GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 8: LOGISTIC REGRESSION

Paul Burkard

11/18/2015

LAST TIME:

- PROBABILITY
- BAYESIAN INFERENCE
- NAIVE BAYES CLASSIFICATION

QUESTIONS?

I. LOGISTIC REGRESSION

HANDS-ON: LOGISTIC REGRESSION

I. LOGISTIC REGRESSION

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

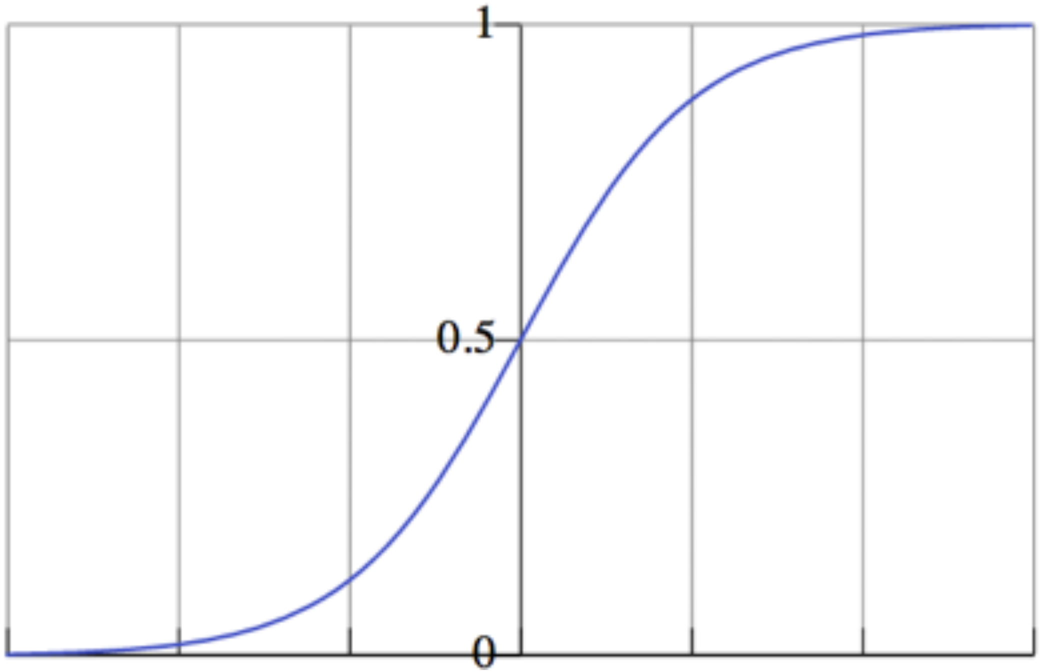
A: A generalization of the linear regression model to classification problems.

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

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

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

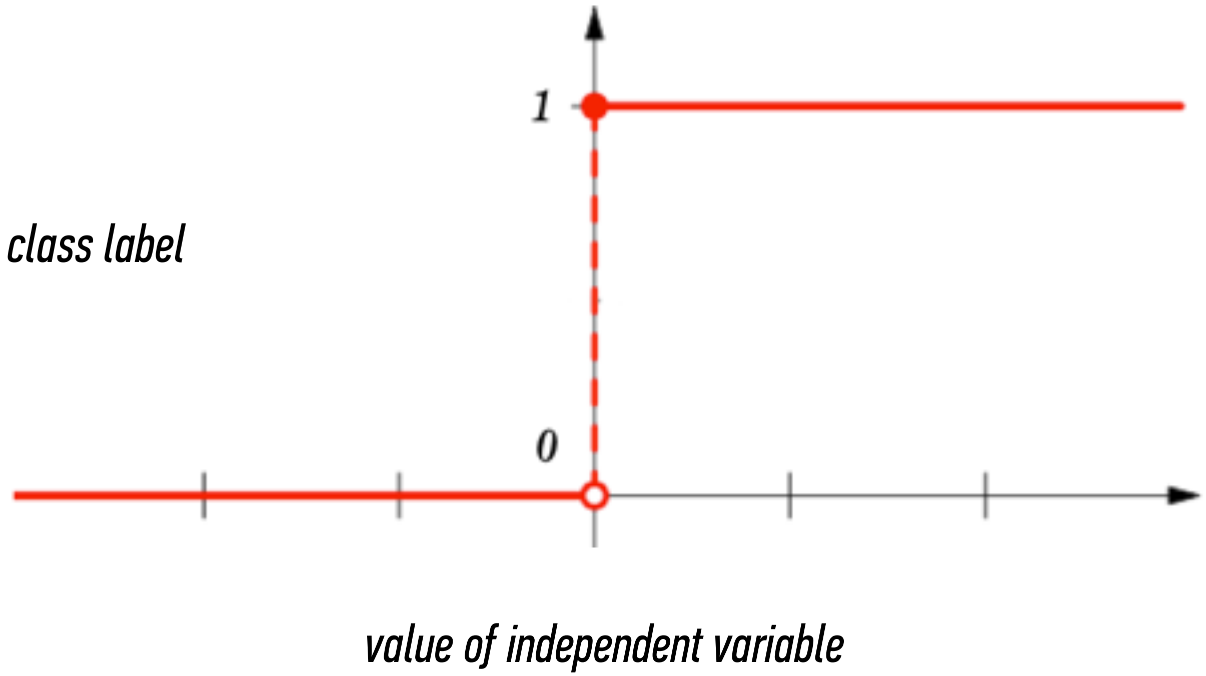
probability of belonging to class



value of independent variable

NOTE

Probability predictions look like this.



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 main difference is in the outcome variable.

The key variable in any regression problem is the **conditional mean** *of the outcome variable y given the value of the feature(s) x:*

$$E(y|x)$$

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

$$E(y|x) = \alpha + \beta x$$

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

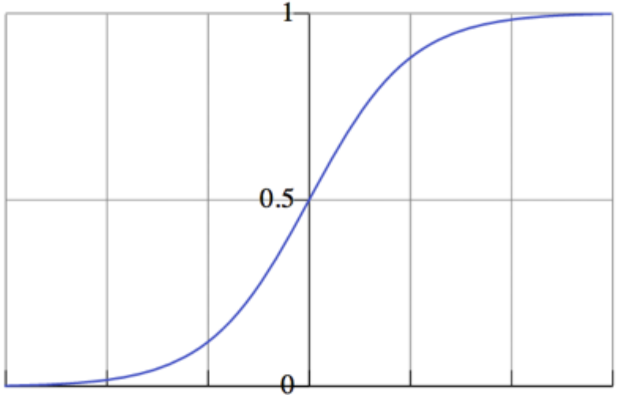
The first step in extending the linear regression model to logistic regression is to map the outcome variable $\mathbf{E}(y\,|\,x)$ *into the unit interval.*

Q: How do we do this?

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

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

We've already seen what this looks like:

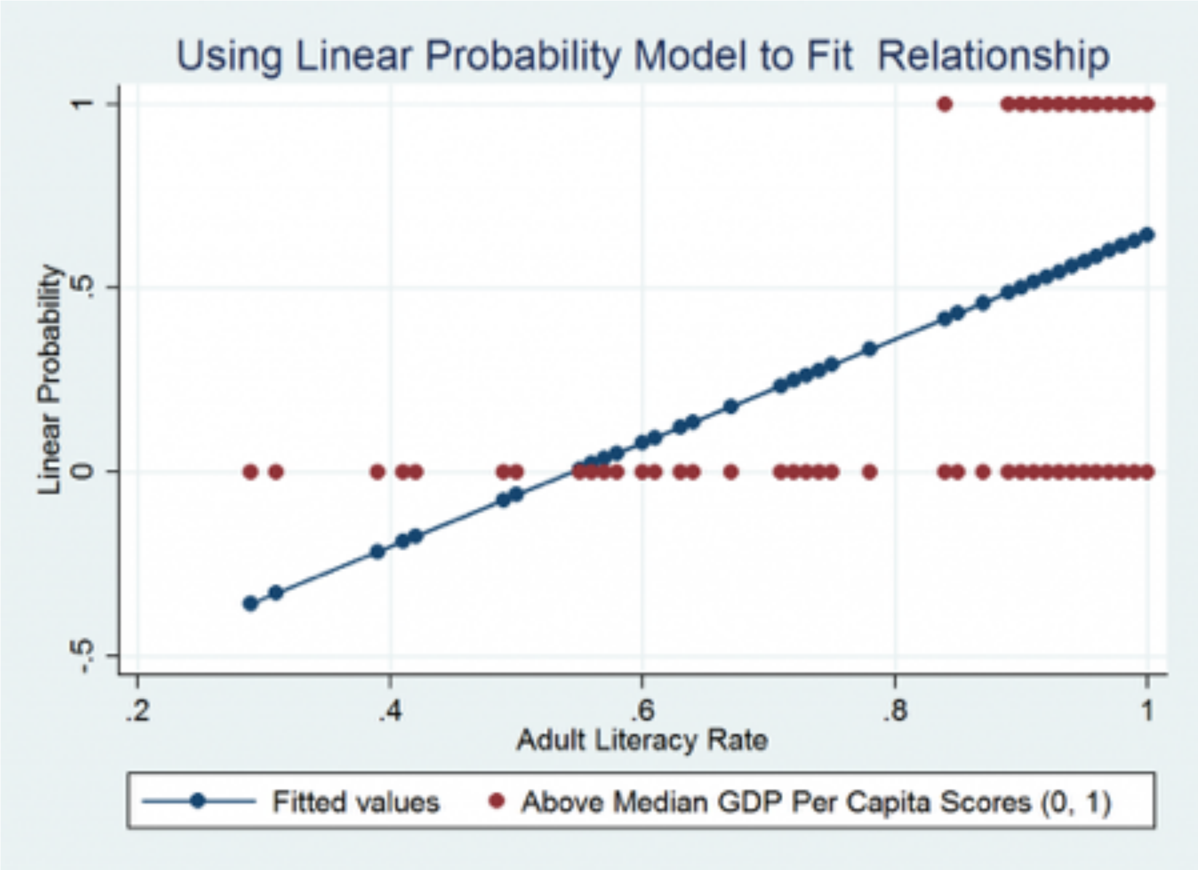


NOTE

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

This is a nonlinear transformation!

WHY LOGISTIC OVER LINEAR?



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**.

NOTE

This name hints at its usefulness in interpreting our results.

We will see why shortly.

We can now state the following:

$$e^{g(x)} = OR = e^{\alpha+\beta x}$$

So that if,

$$e^{\beta_i} = n$$

then the odds ratio is increased by a factor of n for a unit increase of x_i

WHEN TO USE LOGISTIC REGRESSION?

- *Classification Problems - Often good as a baseline method*
- *When we need an estimate of class likelihood ("probabilistic classifier")*
- *Many attributes*

HANDS-ON: LOGISTIC REGRESSION