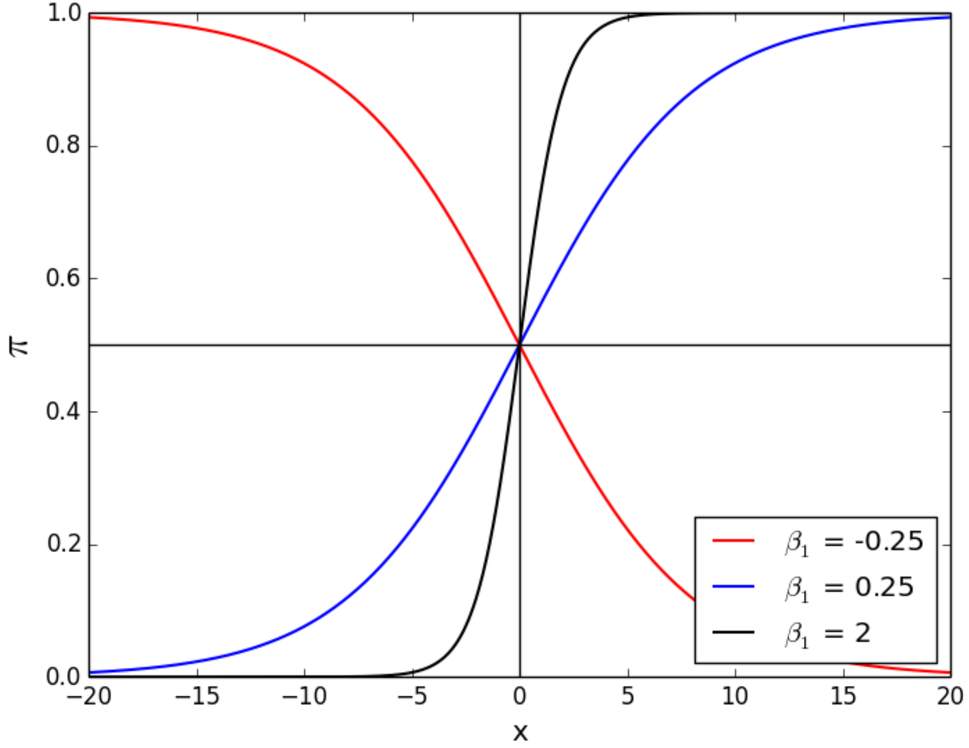
NOTE

This function gives Logistic Regression its name!

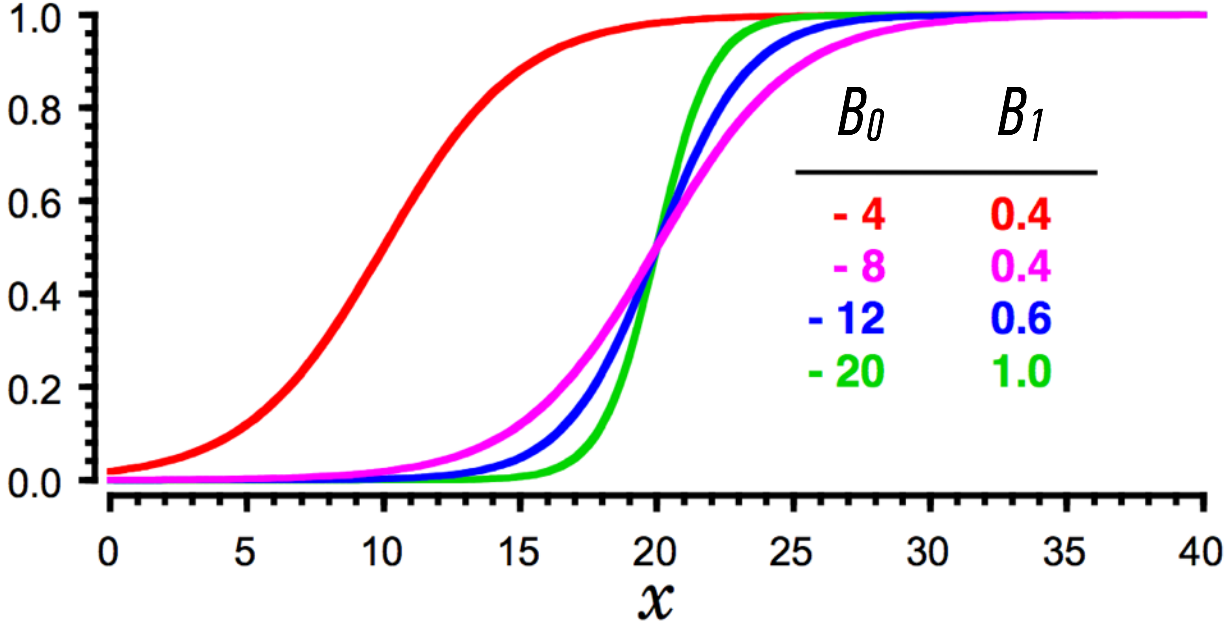
Changing the β_0 value shifts the function horizontally.



Changing the β_1 value changes the slope of the curve

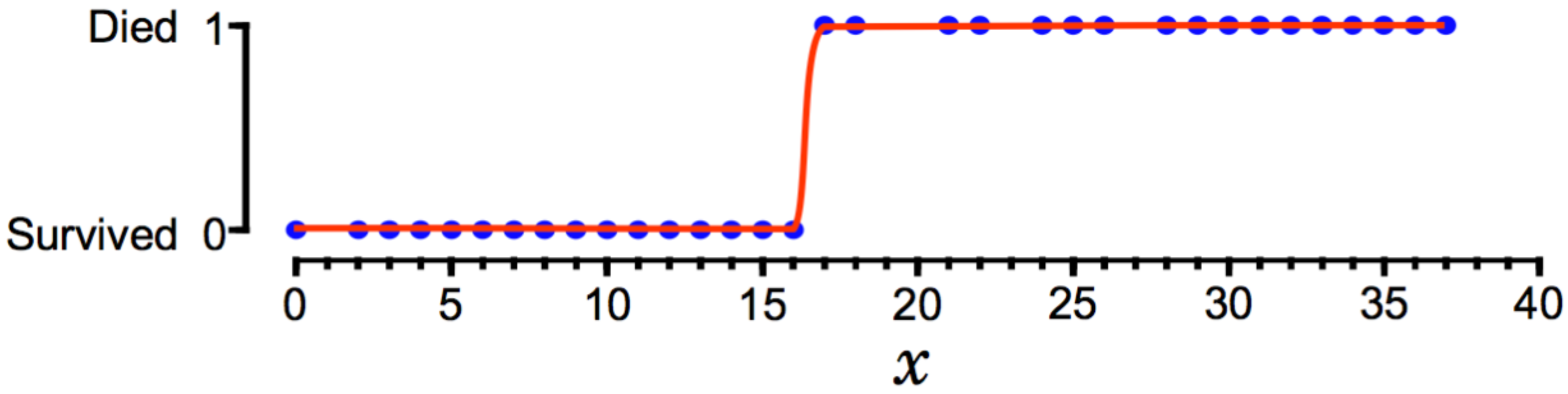


$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

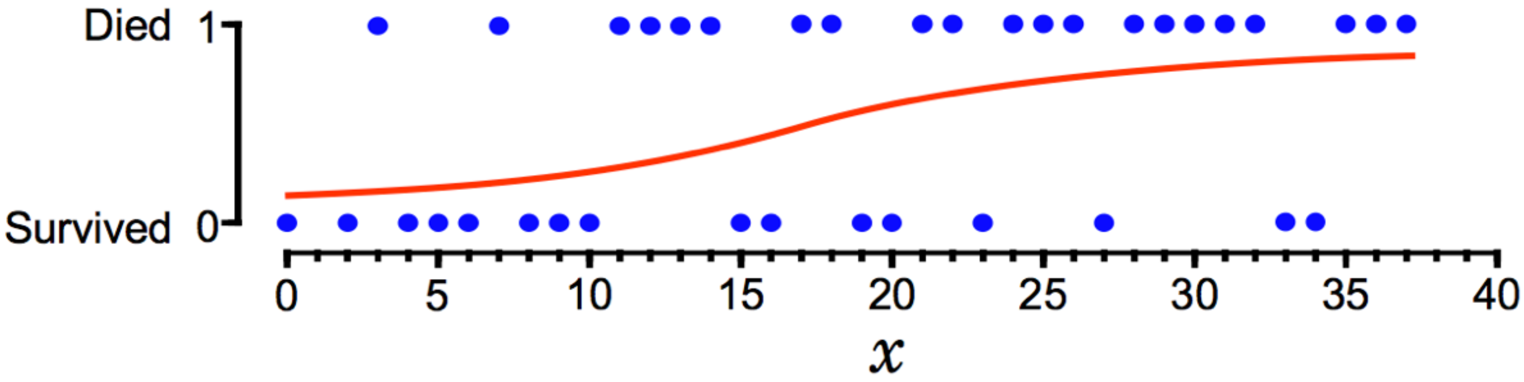


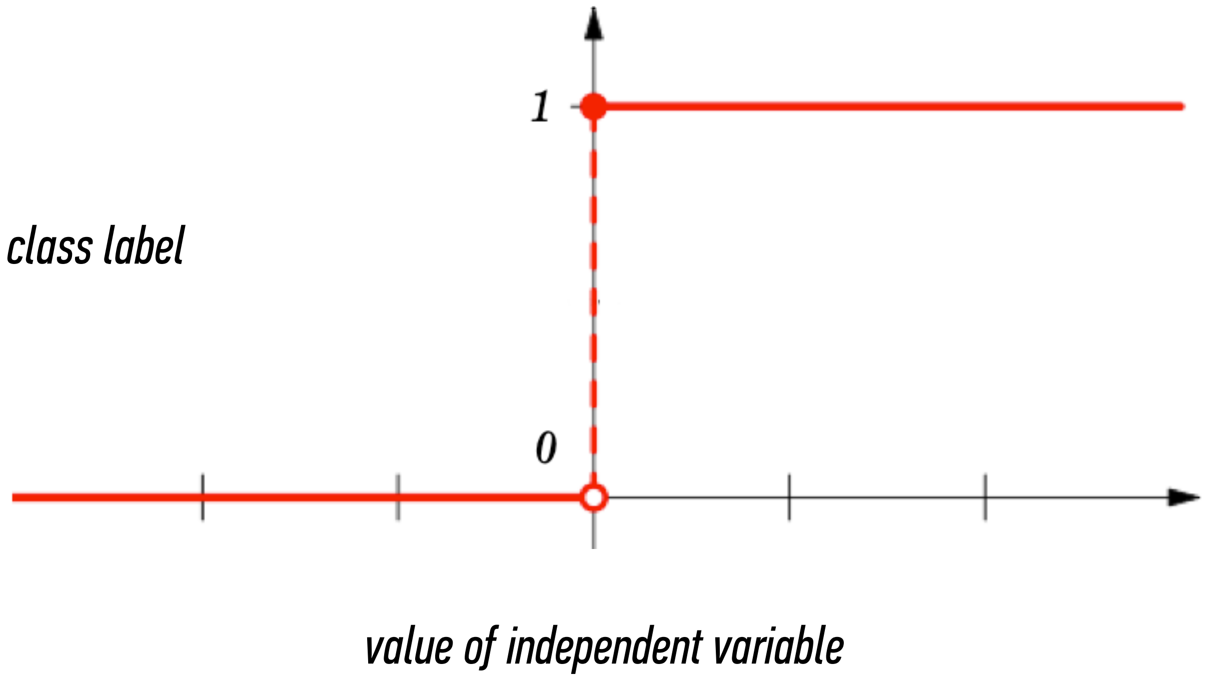
When $B_0 + B_1x = 0$, then $F(x) = 0.5$, which is the inflection point on all these curves.

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

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

III. INTERPRETATION

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

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:

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:

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

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:

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

QUESTION

What is the range of the odds?

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?*

Take 2 minutes and work this out.

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?*

Take 2 minutes and work this out.

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

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?*

Take 2 minutes and work this out.

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

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?*

Take 2 minutes and work this out.

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

NOTE

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

What would happen if we took the odds of the logistic function?

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

What would happen if we took the odds of the logistic function?

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

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

What would happen if we took the odds of the logistic function?

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

Does that exponent look familiar...?

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

Notice if we take the logarithm of the odds, we return a linear equation

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x$$

Notice if we take the logarithm of the odds, we return a linear equation

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x$$

NOTE

What is the range of the logit function?

Notice if we take the logarithm of the odds, we return a linear equation

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(e^{\beta_0 + \beta_1 x}\right) = \beta_0 + \beta_1 x$$

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

*In linear regression, the parameter β_1 represents the change in the **response variable** for a unit change in x.*

In linear regression, the parameter β_1 represents the change in the ***response variable*** *for a unit change in x.*

In logistic regression, β_1 represents the change in the ***log-odds*** *for a unit change in x.*

In linear regression, the parameter β_1 represents the change in the ***response variable*** *for a unit change in x.*

In logistic regression, β_1 represents the change in the ***log-odds*** *for a unit change in x.*

This means that e^{β_1} gives us the change in the ***odds*** *for a unit change in x.*

Q: How to determine whether a coefficient is significant?

A: This is based off of the p-value, just as with the linear regression

Example: *Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote whether phone was an iPhone.*

Example: *Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote whether phone was an iPhone.*

We perform a logistic regression, and we get $\beta_1 = 0.693$.

Example: *Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote whether phone was an iPhone.*

We perform a logistic regression, and we get $\beta_1 = 0.693$.

Q: What does this mean?

Example: *Suppose we are interested in mobile purchase behavior. Let y be a class label denoting purchase/no purchase, and let x denote whether phone was an iPhone.*

We perform a logistic regression, and we get β_1 *= 0.693.*

In this case the odds ratio is exp(0.693) = 2, meaning the likelihood of purchase is twice as high if the phone is an iPhone.

Once we understand the basic form for logistic regression, we can easily extend the definition to include multiple input values.

Logit function

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Once we understand the basic form for logistic regression, we can easily extend the definition to include multiple input values.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Logistic function

$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}$$

IV. LAB

V. Q&A

Q: What is a Generalized Linear Model (GLM)?

A: 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***.*

Q: What is the error distribution and link function for the logistic regression?

A: The error term follows a Bernoulli distribution, and the logit is the link function that connects us to the linear predictor.

Q: Is the logit the only link function used for the Bernoulli distribution?

A: No, other link functions include the probit the tobit model. However, the logit simplifies things nicely and is probably the most commonly used.

Q: What is the difference between $\dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ *and* $\dfrac{1}{1+e^{-\beta_0-\beta_1 x}}$ *?*

A: Nothing, these are equivalent expressions.

If you want to prove this to yourself (a) plot both equations, or (b) multiply both numerator and denominator by $\dfrac{1}{e^{\beta_0+\beta_1 x}}$*.*

Q: Why not use a linear regression to predict probabilities of class membership?

A: The linear regression will make predictions that don't make sense (e.g., probability outside of [0,1])

A: Transforming the linear regression into a step function will produce heteroskedastic errors

Q: How do we derive coefficients using maximum likelihood?
A: We find the coefficients that are the most likely, given the observed data. Formally, we estimate the coefficients that maximize the likelihood function. This is done using an iterative procedure.

Notation for the product of a series

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Check out this link, for details on the estimation of the coefficients.