

# CSCE 421: Machine Learning

## Lecture 6: Logistic Regression

Texas A&M University

Bobak Mortazavi

Ryan King

Zhale Nowroozilarki

# Goals

---

- Learn Logistic Regression!
- See how Gradient Descent helps Logistic Regression!
- Evaluate Measures of Performance with Cross-Validation

# Important Questions

---

- Why do we need it? Classification vs. Regression
- Why not simply use Linear Regression?
- What are odds?
- How do we design Logistic Regression?
- How do we find optimal coefficients for Logistic Regression?

## An example of questions that need classification

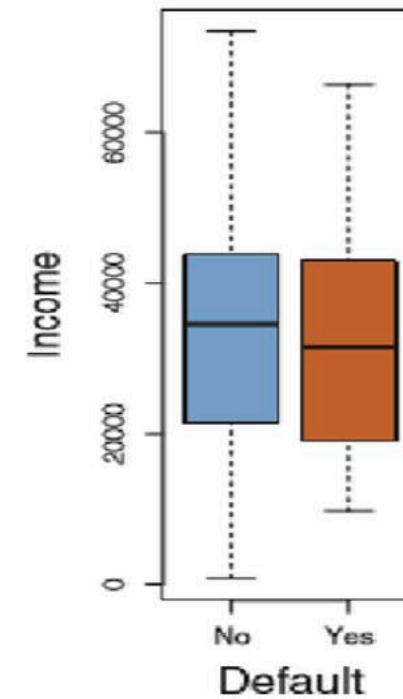
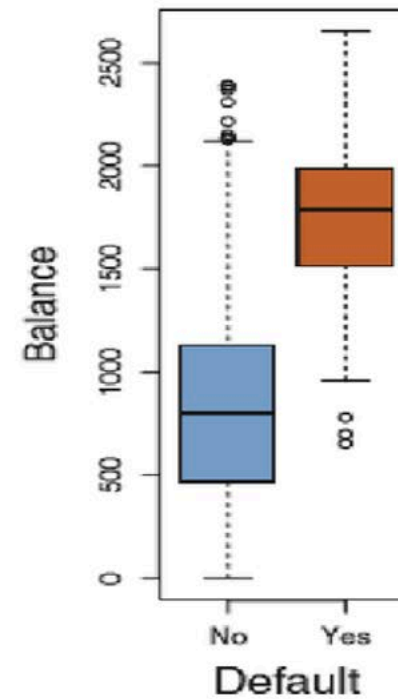
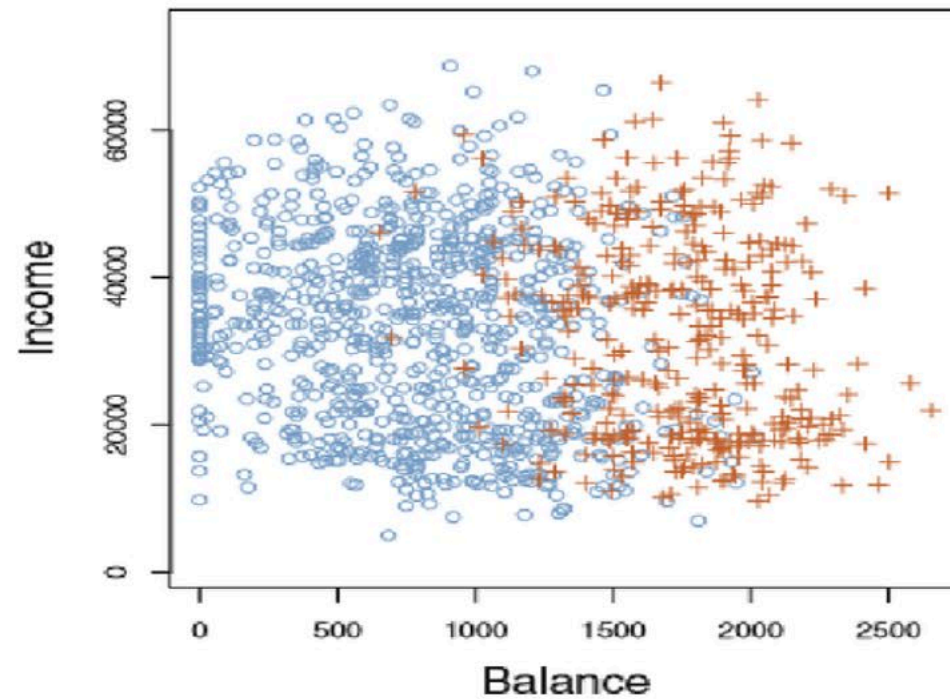
- A person arrives in the emergency room and has symptoms that present as 1 of 3 different conditions, which one is it?
- A bank must determine which transactions are fraudulent
- What is the likelihood someone will default on credit card payments?
- What are some other examples of classification problems?

## An example of questions that need classification

- A person arrives in the emergency room and has symptoms that present as 1 of 3 different conditions, which one is it?
- A bank must determine which transactions are fraudulent
- What is the likelihood someone will default on credit card payments?
- Still in the data scenario of  $D = \{(x_i, y_i)\}_{i=1}^n$ , but now  $y \in \{0,1\}$

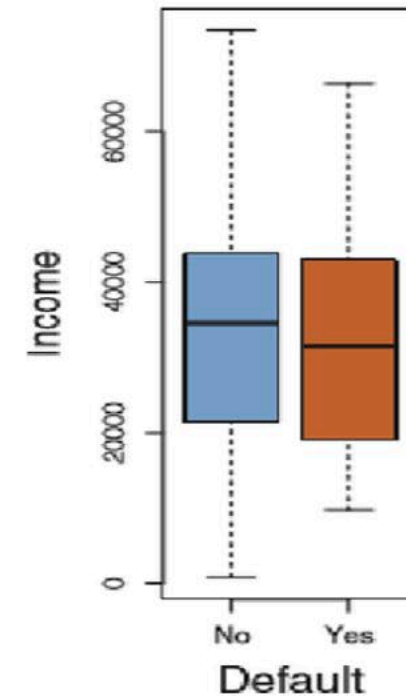
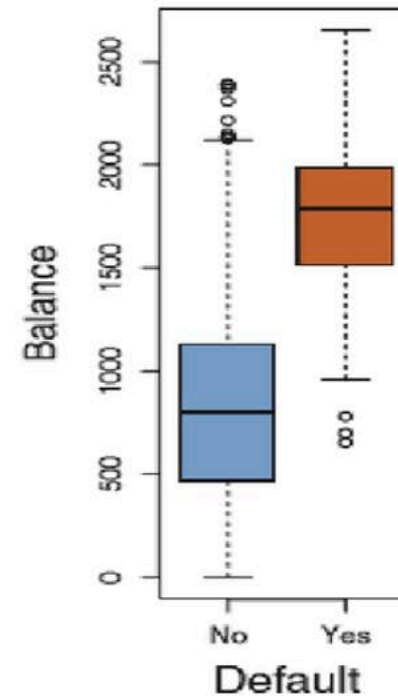
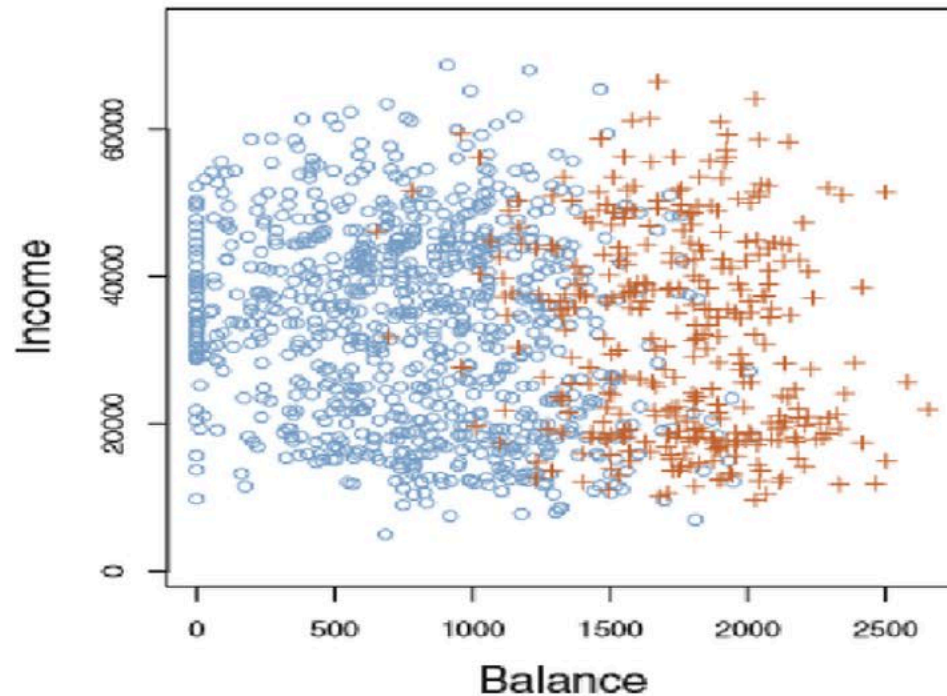
# Credit Default Prediction

- Predict class A vs. class B
- Determine if two classes are separable



# Credit Default Prediction

- Predict class A vs. class B
- Determine if two classes are separable
- Why not simply use linear regression and use values to determine classes?



# Why not linear regression?

- Imagine someone comes into the emergency room, and presents with a likelihood of either heart attack, overdose, or treatment for broken bones - and you need a model to quickly determine which is highest/needs to be treated.



# Why not linear regression?

- Imagine someone comes into the emergency room, and presents with a likelihood of either heart attack, overdose, or treatment for broken bones - and you need a model to quickly determine which is highest/needs to be treated.
- Encoding these outcomes presents a natural ordering (e.g. 1, 2, 3)

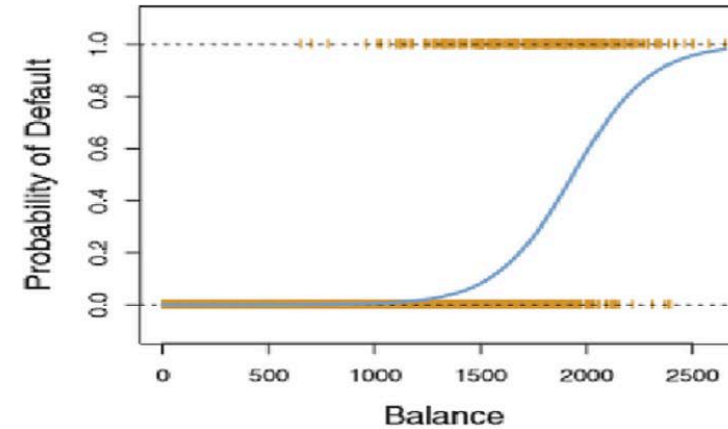
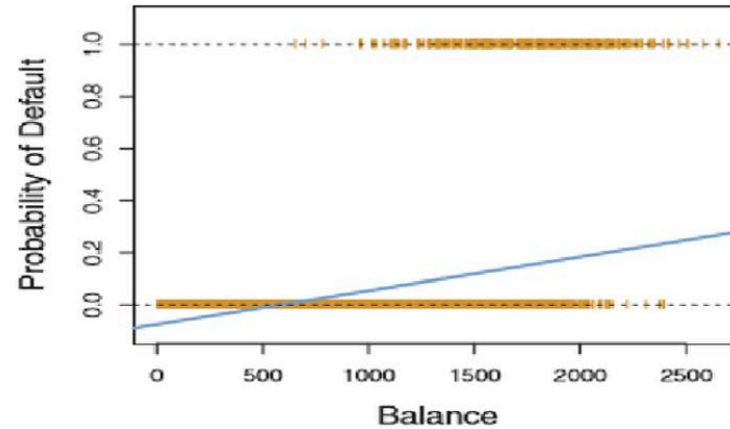
# Why not linear regression?

- Imagine someone comes into the emergency room, and presents with a likelihood of either heart attack, overdose, or treatment for broken bones - and you need a model to quickly determine which is highest/needs to be treated.
- Encoding these outcomes presents a natural ordering (e.g. 1, 2, 3)
- But what is the right encoding to use here?

# Why not linear regression?

- Imagine someone comes into the emergency room, and presents with a likelihood of either heart attack, overdose, or treatment for broken bones - and you need a model to quickly determine which is highest/needs to be treated.
- Encoding these outcomes presents a natural ordering (e.g. 1, 2, 3)
- But what is the right encoding to use here?
- When there are natural order to data it isn't challenging – for example mild moderate or severe infection.
- In this case a linear regression can be calculated to model against 1, 2, and 3, where the gap between 1 and 2 and the gap between 2 and 3 would be considered the same.
- So perhaps a single binary 0 vs. 1 decision could use a linear regression?

# Linear vs. Logistic Regression

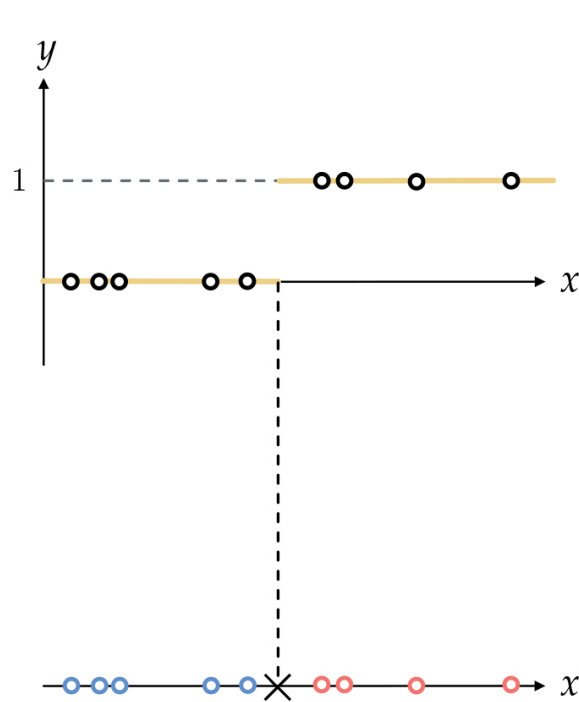


- Linear regression could pick a value such that  $\hat{y} > \tau$  results in a class decision
- Hard to interpret. Look at the figure on the right – this is easy to predict as a probability  $p(\text{default} \mid \text{balance})$
- It would be nicest to predict the probability of  $y$  belonging to a class, then classifying if the probability is  $> 50\%$ .

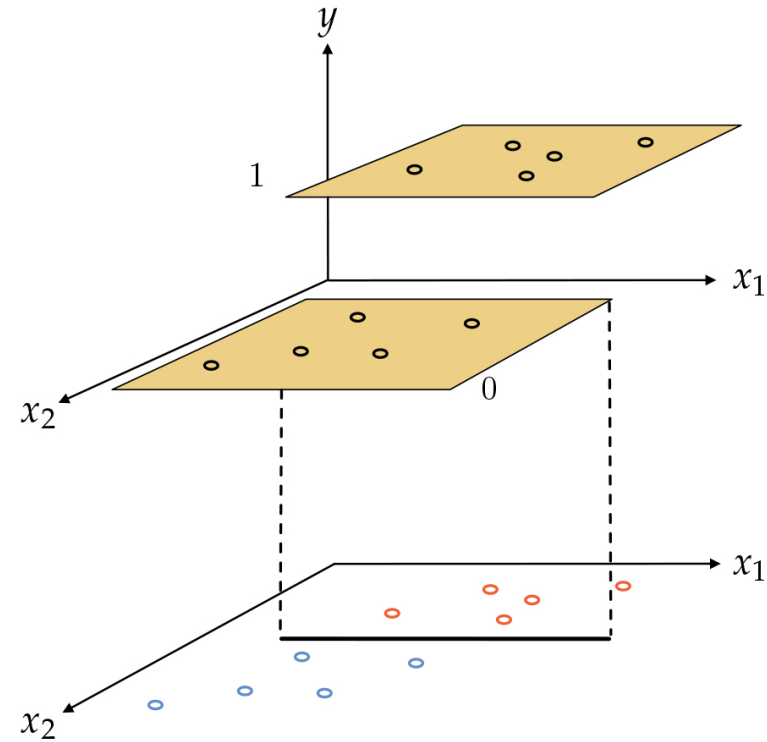
# Decision Boundaries (Discriminative vs. Generative Models)

---

# Classification – Step Functions

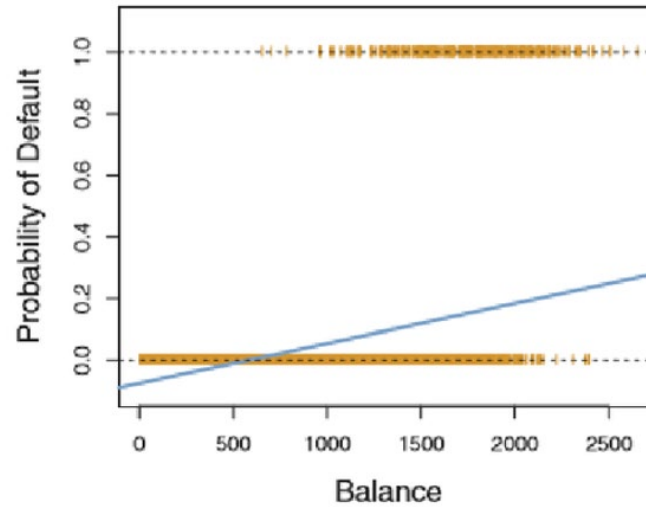


one-dimensional input:  
decision boundary is a single point



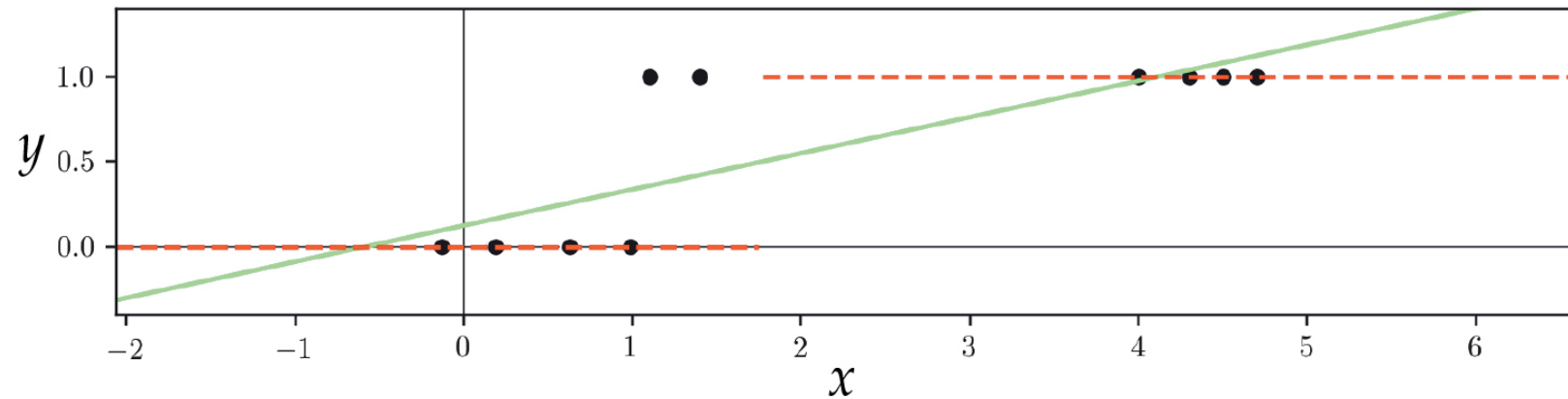
two-dimensional input:  
decision boundary is a line

# Linear Regression for Classification



If you set  $y > 0.5$  as a decision rule – not clear it works  
Hard to interpret – results are not contained in  $[0,1]$

# Step function from Linear Regression Boundary





# Bernouli Distribution

- We can create a probability density function that represents a single experiment asking yes/no

$$Y \sim \text{Bernouli}(\theta), Y \in \{0,1\}$$

$$p(y|\theta) = \theta^{I(y=1)}(1 - \theta)^{I(y=0)} = \begin{cases} \theta, y = 1 \\ 1 - \theta, y = 0 \end{cases}$$

- Can think of this as a coin toss experiment and the likelihood of heads vs. tails

# Logistic Regression

- **Parametric classification** method (**not regression**), which is sometimes referred to as a “generalization” of linear regression because
  - We still compute a linear combination of feature inputs,  $\mathbf{x}^T \mathbf{w}$  (sometimes written  $\mathbf{w}^T \mathbf{x}$ )
  - However, instead of estimating the continuous output variable, we pass this into a function

$$\mu(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$

- Where  $0 \leq \mu(\mathbf{w}^T \mathbf{x}) \leq 1$ , and where the Gaussian noise of linear regression is replaced by the Bernoulli Distribution so that

$$p(y|\mathbf{x}, \mathbf{w}) = \text{Ber}(y|\mu(\mathbf{w}^T \mathbf{x}))$$

- Therefore, the output belongs to a class 1 ( $y = 1$ ) with probability  $\mu(\mathbf{w}^T \mathbf{x})$ , and class 0 ( $y = 0$ ) with probability  $1 - \mu(\mathbf{w}^T \mathbf{x})$

# Why use a Sigmoid Function?

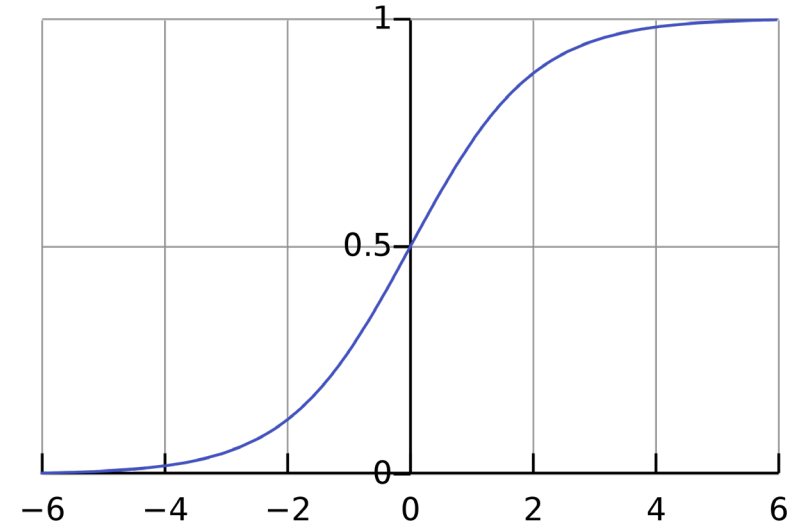
$$\sigma(\eta) = \frac{1}{1 + e^{-(\eta)}} = \frac{e^{\eta}}{1 + e^{\eta}}$$

- Has very nice properties for classification
  - Bounded between 0 and 1 <- thus interpretable as a probability
  - Monotonically increasing <- thus can be used for classification rules

$\sigma(\eta) > 0.5$  , positive class ( $y = 1$ )

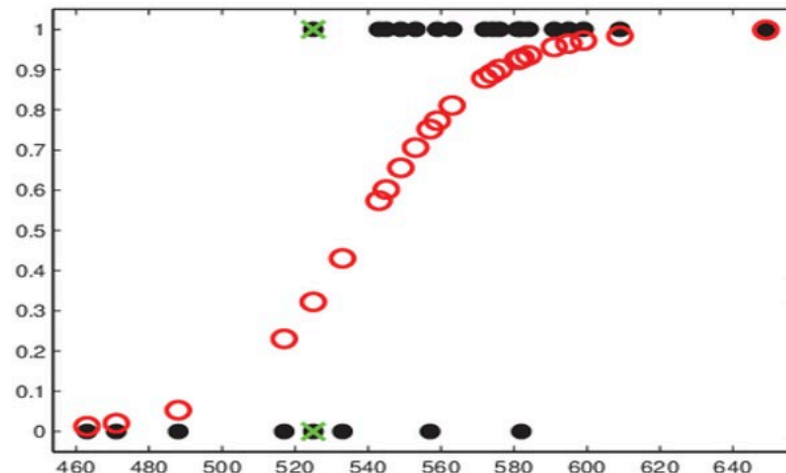
$\sigma(\eta) \leq 0.5$  , negative class ( $y = 0$ )

- Nice computational properties for optimizing criterion function



# Using logistic function for probability

- Classification task: Predict whether a student will pass a first year class or not
- Features: SAT Scores
- Data: set of SAT Scores and pass/fail labels
- Logistic Regression: Assigns each score to a pass probability (red circle), and assigns label with probability greater than 50%



# Logistic Regression: Representation

- Setup classification problem for two classes
  - Input:  $\mathbf{x} \in \mathbb{R}^p$
  - Output:  $y \in \{0, 1\}$
  - Training Data:  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$
  - Model parameters:  $\mathbf{w}$  (weights)
  - Model:

$$p(y | \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}), \sigma(\eta) = \frac{1}{1 + e^{-(\eta)}} = \frac{e^\eta}{1 + e^\eta}$$

$$y = \begin{cases} 1, & p(y | \mathbf{x}, \mathbf{w}) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

# Logistic Function

- Let's look at an example with a single variable  $x$

$$p(y|x) = p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

# Logistic Function

- Let's look at an example with a single variable  $x$

$$p(y|x) = p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

- All values are between 0 and 1
- After some algebraic manipulation, we can re-write this as

$$\frac{p(x)}{1 - p(x)} = e^{w_0 + w_1 x}$$

# Logistic Function

- Let's look at an example with a single variable  $x$

$$p(y|x) = p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

- All values are between 0 and 1
- After some algebraic manipulation, we can re-write this as

$$\frac{p(x)}{1 - p(x)} = e^{w_0 + w_1 x}$$

- These are called odds



## Briefly: Odds

- Odds can be between 0 and infinity
- For example: 1 in 5 people will default. This is then with odds 1 out of 4

$$p(x) = 0.2 \text{ and then odds } \frac{0.2}{0.8} = \frac{1}{4}$$

## Briefly: Odds

- Odds can be between 0 and infinity
- For example: 1 in 5 people will default. This is then with odds 1 out of 4

$$p(x) = 0.2 \text{ and then odds } \frac{0.2}{0.8} = \frac{1}{4}$$

- If we change this to 9 in 10 people

$$p(x) = 0.9 \text{ and then odds } \frac{0.9}{0.1} = \frac{9}{1} - \text{the odds are 9 to 1}$$

# Logistic Function

- Let's look at an example with a single variable  $x$

$$p(y|x) = p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

- All values are between 0 and 1
- After some algebraic manipulation, we can re-write this as

$$\frac{p(x)}{1 - p(x)} = e^{w_0 + w_1 x}$$

- These are called odds
- Then we can take the log of the odds (the log-odds or logit)

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = w_0 + w_1 x$$

# Multiple Logistic Regression

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = w_0 + w_1x_1 + \cdots + w_px_p$$

- $p(y|x, w) = \text{Ber}(y|\sigma(w^T x))$ , with a linear decision boundary at  $p(x) > 0.5$
- We can then calculate the negative of the log of the likelihood (negative log likelihood)

$$NLL(w) = - \sum_{i=1}^n (y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i)))$$

# Multiple Logistic Regression

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = w_0 + w_1x_1 + \cdots + w_px_p$$

- $p(y|x, w) = \text{Ber}(y|\sigma(w^T x))$ , with a linear decision boundary at  $p(x) > 0.5$
- We can then calculate the negative of the log of the likelihood (negative log likelihood)

$$NLL(w) = - \sum_{i=1}^n (y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i)))$$

- Now suppose we recast  $\tilde{y} = \{-1, 1\}$ , then

$$NLL(w) = - \sum_{i=1}^n \left( \log \left( 1 + e^{-\tilde{y}_i w^T x_i} \right) \right)$$

- Which has no closed form solution

# Logistic Regression: Evaluation

- Data likelihood (1 training sample)

$$p(y|x) = \begin{cases} \sigma(w^T x) & y = 1 \\ 1 - \sigma(w^T x) & \text{otherwise} \end{cases} = \sigma(w^T x)^y (1 - \sigma(w^T x))^{1-y}$$

- Data likelihood (all training samples)

$$L(D, w) = \prod_{i=1}^n p(y_i|x_i, w) = \prod_{i=1}^n \sigma(w^T x_i)^{y_i} (1 - \sigma(w^T x_i))^{1-y_i}$$

- Log Likelihood

$$l(D, w) = \sum_{i=1}^n (y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i)))$$

- Negative Log Likelihood (the cross-entropy error)

$$nll(w) = - \sum_{i=1}^n (y_i \log \sigma(w^T x_i) + (1 - y_i) \log(1 - \sigma(w^T x_i)))$$

## Next Time

---

- Continued review of logistic regression
- Discuss why logistic regression has no closed form solution
- Discuss what makes optimizing logistic regression necessary
- Optimizing logistic regression to identify coefficients to variables