Intro to Data Science:

Logistic Regression

Ed Podojil Data Engineer/Scientist, Animoto

Warm Up:

Agenda

I. Logistic Regression

II. Outcome Variables

III. Error Terms

IV. Interpreting Results

V. Implementing a Logistic Fit in R

I. Logistic Regression Regression

Goals

- Define **logistic regression**
- identify the differences in classification algorithms
- understand what logistic regression outputs for classification

	continuous	categorical
Supervised		
Unsupervised		

	continuous	categorical
Supervised	regression	classification
Unsupervised	dimension reduction	clustering

What is **logistic regression**?

A **generalization** of linear regression to **classification problems**.

Logistic regression is used to solve similar problems as KNN or Naïve Bayes.

Logistic Regression vs KNN

KNN is always distance based and not (always) linear

Logistic regression is more generalization and linear

Generalization: statement or concept obtained by inference from other data

Logistic Regression vs Naïve Bayes

Bayes is **generative**: based on general assumptions given about the data to classify answers

Logistic regression is **discriminative**: classification does not depend completely on the data set

How do I know which algorithm to use?

For now, trial and error (like in your homework!)

Learn more about data attributes

better answer: usually you want better data, not better models

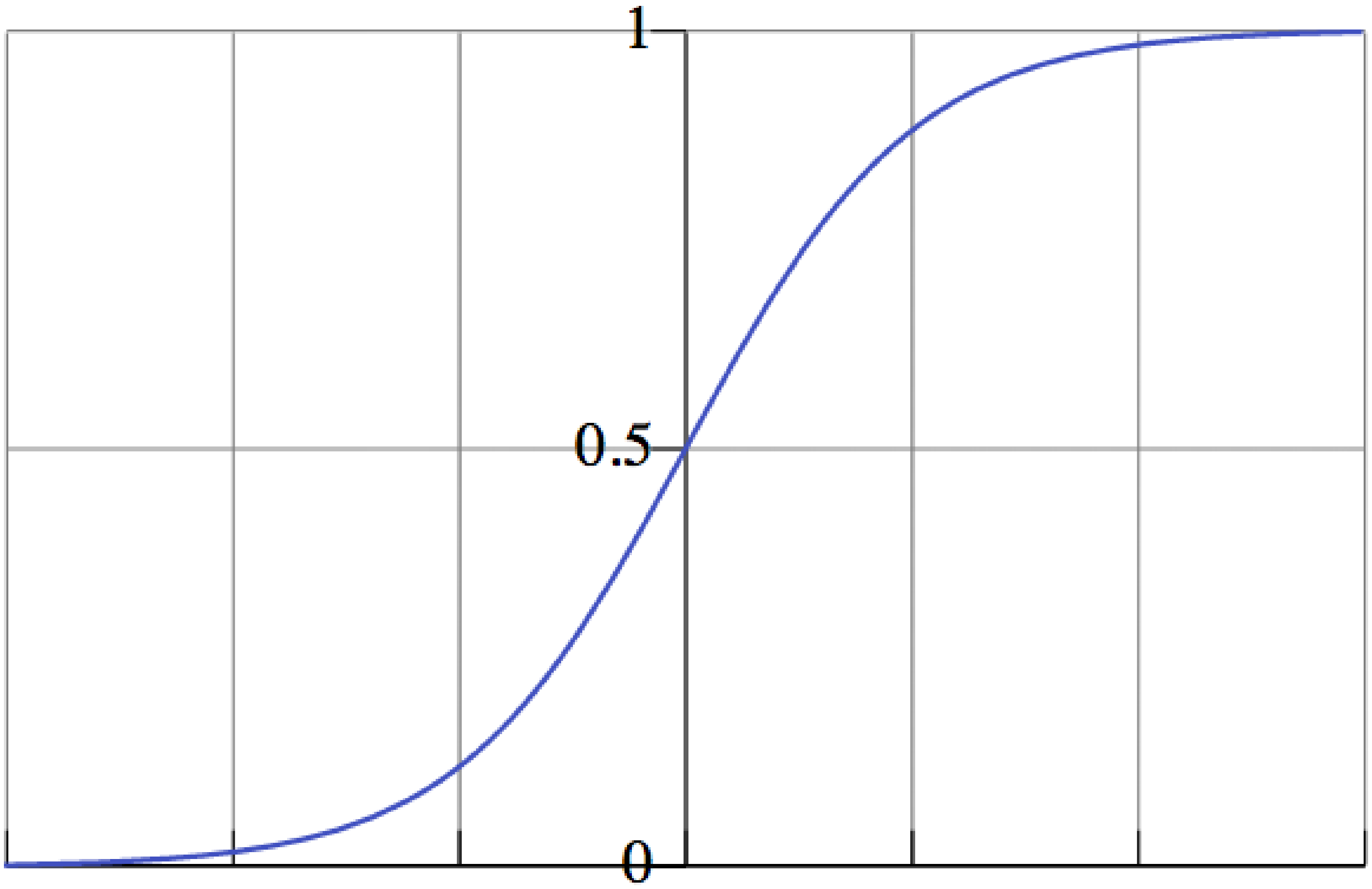
More about Logistic Regression

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

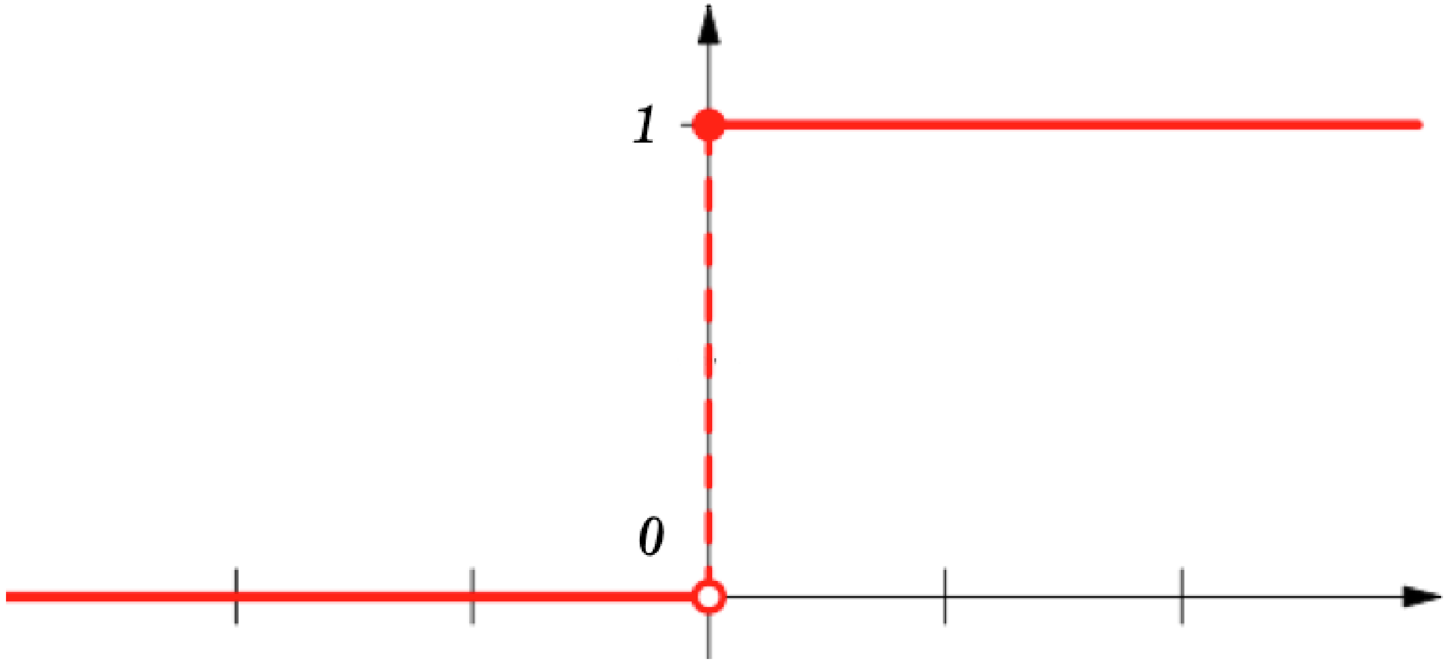
In logistic regression, we use a set of covariates to predict probabilities of **(binary) class membership**.

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

We use logistc regression to convert probabilities...



...into class labels!



Logistic vs Linear Regression

Both of these models work similarly. In fact, we can think of logistic regression as an extension of linear regression.

There are a couple important differences, however

- Difference 1: Outcome Variables
- Difference 2: Error Terms

II. Outcome Variables

Goals:

- review conditional mean
- Interpretation of the logistic formula

Outcome Variables

The key variable in any regression problem is the conditional mean of the outcome variable y given the value of the covariate x:

$$E(y|x)$$

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

$$E(y|x) = \alpha + ß x$$

Outcome Variables

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 linear regression to logistic regression is to map the outcome variable `E(y|x)` into the unit interval.

How do we do this?

By using a transformation called the logistic function

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

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

Here, π(x) (pi of x) is the probability of x occuring

(The logit function is also called the log-odds function)

III. Error Terms

Goals

- Understand the difference between Gaussian distribution and Bernoulli distribution

Error Terms

Review:

What error terms have we already discussed as a class?

What do each of those error terms mean?

(take a minute or two and find them in your notes)

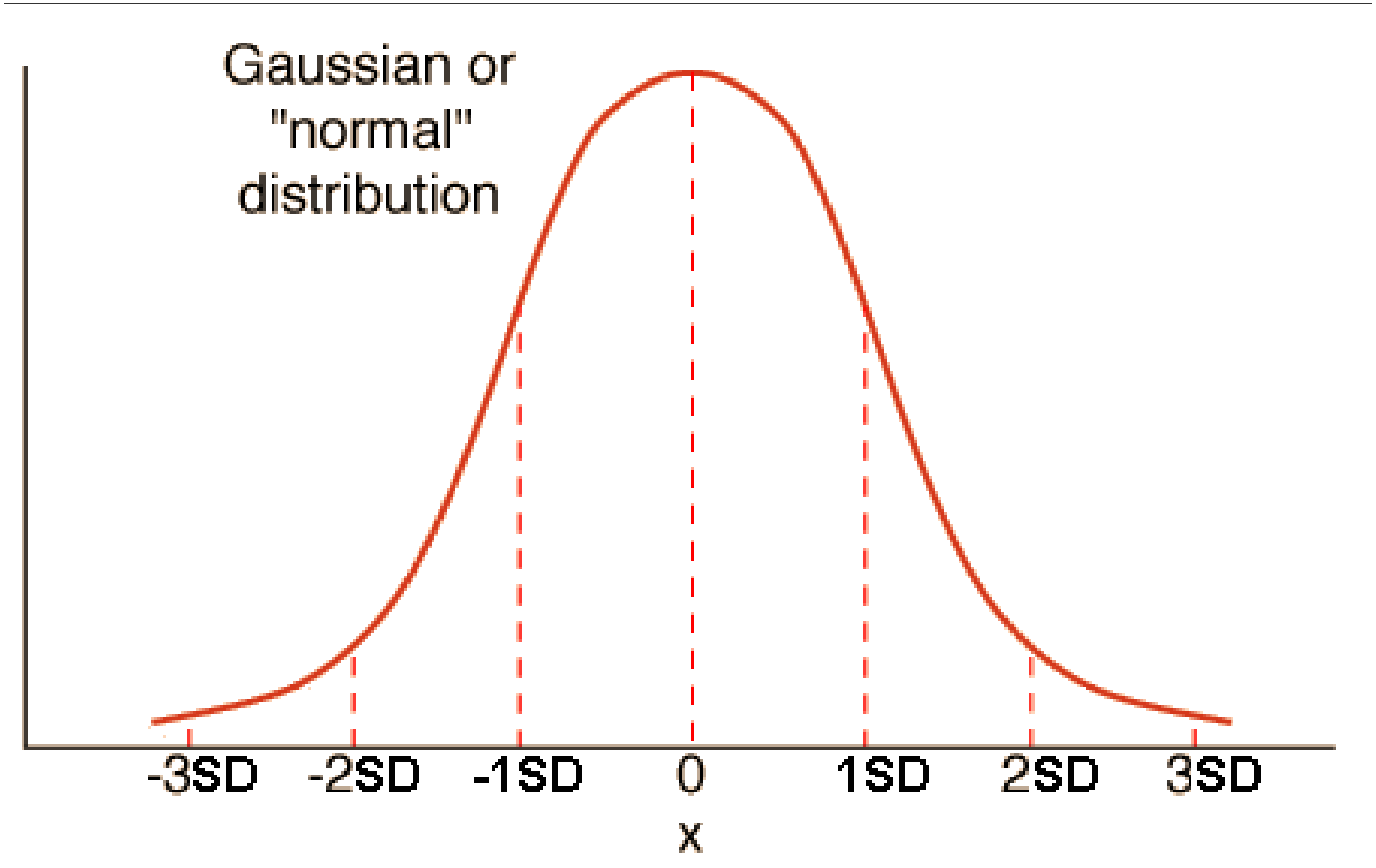
Typically, linear regressions will follow **independent Gaussian distribution** (normal, or bell curve) distrubution with **zero mean and constant variance**

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

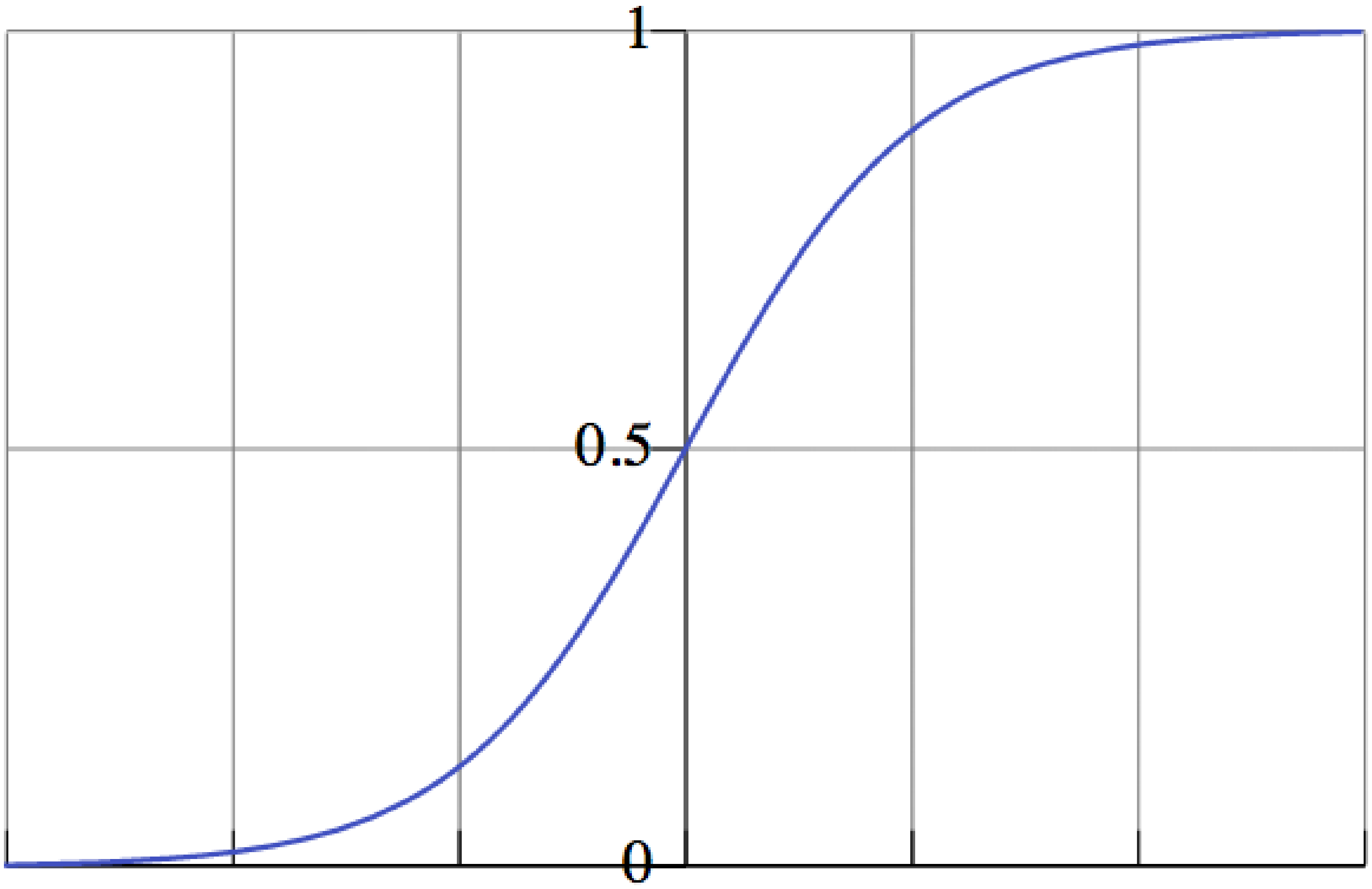
So, instead of following a Gaussian distribution, the error term in logistic regression follows Bernoulli distrubution

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

Gaussian distribution



Bernoulli distribution



Class break

IV. Interpreting Results

Goals

- Understand and define odds and odds ratio
- Understand how an odds ratio helps us interpret logistic regression

Interepeting Results

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.

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 $\pi(\mathbf{x})$ into this equation yields:

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

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

How do we interpret this?

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

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).

In this case, an odds ratio of 2 (eg, $\beta = \log(2)$) indicates that a purchase is twice as likely for an iOS user as for a non-iOS user.

V. Logistic Regression

Group Discussion

Questions to ponder about:

What data do you use in your work that depends on predicting probability?

There's a general movement of data scientists using less Naïve Bayes and more logistic regression. Why do you think that's the case?

What do we do when we want to classify data but don't have the class labels yet?