Logistic Regression

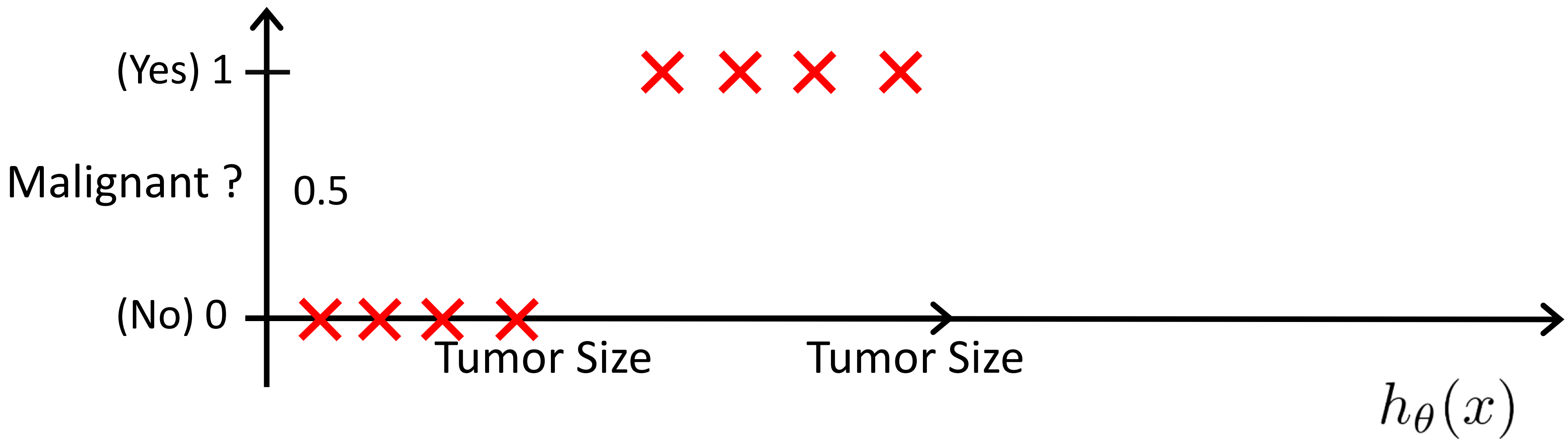
Jay Urbain, PhD

Topics

- Linear Regression for Classification
- Logistic regression
- Model interpretation
- Predicting default rates
- Q & A

Linear Regression for classification

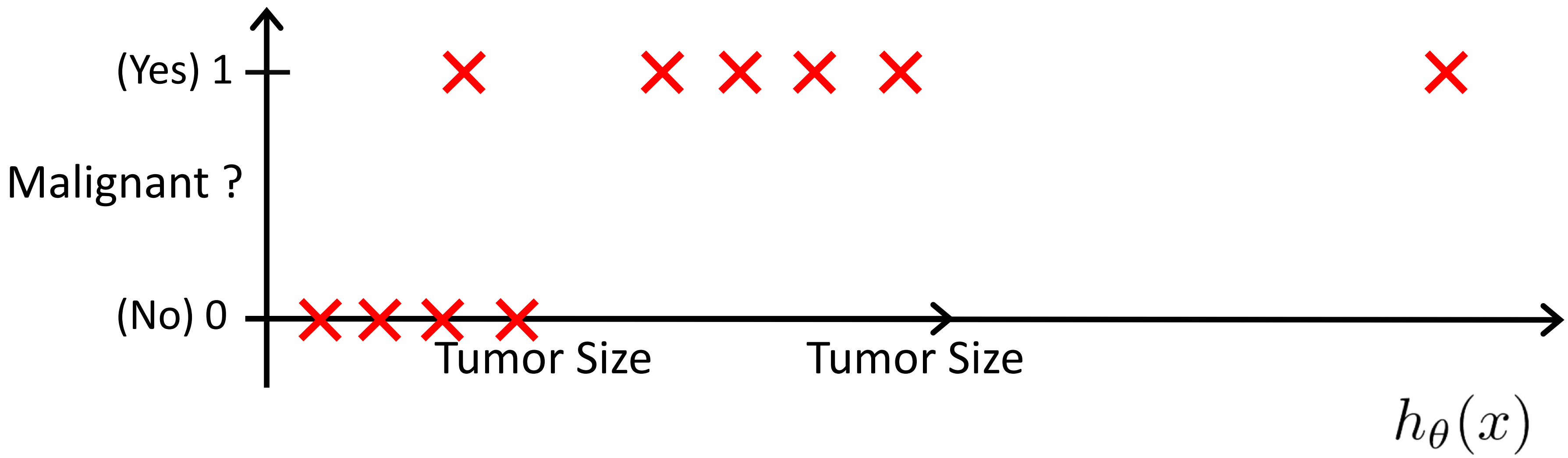
- Linear regression can be used for classification in domains with numeric attributes.
  - Perform a regression for each class, set output to 1 for instances that belong to the class, and 0 for those that do not.
  - The result is a linear expression for each class.
  - Then, given a test instance of an unknown class, calculate the value of each linear expression and choose the one that is **largest**.
  - *Called multinomial linear regression.*
  - Problems: output is not a proper probability, assumes errors are not statistically significant.



Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"



Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

Classification: y = 0 or 1

$h_\theta(x)$ can be > 1 or < 0

With regression

Logistic *Regression*: $0 \leq h_\theta(x) \leq 1$

Want proper probability

Classification

Logistic Regression

Q: What is logistic regression?

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

Logistic Regression – basic form

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

Logistic Regression – basic form

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 (category) membership.

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

Logistic Regression – basic form

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.

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

Logistic Regression – 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:

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

Logistic function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Logistic Regression – 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:

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

Probability of y = 1, given x

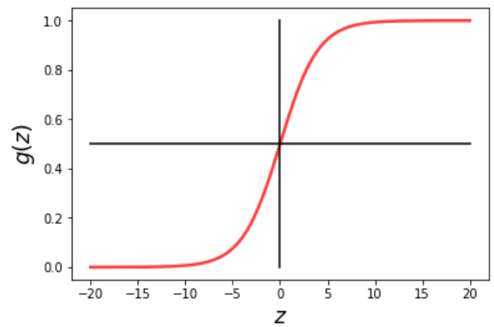
Logistic Regression – basic form

Optional in-class exercise: 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?

Figure 1. Logistic Function

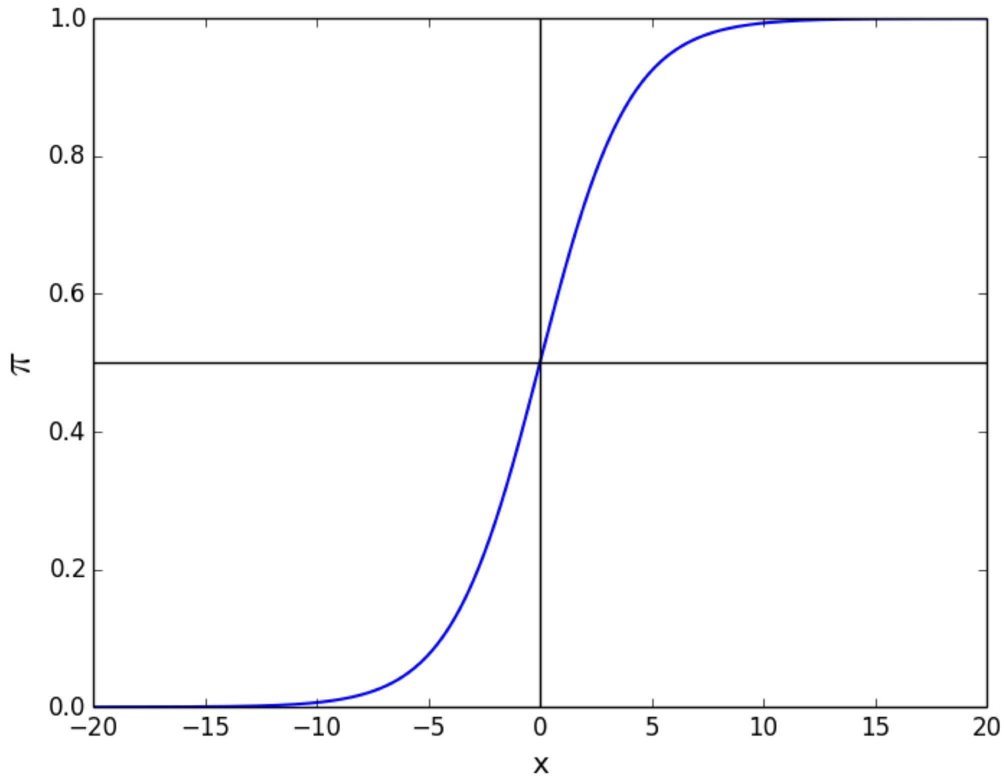


Logistic Regression – basic form

Why does this shape make sense?

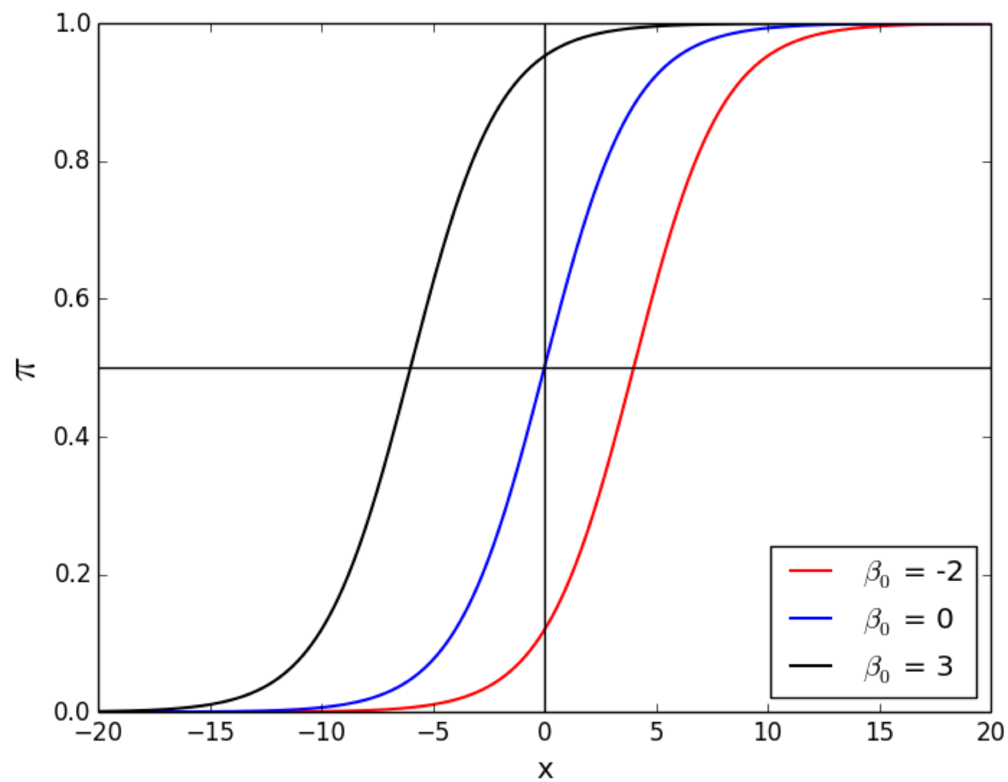
The logistic function takes on an "S" (sigmoid) shape, where y is bounded by [0,1]

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



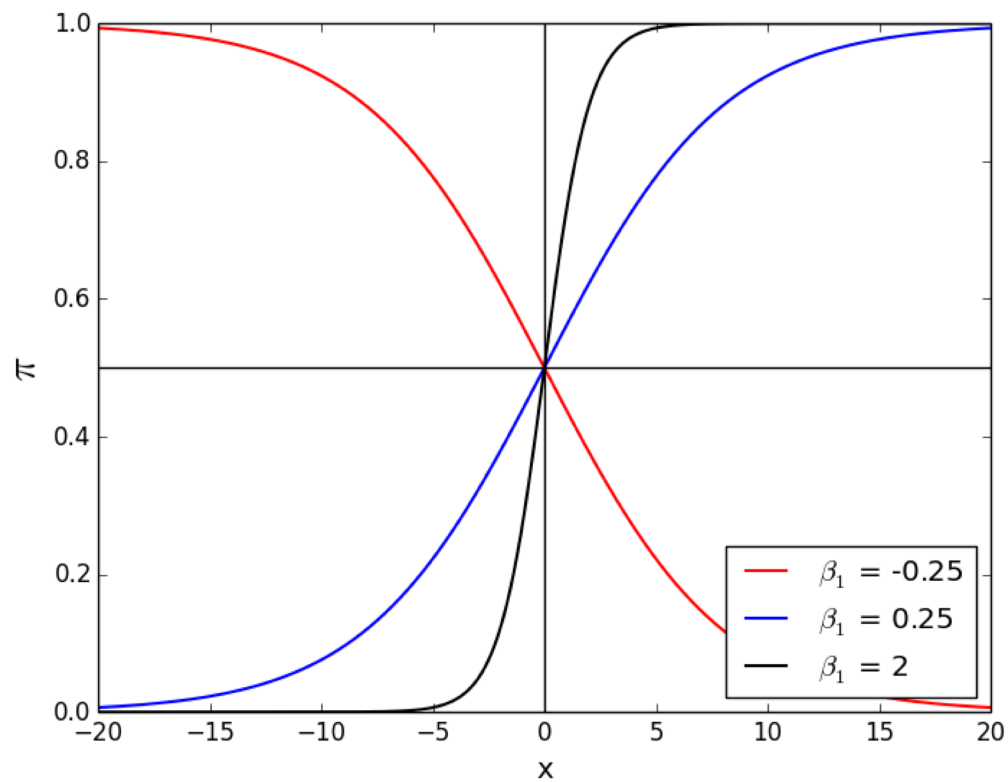
Logistic Regression – basic form

Changing the β_0 value shifts the function horizontally.



Logistic Regression – basic form

Changing the β_1 value changes the slope of the curve



Interpreting results

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:

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

$\boxed{\text{What is the range of the odds ratio?}}$

Interpreting results

Question: 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.

Interpreting results

Question: 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.

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

Interpreting results

Question: 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.

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

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

Interpreting results

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

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

Interpreting results

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

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

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

Interpreting results

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

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

What is the range of the logit function?

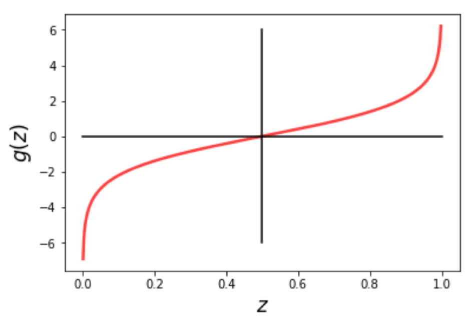
Interpreting results

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

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

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

Figure 1. log(x/(1-x)



The value of the logit function heads towards infinity as *p* approaches 1 and towards negative infinity as it approaches 0.

Interpreting results

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

Interpreting results

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.

Interpreting results

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.

Interpreting results

Q: How to determine whether a coefficient is significant?

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

Interpreting results

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

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

Q: What does this mean?

Interpreting results

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?

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.

Interpreting results

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

Logit function

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

Interpreting results

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

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

Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m examples $x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix}$ $x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

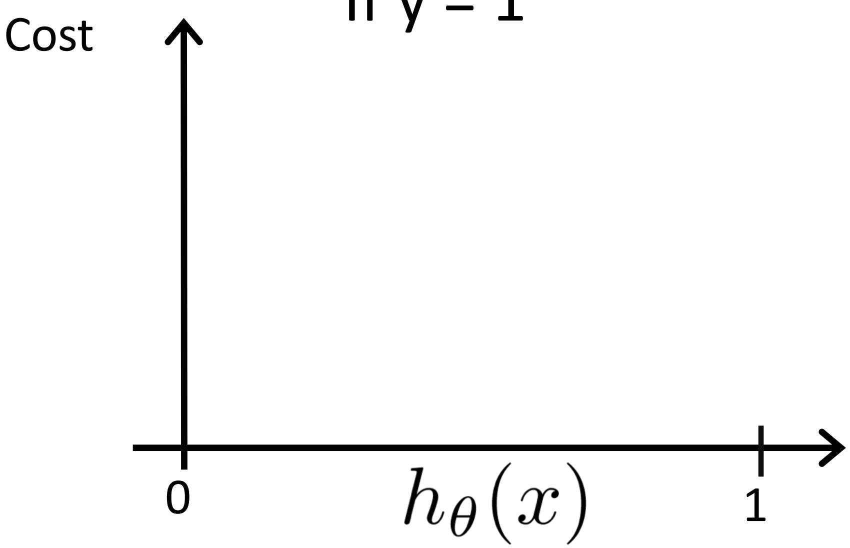
How to choose parameters θ ?

Logistic regression cost function

```
>>> -math.log(0.0000001,2)
23.25
>>> -math.log(1,2)
-0.00
```

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

If y = 1



$\mathrm{Cost} = 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \rightarrow 0$

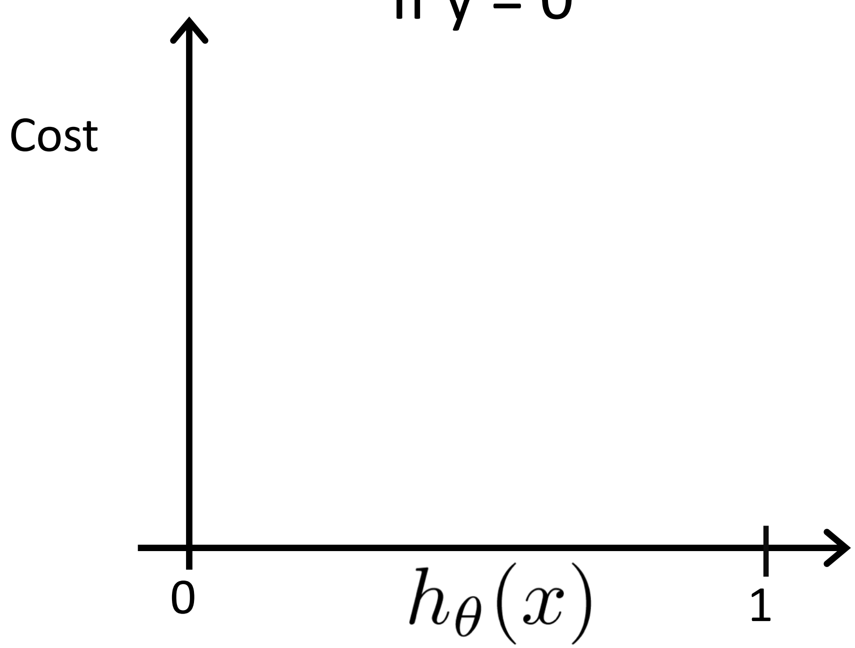
$Cost \rightarrow \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

Logistic regression cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

If y = 0



Logistic regression cost function with gradient Descent

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: $y = 0$ or 1 always

Logistic regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log\left(1-h_\theta(x^{(i)})\right)]$$

To fit parameters θ:

$$\min_\theta J(\theta)$$

To make a prediction given new *x* :

Output $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all θ_j)

$\}$

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

Algorithm looks identical to linear regression!

Predicting default – Go to lab

This data set contains 10,000 records associated with credit card accounts with the following four fields:

Default	Binary variable indicating whether the credit card holder defaulted on their credit card obligations
Student	Binary variable indicating whether the credit card holder is a student
Balance	Continuous variable recording the credit card holders current outstanding balance
Income	Continuous variable representing the total annual income for the credit card holder

Predicting default

Part I: Exploration

1) Read in Default.csv and convert all data to numeric
2) Split the data into train and test sets
3) Create a histogram of all variables
4) Create a scatter plot of the income vs. balance
5) Mark defaults with a different color (and symbol)
6) What can you infer from this plot?

Predicting default: hands-on

Part II: Logistic Regression

1) Run a logistic regression on the balance variable
    - Use the training set
    - Use the `scikit-learn`
2) Interpret the results

Predicting default: review

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

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

Predicting default: review

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

When the scatter of the errors is different, varying depending on the value of one or more of the independent variables, the error terms are *heteroskedastic*.

Predicting default: review

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.

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

Notation for the product of a series

Check out http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf for details on the estimation of the coefficients.

Predicting default: review

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.

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

Notation for the product of a series

Check out http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf for details on the estimation of the coefficients.