GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 11: LOGISTIC REGRESSION

I. LOGISTIC REGRESSION

	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.**

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.

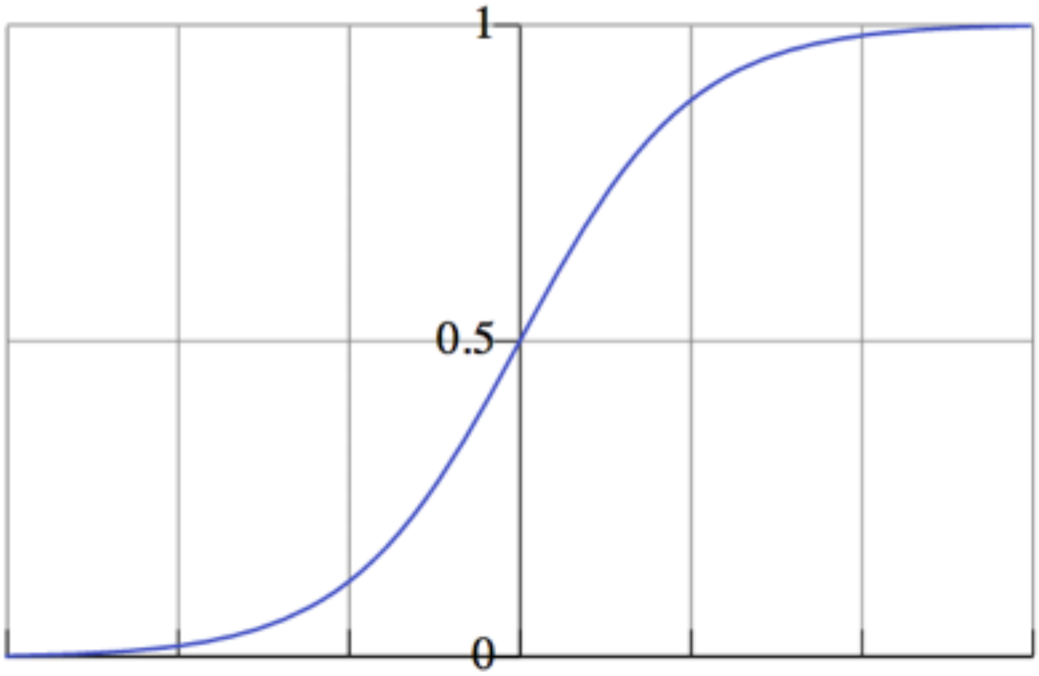
In logistic regression, we use a set of covariates to predict probabilities of (binary) class membership.

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.

probability of belonging to class



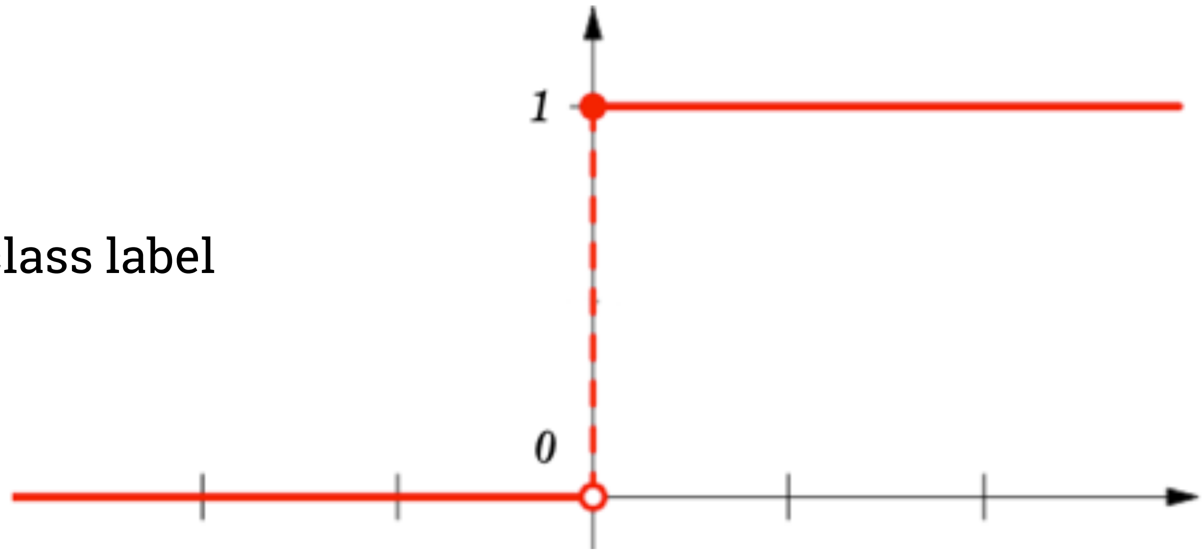
value of independent variable

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.

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 term.

II. 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: $E(y|x)$

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 $(-\infty, +\infty)$:

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

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

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 $E(y|x)$ 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 variable E(y|x) into the unit interval.

Q: How do we do this?

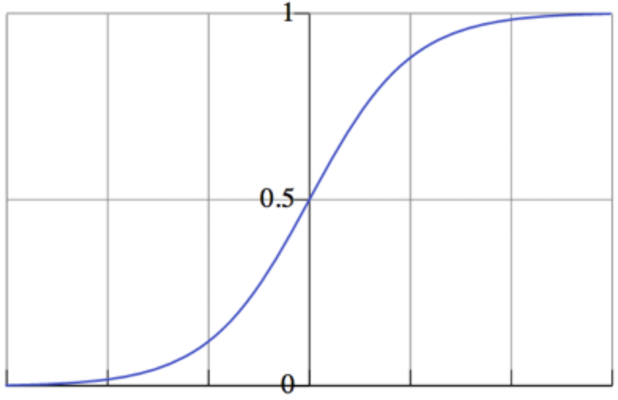
A: By using a transformation called the logistic function:

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

A: By using a transformation called the logistic function:

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

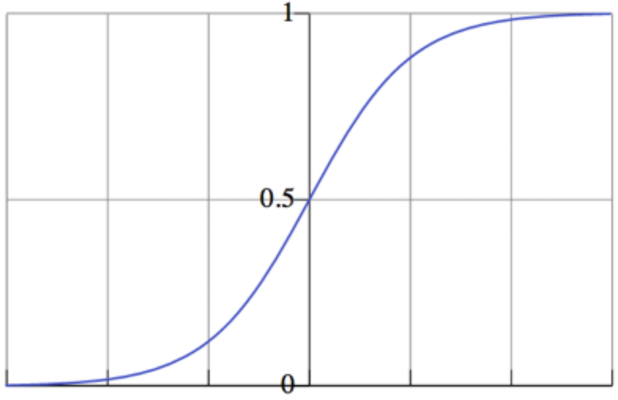
We've already seen what this looks like:



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!

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!

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

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.

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.

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

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.

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

Since the Bernoulli distribution and the logit function share a common parameter π, we say that the logit is the canonical link function for the Bernoulli distribution.

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

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

NOTE

This terminology is just FYI!

Since the Bernoulli distribution and the logit function share a common parameter π, we say that the logit is the canonical link function for the Bernoulli distribution.

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.

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 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.

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, 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.

Pros:

- Nice probabilistic interpretation (odds ratios) unlike decisions trees
- Many ways to regularize your model (we'll talk about this next time)

Cons:

- Usually less accurate than random forests
- NB is always faster (NB is great if conditional independence holds)
- Becomes computationally intense as number of outcomes increases

But keep in mind…

- Better data always beats better algorithms
- Creating quality features goes a long way
- If you have a very large dataset, accuracy differences between algorithms will be minimal, so choose based on speed
- Ensembles rule (we'll talk about these in a couple weeks)

EX: LOGISTIC REGRESSION