Logistic Regression

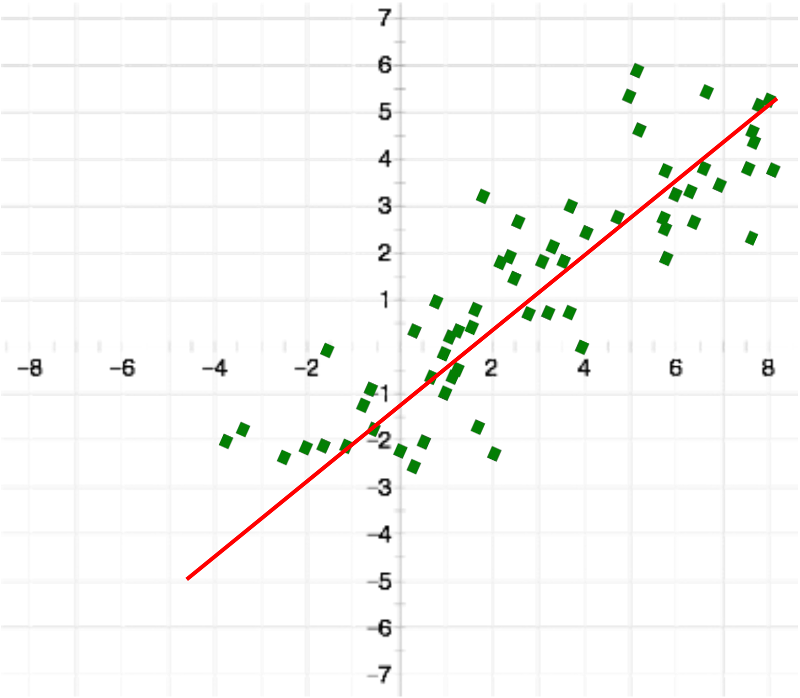
Intro to Machine Learning: Beginner Track #3

Slides: tinyurl.com/f20btrack3
Attendance code: **findingnemo**
Discord: bit.ly/ACMdiscord
Internship: bit.ly/ACMInternApplications2020

acm.ai

Linear regression recap

What is linear regression?



- Goal: to find the equation of a line that best fits our data
  - We want to be able to use this line to predict outputs from given inputs
- Notice that the outputs are continuous
  - Classification or regression?

Linear Regression

$$\hat{y}(x) = b + w_1x_1 + w_2x_2 + \ldots\ldots w_nx_n$$

An input **X** is an **n-dimensional vector** for the n features in the example

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

The weight **W** is also an n-dimensional vector.

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix}$$

The bias **b** is a real number.

$$b$$

acm.ai

Loss function

$$L(\hat{y_1}, \hat{y_2}, \ldots \hat{y_m}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2$$

yhat$_i$ is the output of your model (**prediction**),

y$_i$ is the actual value (**target**),

all for training example number **i**

acm.ai

Gradient Descent

Use gradient descent on loss function to determine how to change each of the weights!

$$\frac{\delta J}{\delta w_j} = \frac{2}{m}\sum_{i=1}^{m}(\hat{y_i} - y_i)x_{ij}$$

$$w_j = w_j - \alpha\frac{\delta J}{\delta w_j}$$

$$\frac{\delta J}{\delta b} = \frac{2}{m}\sum_{i=1}^{m}(\hat{y_i} - y_i)$$

$$b = b - \alpha\frac{\delta J}{\delta b}$$

i refers to the ith training sample, **j** refers to the jth feature

Quick Poll: Linear Regression Review

What is linear regression well suited for?

a. Determining whether an image is a dog or a cat
b. Predicting whether a tumor is benign or harmful
c. Creating Siri
d. Predicting MCAT scores based on college GPA

Logistic regression

Motivation

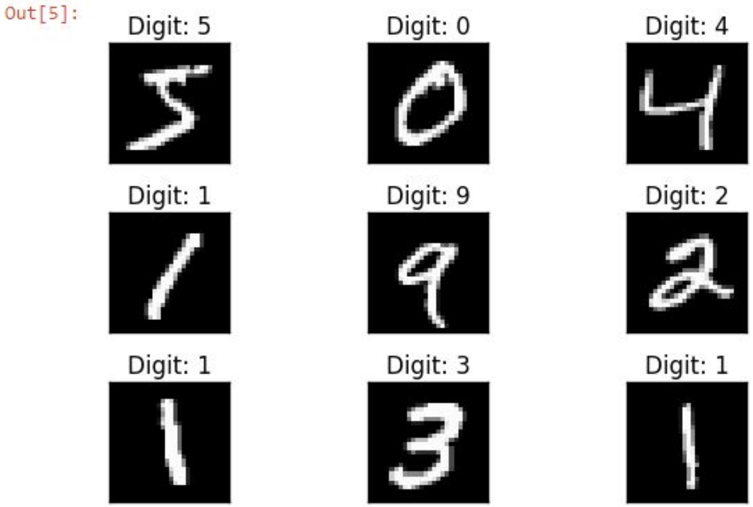
- Linear Regression: predict continuous data (eg: house price)
- Now, want to classify data (this or that)
- Use Logistic Regression

Output: $250,000

Output: Cat

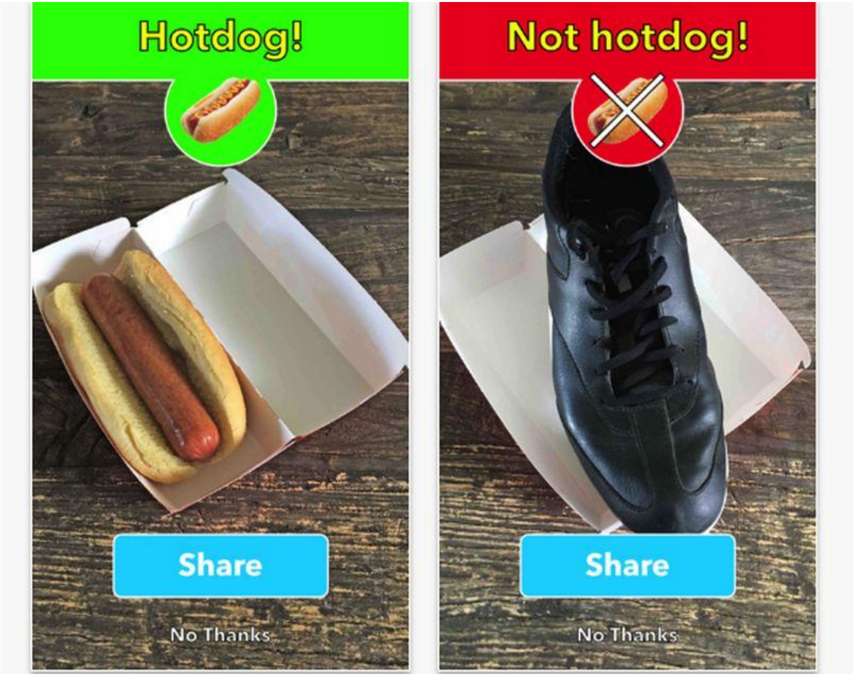
Examples of classification tasks

- Tumors - benign or malignant
- Numerical digits - one or two or three or …



Input and Output for binary classification

- Input for image classification task:
  - Array of pixels - the image
  - Each pixel is a feature
- Output for image classification task:
  - Class label: **0** or **1**
  - 0 - Cat, 1 - dog
  - 0 - benign, 1 - malignant
  - 0 - hotdog, 1 - not hotdog



acm.ai

An Example Problem

- Suppose we want to predict if it is going to rain today or not.

- What kind of features would we look at?
  - Location
  - Temperature
  - Wind Speed
  - Humidity
  - Cloud Cover

- Given these *features* how would we determine whether it is going to rain?
- **Logistic Regression!**

acm.ai

Example Problem

Suppose we are given the following data:

- Location: Los Angeles
- Temperature: 105F
- Cloud cover : Low
- Humidity: 50%
- Wind Speed: 10mph

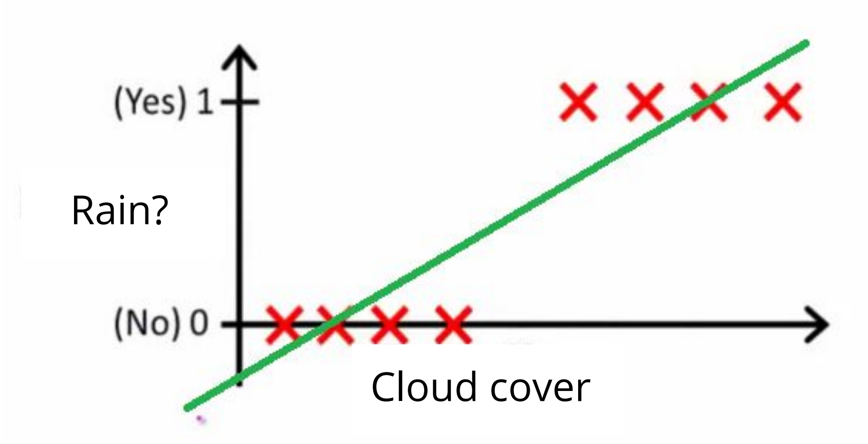
Is it more likely to rain or not to rain?

It's probably not going to rain!
I.e. the probability of it raining is less than **50%**

acm.ai

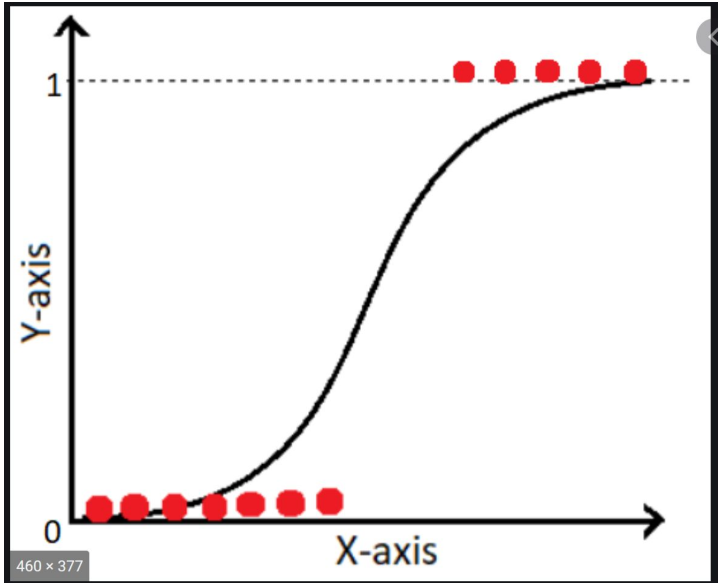
Logistic Regression

- We want our hypothesis to be a function that learns a boundary between two classes.
- But our linear function from before doesn't do the job that well.



Logistic Regression

- The solution is to **activate** the output of our previous linear model using a **nonlinear** function.



Activation Functions

acm.ai

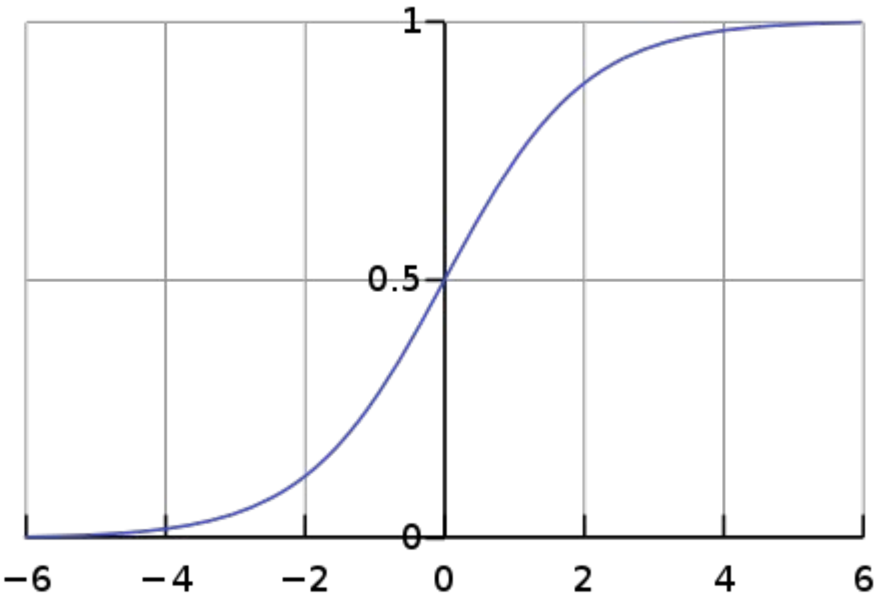
Sigmoid Function

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

If x is negative,
sigma(x) < 0.5

If x is positive,
sigma(x) > 0.5

Also,
0 <= sigma(x) <= 1



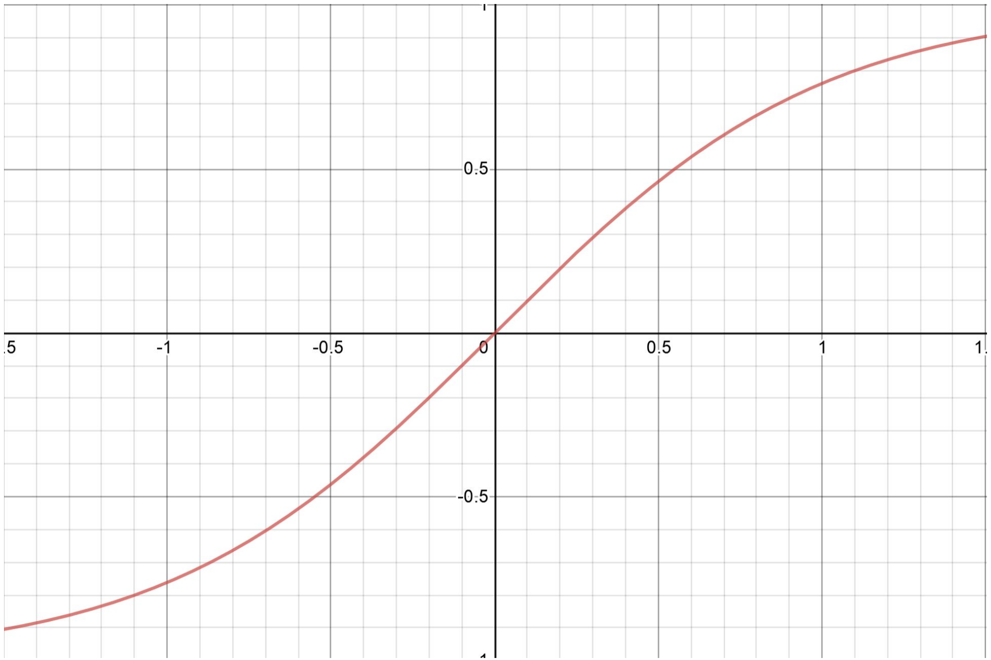
Hyperbolic Tangent (tanh)

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

If x is negative,
tanh(x) is negative

If x is positive,
tanh(x) is positive

-1 <= tanh(x) <= 1

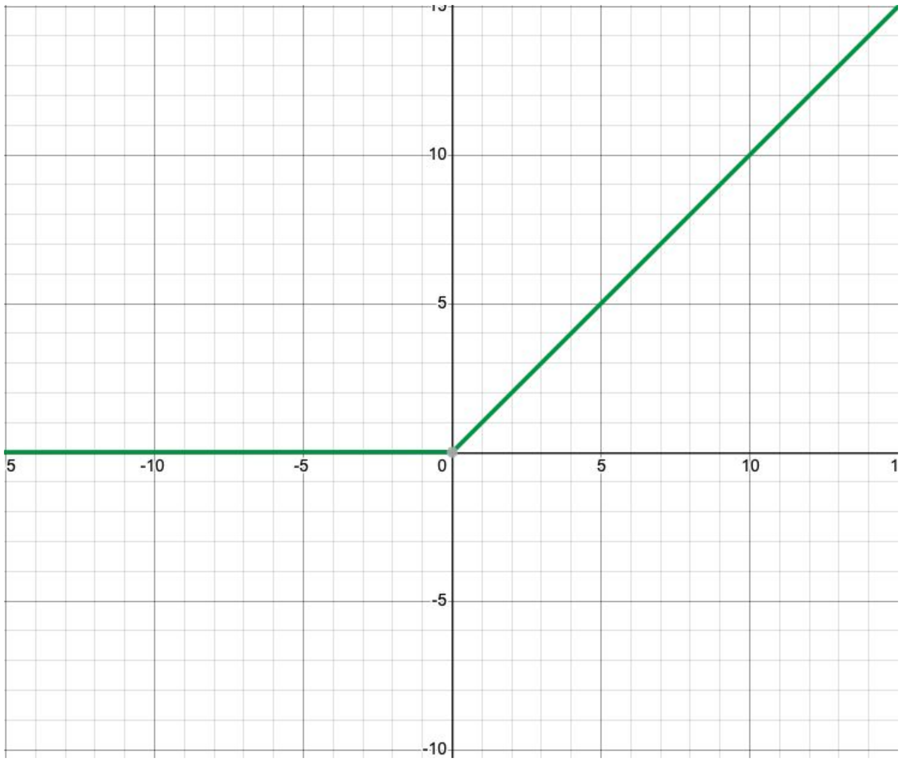


Rectified Linear Unit (ReLU)

ReLU(x) = max(0,x)

If x is negative,
ReLU(x) = 0

If x is positive,
ReLU(x) = x



acm.ai

Quick Poll: Sigmoid Function

What is the correct formula for a sigmoid function?

A) $$\frac{1}{1+e^{x}}$$

B) $$\frac{1}{1+e^{-x}}$$

C) $$\frac{1}{1-e^{-x}}$$

D) $$\frac{1}{1-e^{x}}$$

Quick Poll: RELU Function

What is the output range of the RELU function ?

a. [0, inf)
b. (-inf, 0]
c. [0, 1]
d. (1, inf)

acm.ai

Building the model

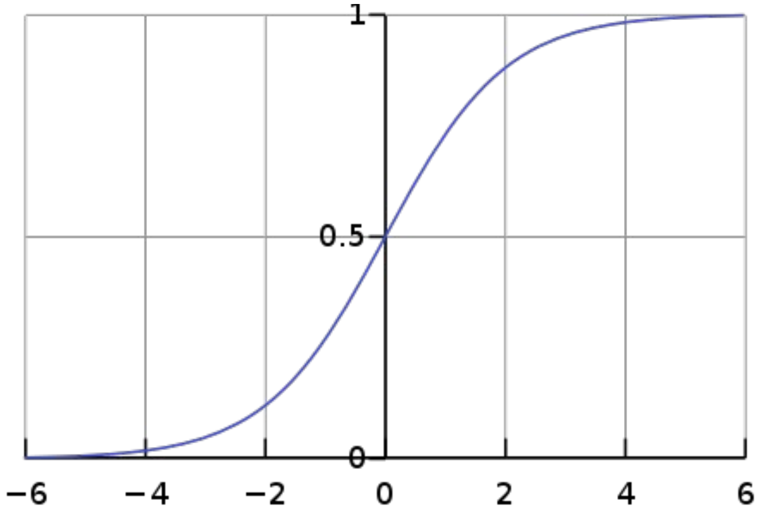
The hypothesis

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

Instead of x, let's use our old linear function as the input

$$h(x) = W^T X + b$$

$$\hat{y}(x) = \sigma(W^T X + b) = \frac{1}{1+e^{-(W^T X + b)}}$$



So depending on the values of **W** and **b**, an input **X** will result in a prediction **yhat** that is either greater than 0.5 or lesser than 0.5

If yhat<0.5 we can classify it as **0**
If yhat>0.5 we can classify it as **1**

acm.ai

Probability of being a particular class

- Think of the output of the model as the **probability** of the input being class 1 given the features X.

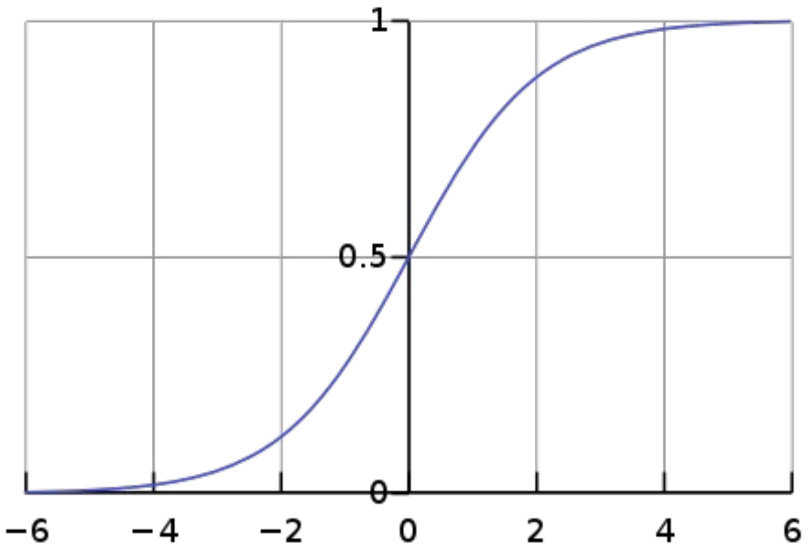
$$\hat{y}(x) = P(Y = 1|X)$$

- This is read as “Probability that the label **Y** is **1** given the features **X** we have”

So what do we need to find?

$$\hat{y}(x) = \frac{1}{1 + e^{-(W^T X + b)}}$$

- We need to find the **decision boundary**
- That is, we must learn **W** and **b** such that an input **X** when transformed, is correctly classified as **0** or **1**
- How do we do this?
- Gradient descent on cost function!



acm.ai

Cost Function: Binary Cross-Entropy Loss

- Why not use the linear regression cost function (MSE)?

- Instead, we use **Binary Cross-Entropy Loss** or **Log Loss**

$$L(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

Cost Function: Binary Cross-Entropy Loss

$$L(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

- $\hat{y}$: prediction.
- y : label
- What happens when **y** is **1**? $L(\hat{y}, y) = -\log(\hat{y})$
- What happens when **y** is **0**? $L(\hat{y}, y) = -\log(1 - \hat{y})$

Cost Function: Binary Cross-Entropy Loss

So the total cost across all the samples becomes :

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y_i}, y_i)$$

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} -y_i \log(\hat{y_i}) - (1-y_i)\log(1-\hat{y_i})$$

acm.ai

Quick poll: Correct Loss Function

Which of the following is the correct loss function for binary classification?

a. (1-y)log(yhat) + (y)log(1-yhat)
b. -(y)log(yhat) - (1-y)log(1-yhat)
c. (y)log(yhat) + (1-y)log(1-yhat)
d. -(1-y)log(yhat) - (y)log(1-yhat)

Gradient Descent

The derivatives **dJ /dw** and **dJ / db** are similar to those in linear regression.

$$\frac{\delta J}{\delta w} = \frac{1}{m}(\hat{Y} - Y)X^T \qquad w = w - \alpha\frac{\delta J}{\delta w}$$

$$\frac{\delta J}{\delta b} = \frac{1}{m}\sum(\hat{Y} - Y) \qquad b = b - \alpha\frac{\delta J}{\delta b}$$

Yhat is a row vector (1 x n_samples) containing all the predictions, ***Y*** is a row vector (1 x n_samples) with the labels, and **X** is the matrix of features (n_features *x* n_samples)

Answer to Polls

1. D
2. B
3. A
4. B

Thank you! We'll see you next week!

Please fill out our feedback form: tinyurl.com/btrackfeedback3

Next week: Multiclass Classification & K-Nearest Neighbors

Don't worry, it's okay if data points are close together!

Today's event code: **findingnemo**

FB group: facebook.com/groups/uclaacmai

Github: github.com/uclaacmai/beginner-track-fall-2020

acm.ai