

# INTRO to DATA SCIENCE

## LOGISTIC REGRESSION

**0. BASIC FORM**

**I. INTERPRETATION**

**II. EXERCISE: PREDICTING DEFAULT RATES**

**III. Q&A**

# **0. BASIC FORM**

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	???	???
<i>unsupervised</i>	???	???

*Q: Where does logistic regression belong in this diagram?*

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	<i>regression</i>	<i>classification</i>
<i>unsupervised</i>	<i>dimension reduction</i>	<i>clustering</i>

*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.*

*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.*

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

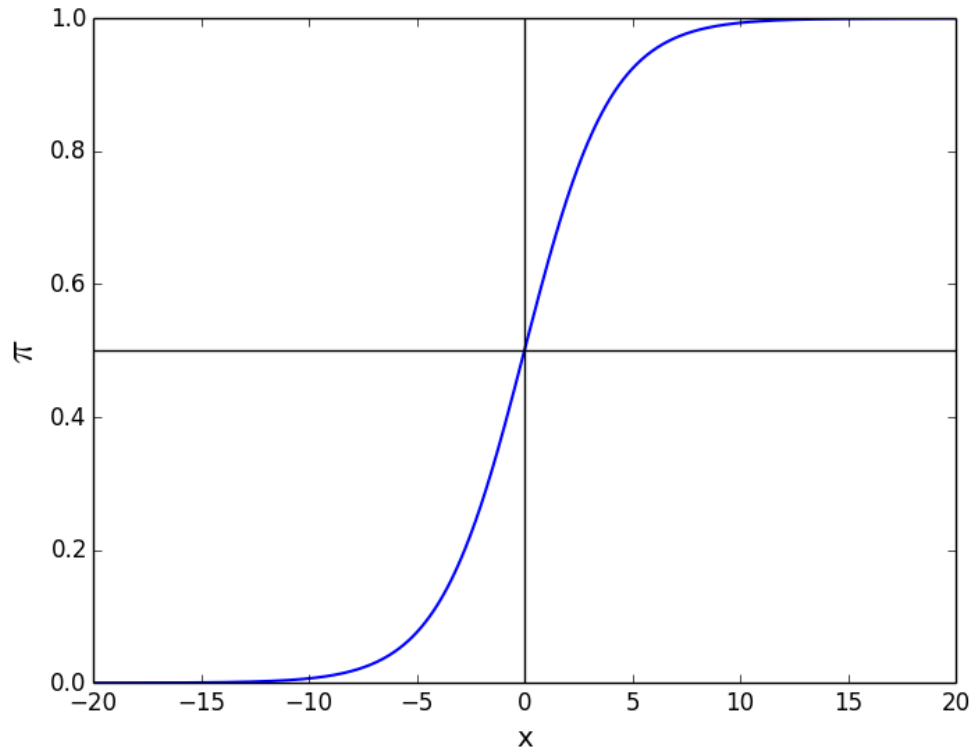
*Quiz: Create a plot of the logistic function.*

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

*How would you describe the shape of the function?*

*The logistic function takes on an “S” shape, where  $y$  is bounded by  $[0, 1]$*

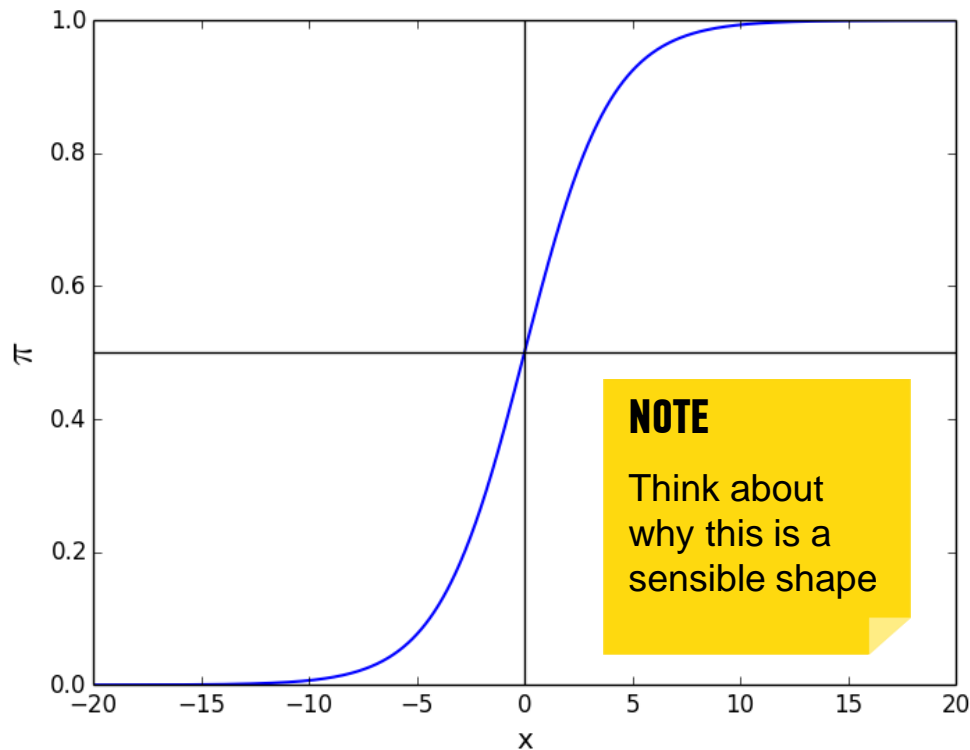
$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



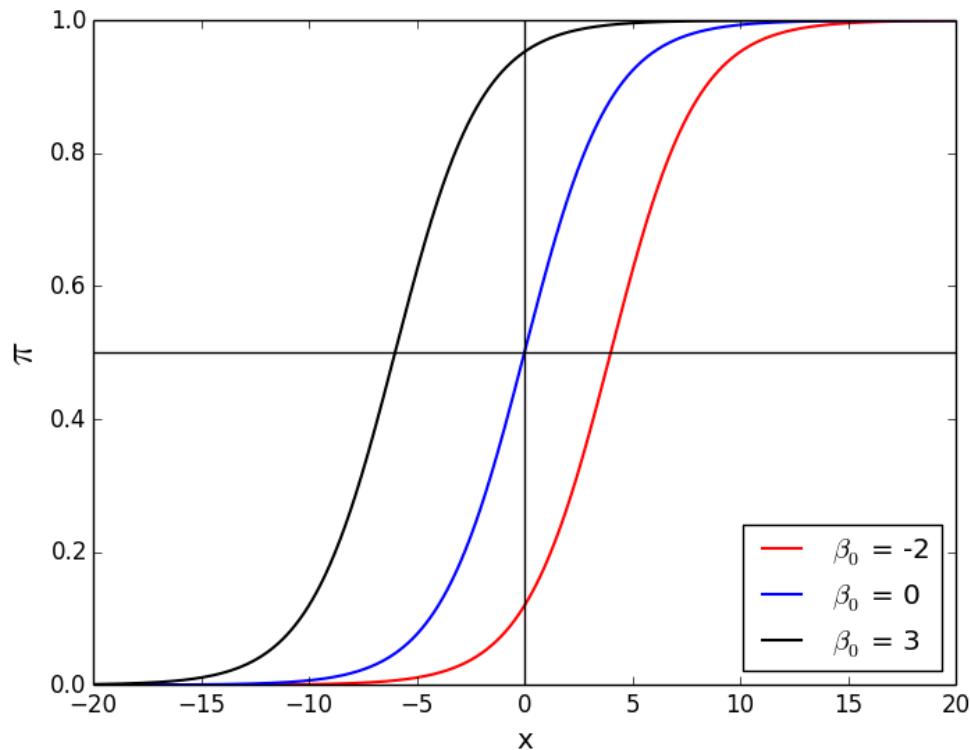


*The logistic function takes on an “S” shape, where  $y$  is bounded by  $[0, 1]$*

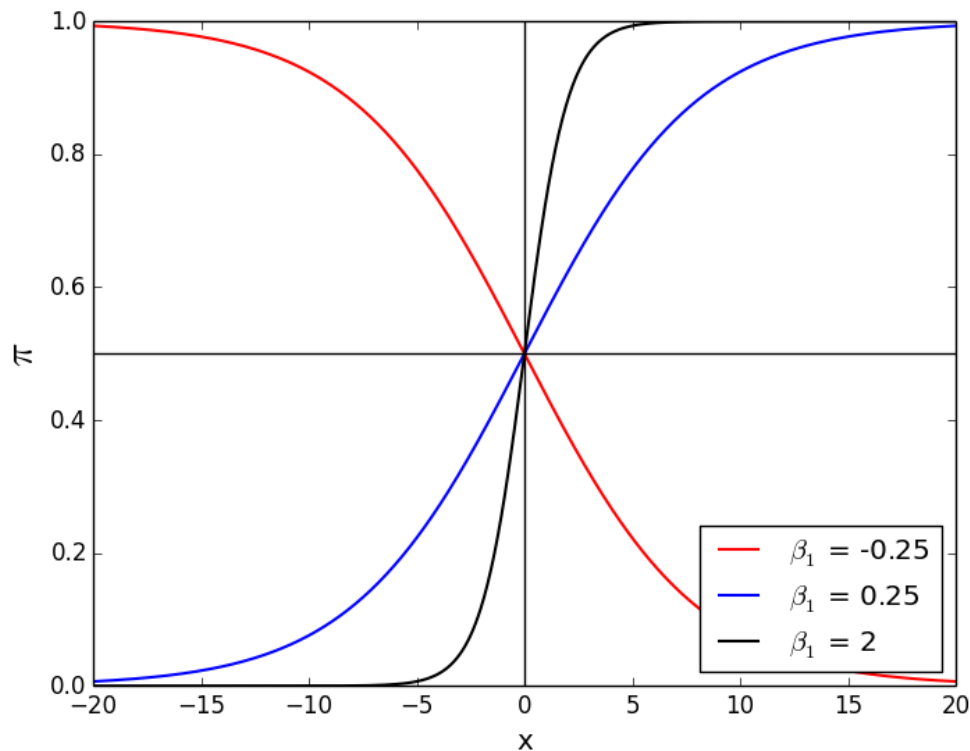
$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



*Changing the  $\beta_0$  value shifts the function horizontally.*



*Changing the  $\beta_1$  value  
changes the slope of  
the curve*



# **I. 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?*

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

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

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

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

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

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

*This simple relationship between the odds ratio and the parameter  $\beta$  is what makes logistic regression such a powerful tool.*



*In linear regression, the parameter  $\beta_1$  represents the change in the response variable for a unit change in  $x$ .*

*In linear regression, the parameter  $\beta_1$  represents the change in the **response variable** for a unit change in  $x$ .*

*In logistic regression,  $\beta_1$  represents the change in the **log-odds** for a unit change in  $x$ .*

*In linear regression, the parameter  $\beta_1$  represents the change in the **response variable** for a unit change in  $x$ .*

*In logistic regression,  $\beta_1$  represents the change in the **log-odds** for a unit change in  $x$ .*

*This means that  $e^{\beta_1}$  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  $\beta_1 = 0.693$ .*

*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


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

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

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

Logistic function



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

# **II. EXERCISE: PREDICTING DEFAULT**

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

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

### *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?*

## ***Part II: Logistic Regression***

- 1) *Run a logistic regression on the balance variable*
  - *Use the training set*
  - *Use the `statsmodels.formula.api` module and `smf.logit()` function*
- 2) *Is the  $\beta$  value associated with balance significant?*
- 3) *Predict the probability of default for someone with a balance of \$1.2k and \$1.5k*
- 4) *Plot the fitted logistic function overtop of the data points*
- 5) *Create predictions using the test set*
- 6) *Compute the overall accuracy, the sensitivity and specificity*

---

## **INTRO TO DATA SCIENCE**

---

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

*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.*