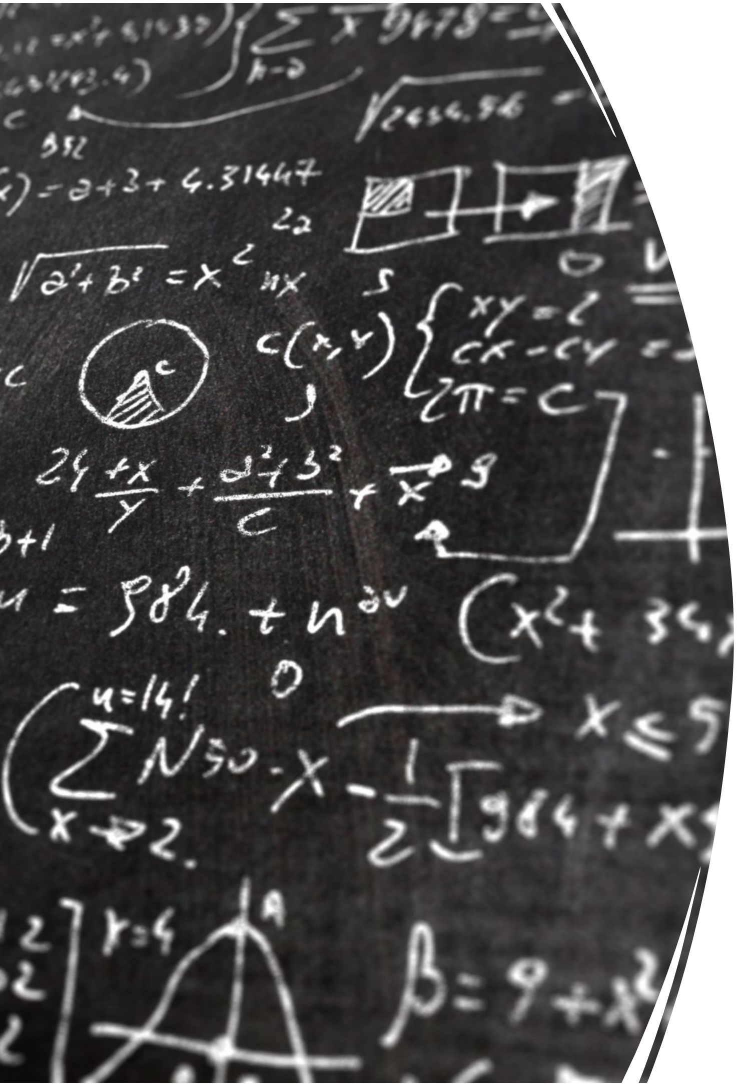


CS 383 – Computational Text Analysis

Lecture 8
Logistic Regression

Adam Poliak

02/13/2023

Slides adapted Dan Jurafsky, Jordan Boyd-Graber, Nate Chambers

Announcements

- Reading 03 released Monday
  - Due Monday 02/13
- HW03 due Wednesday 02/15
- Reading 04
  - Due Monday 02/20 – Dictionary Methods
- HW04
  - Likely due next Friday
  - depends on today and Wednesday’s progress
  - Not committing

Course Outline

- Unsupervised approaches
  - LMs
  - DTM
    - Tf-idf
  - Clustering
    - Dimensionality reduction
    - Topic modeling
- Prediction
- Data Collection
- Hypothesis Testing

Outline

Linear Regression

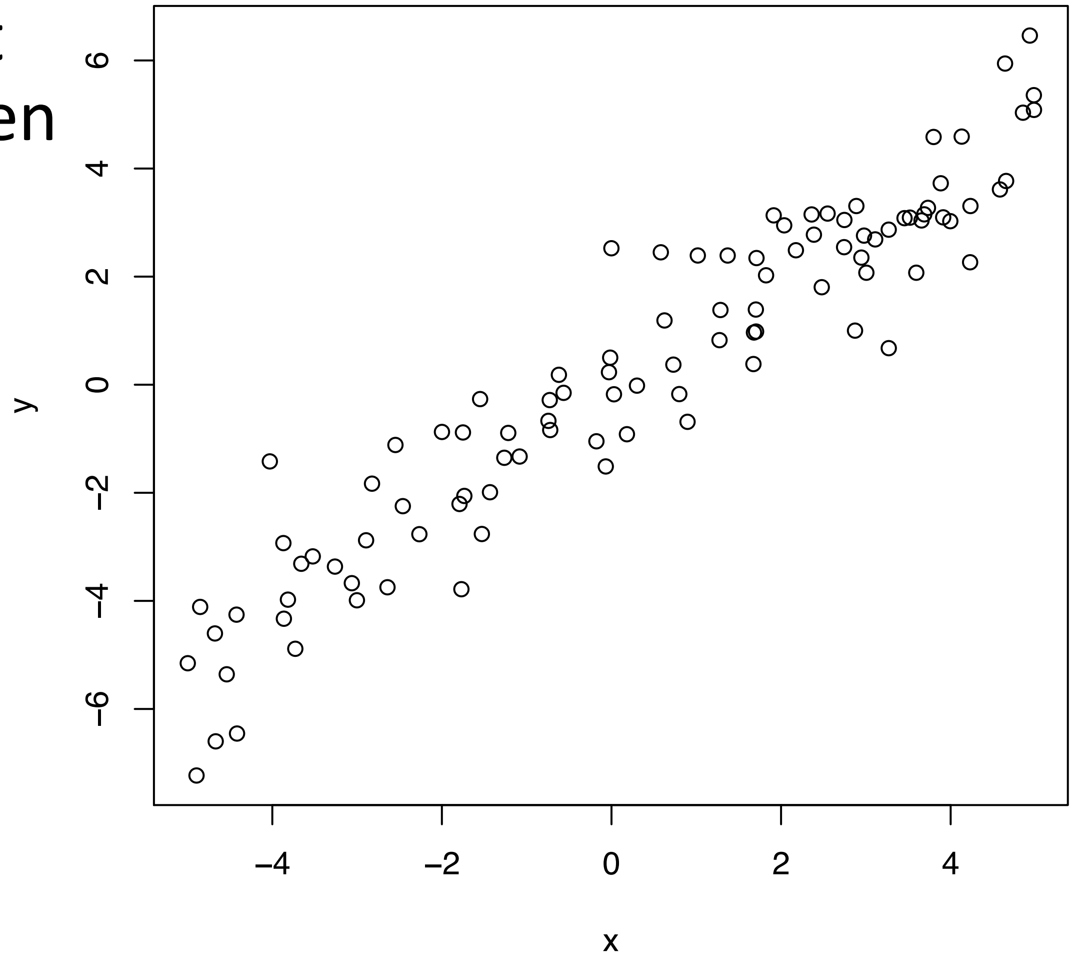
Evaluation

Logistic Regression

Learning weights

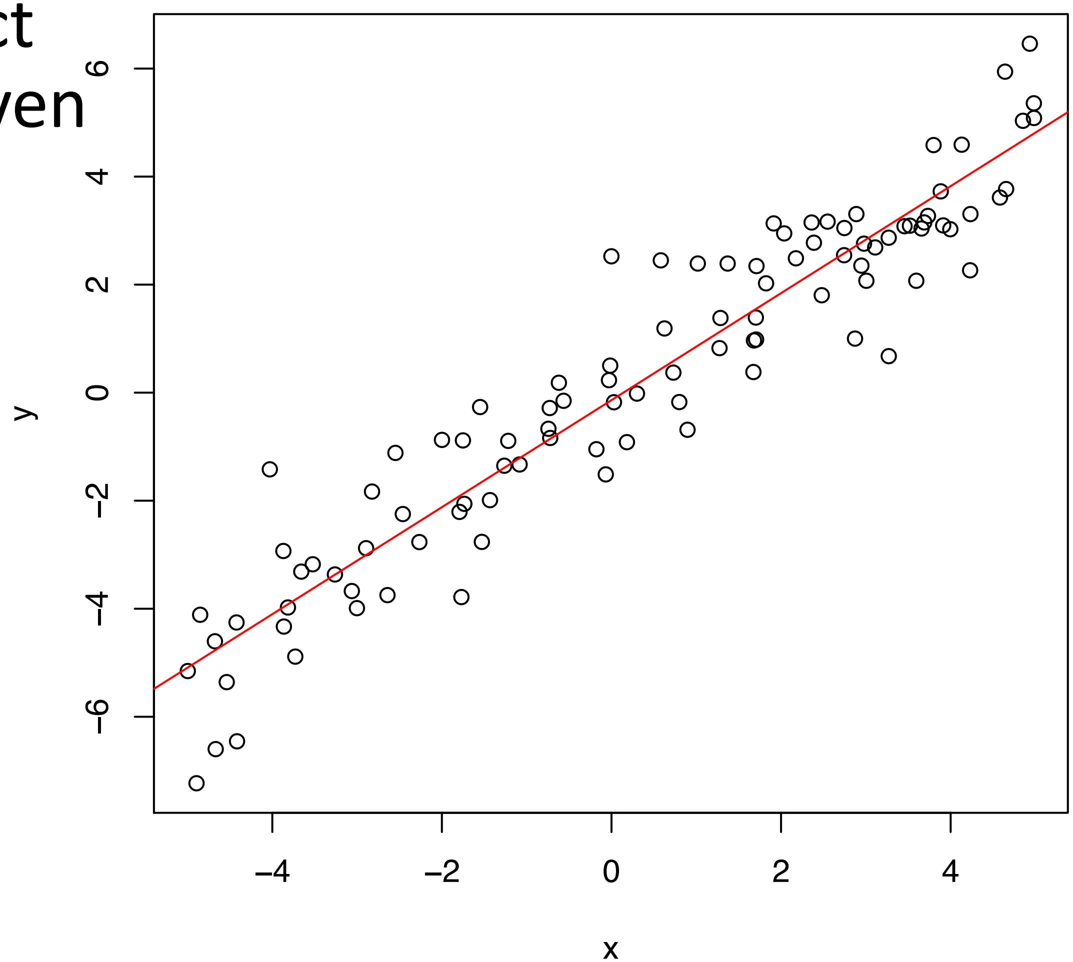
Linear Regression

- Goal is to predict real-valued y given x using a linear function



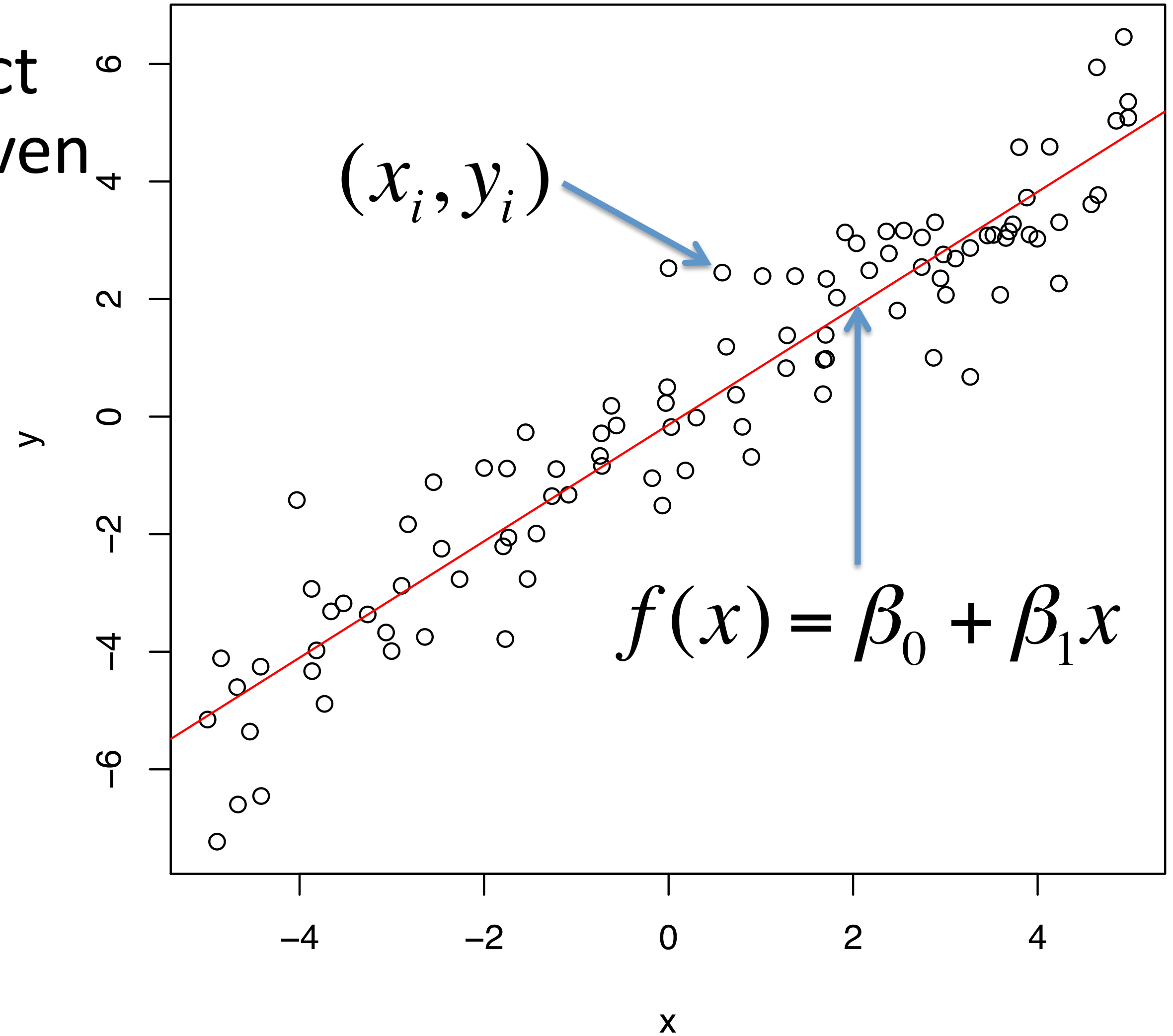
Linear Regression

- Goal is to predict real-valued y given x using a linear function



Linear Regression

- Goal is to predict real-valued y given x using a linear function



Linear Regression

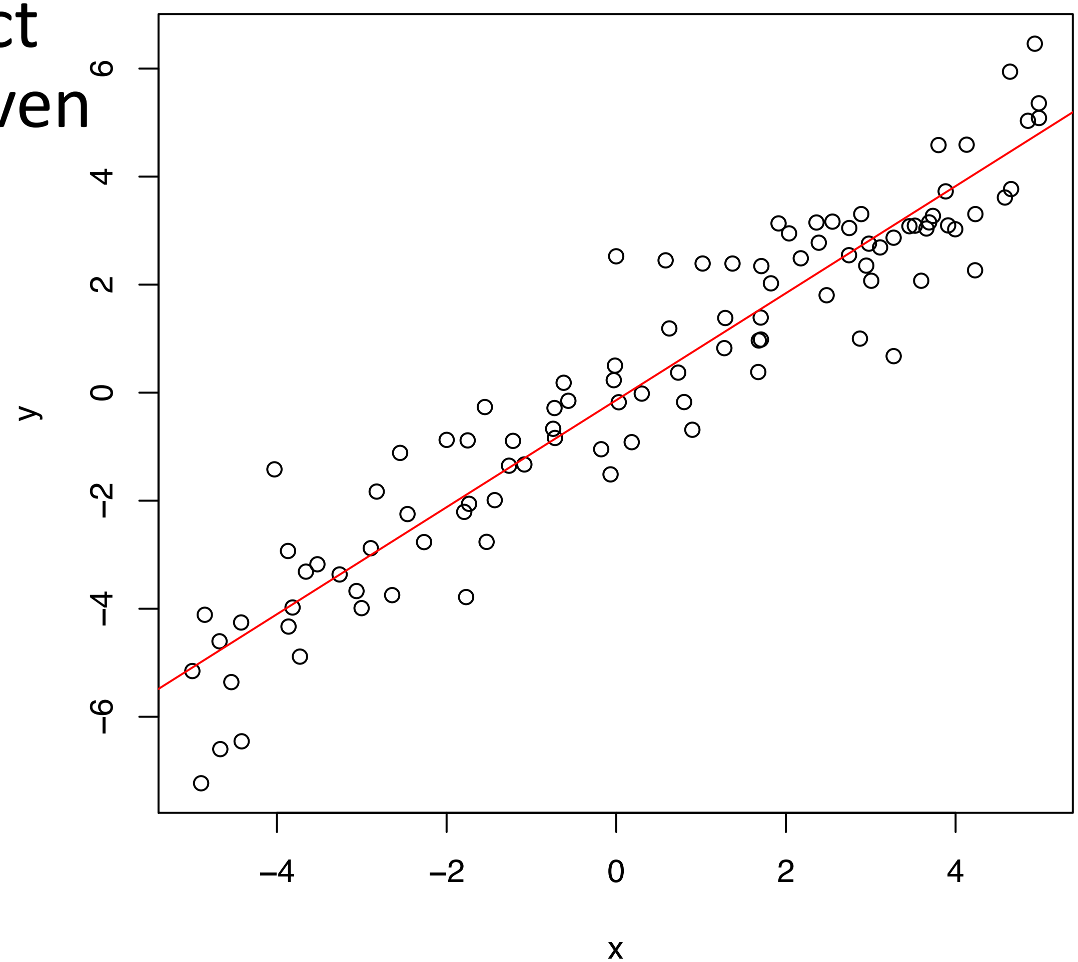
- Goal is to predict real-valued y given x using a linear function

- Examples:
  - Given browsing history, how long will a user stay on a webpage
  - Given a tweet, predict the sentiment
  - …

Linear Regression

- Goal is to predict real-valued y given x using a linear function

- What is x?



Multiple features/covariates

- Represent each datapoint as a vector, each value in the vector represents a *feature*

$$x = (x_0, x_1, x_2, x_3, \ldots., x_p)$$

- Predict *y* by fitting a function that is a linear combination of the

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Multiple features/covariates

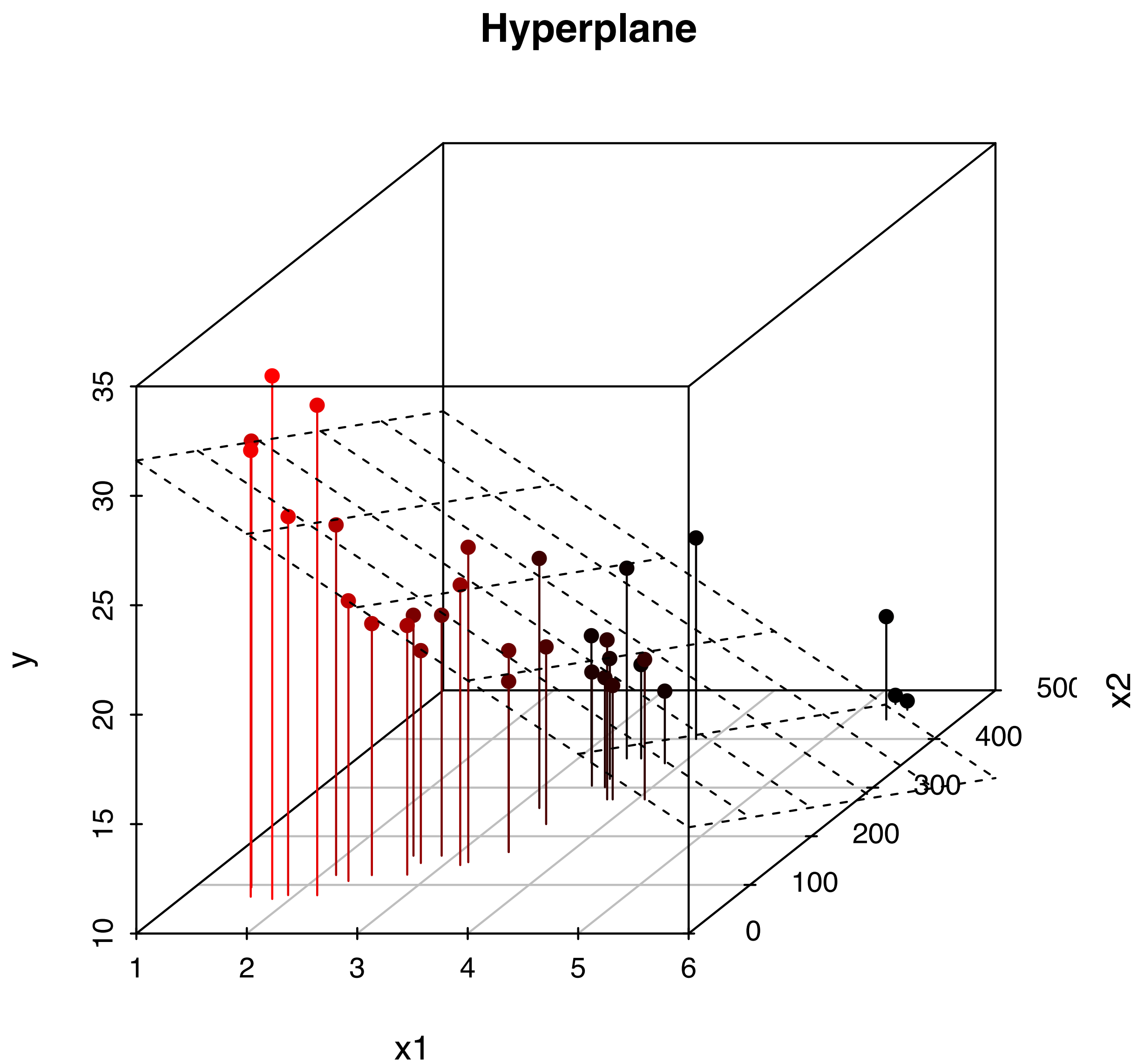
- Predict *y* by fitting a function that is a linear combination of the

$$f(x) = \sum_{j=1}^{p} \beta_j x_j$$

- Since x is a vector, so is β

- What then is the equation?
  - Dot-product

Multiple features/covariates

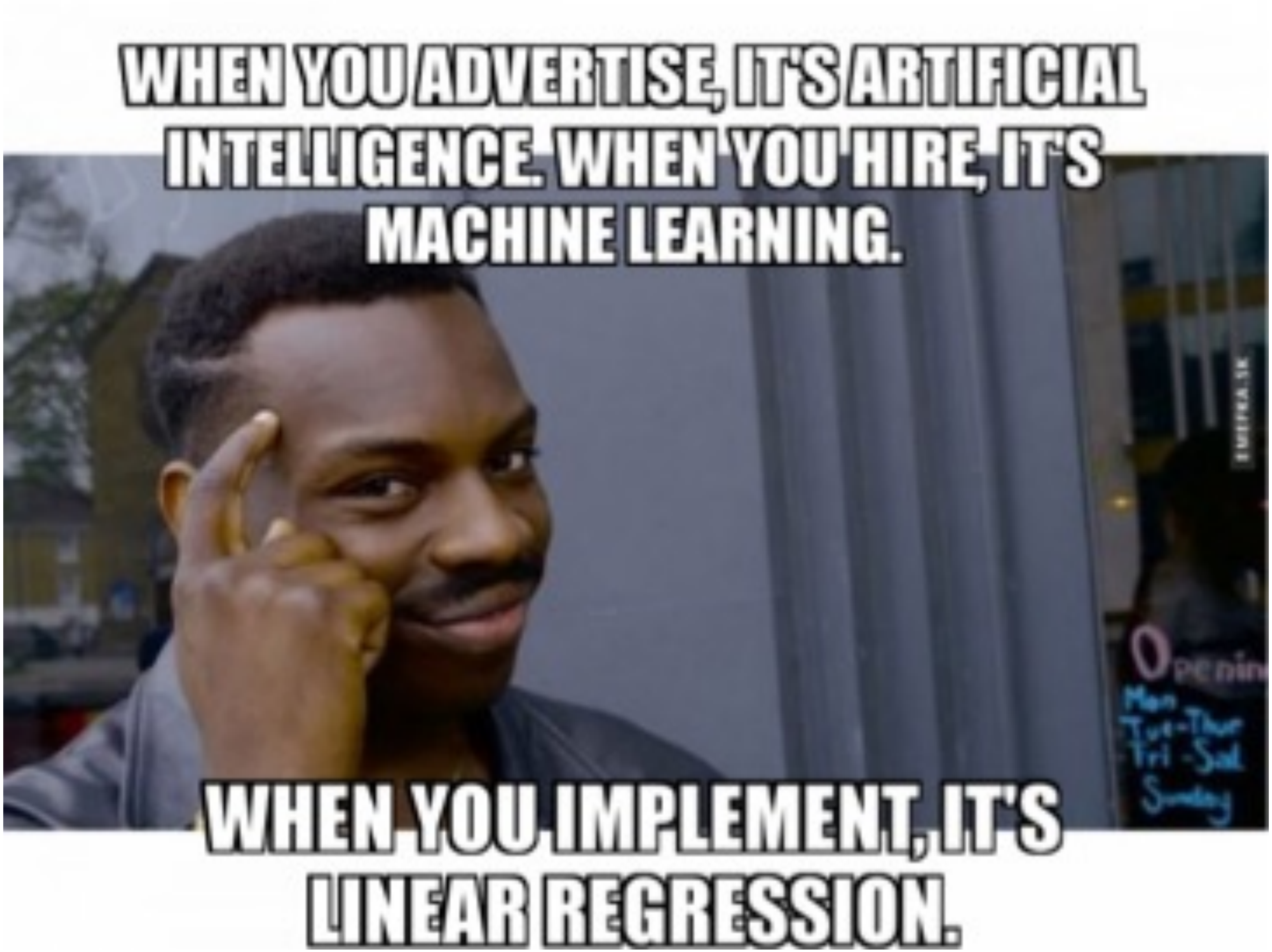


Features/Covariates

- When predicting text, what might the features be?
  - Counts of n-grams

- They can be other values because for word counts:
  - Transformations:
    - tf-idf values
    - log of counts
  - Indicator variables
    - Does the sentence mention X
  - Interactions of variables
    - Number of times mentions function words

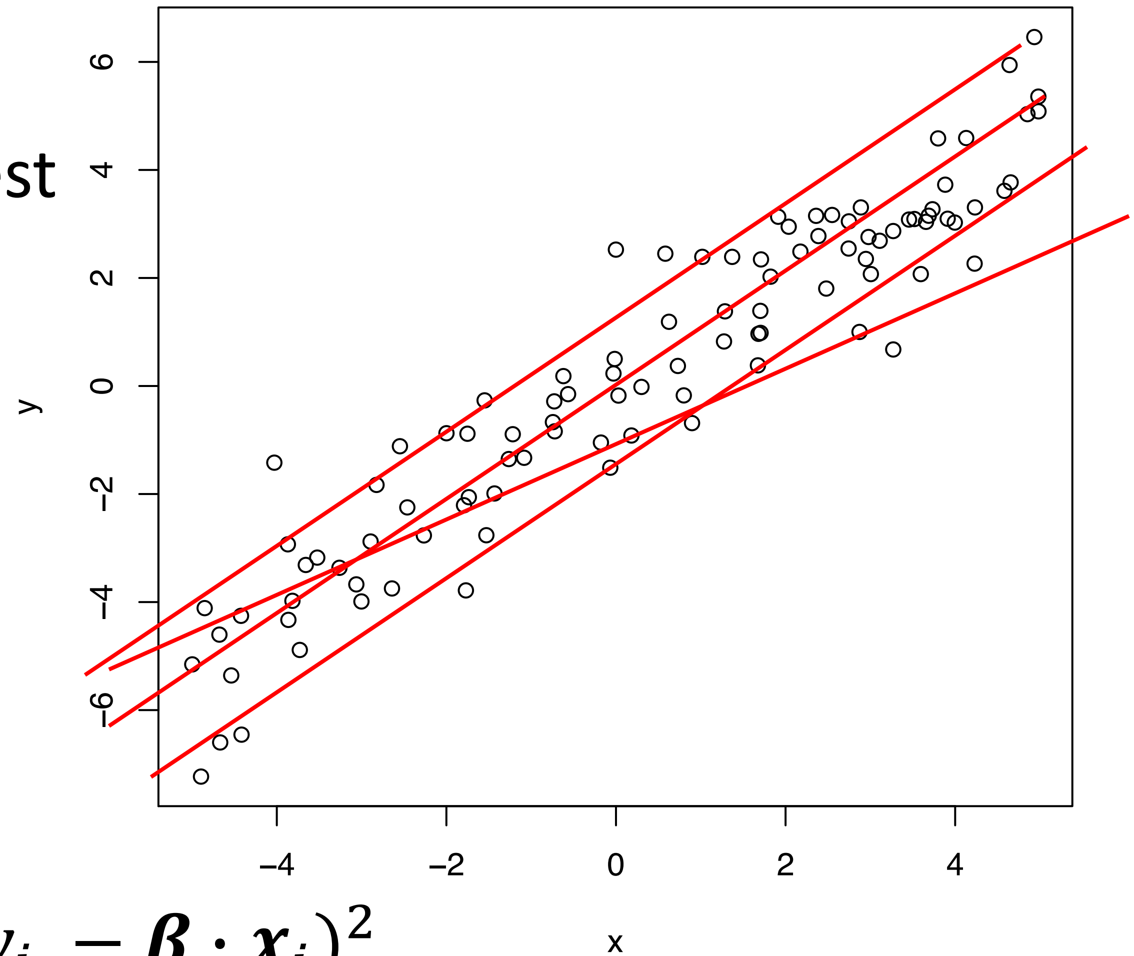
- Because of its simplicity and flexibility, linear regression is one of the most widely implemented regression techniques


WHEN YOU ADVERTISE, IT'S ARTIFICIAL INTELLIGENCE. WHEN YOU HIRE, IT'S MACHINE LEARNING.
WHEN YOU IMPLEMENT, IT'S LINEAR REGRESSION.

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

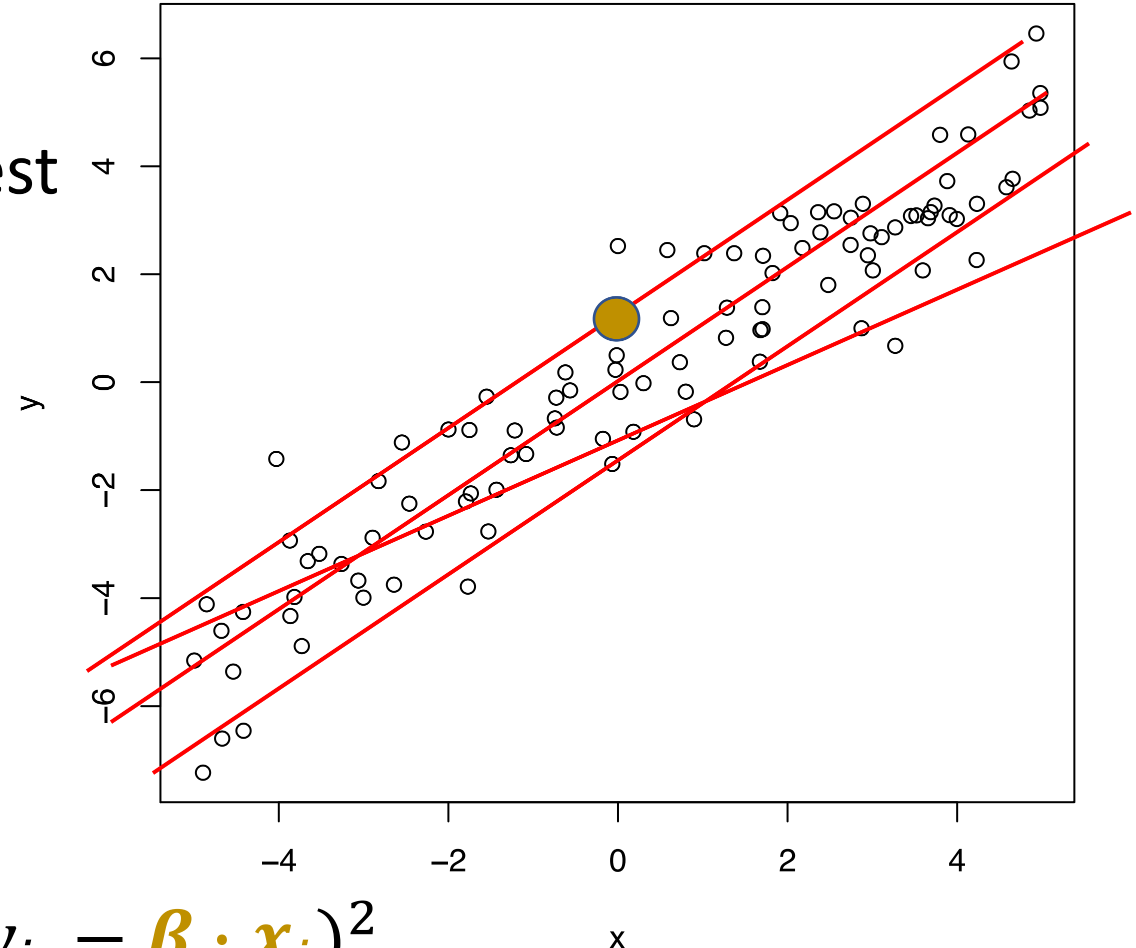


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

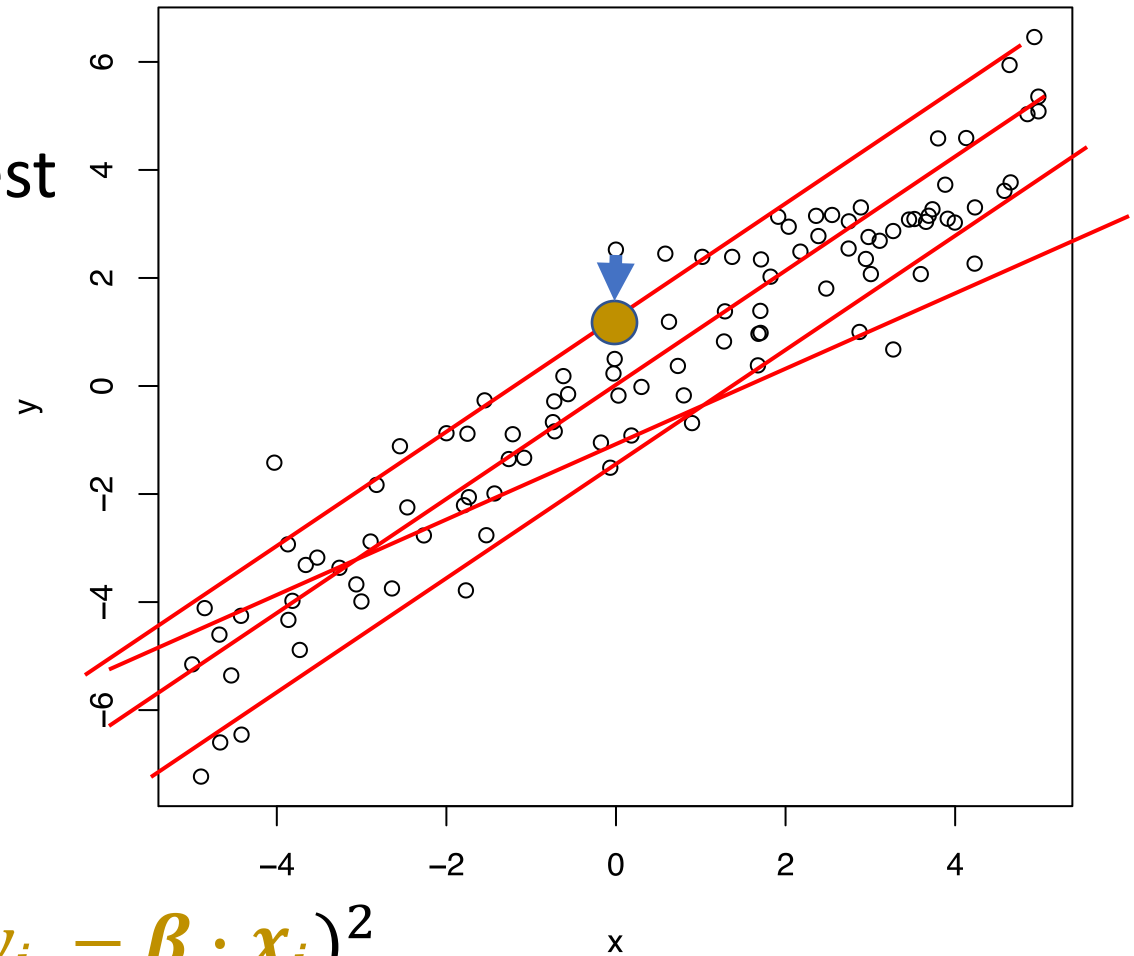


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

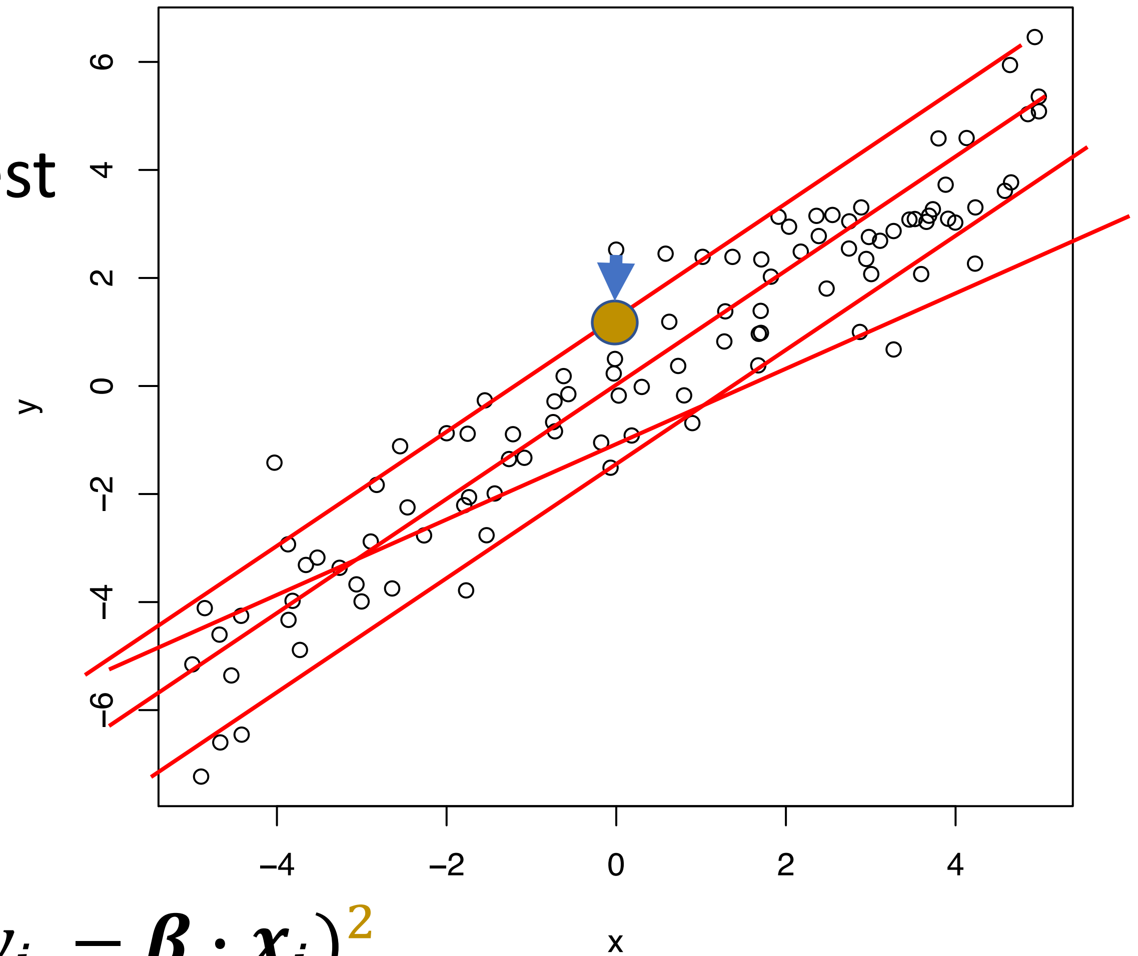


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

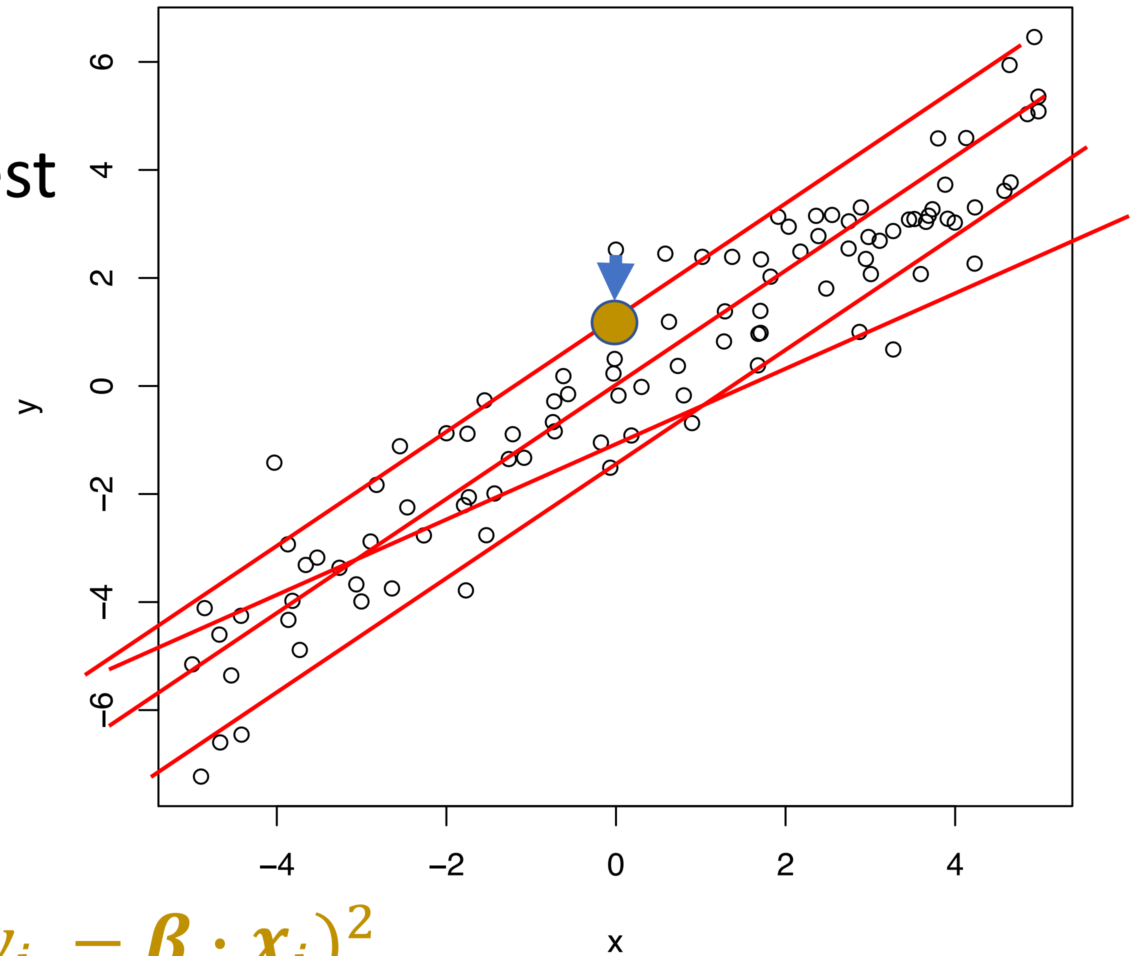


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}\cdot\boldsymbol{x}_i)^2$$

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

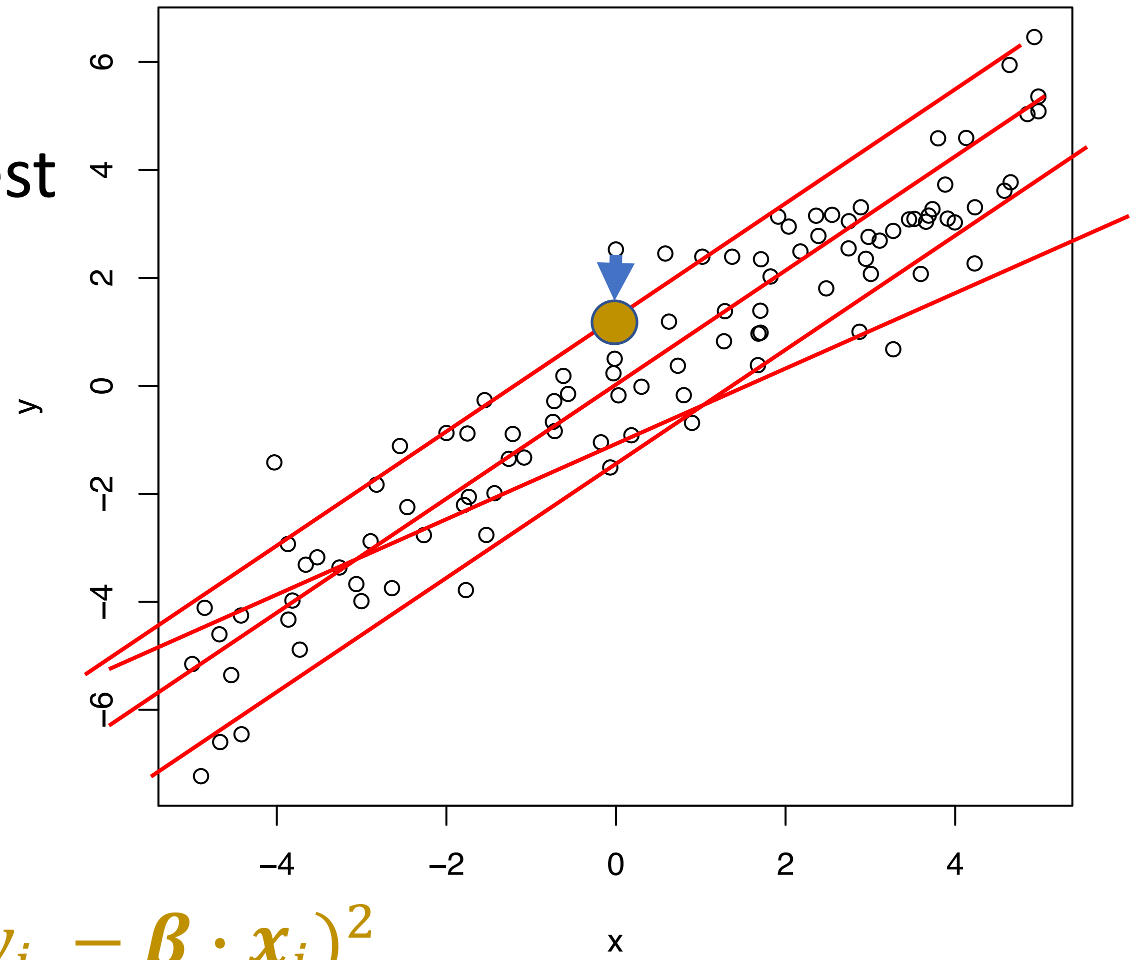


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

Which line?

Which line is the best "fit" to describe the data?

Idea: minimize the Euclidean distance between data and fitted line

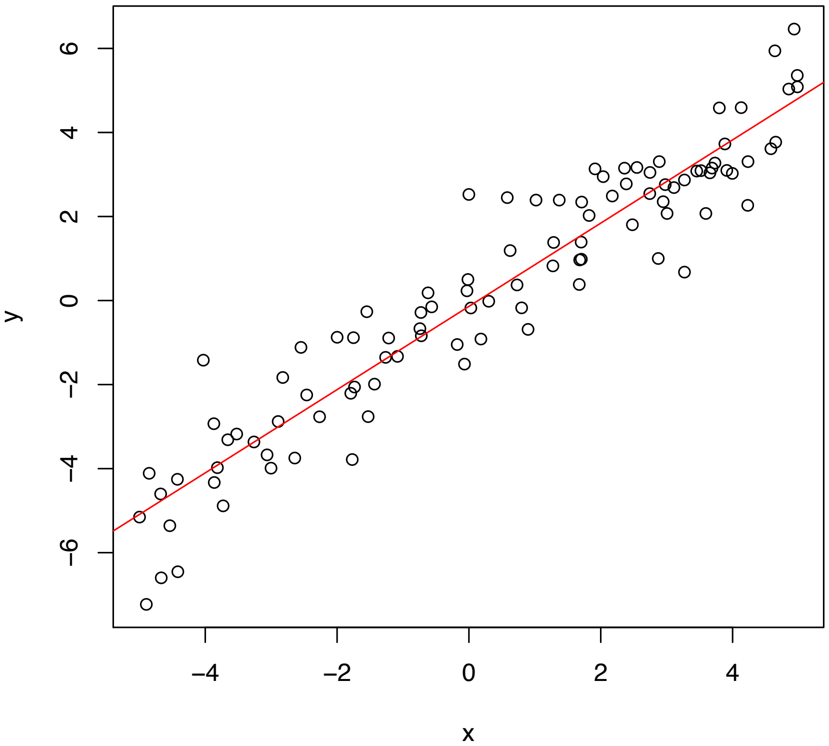


$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

Finding β?

- Use calculus to find β that minimizes RSS
- Or use the closed-form solution:

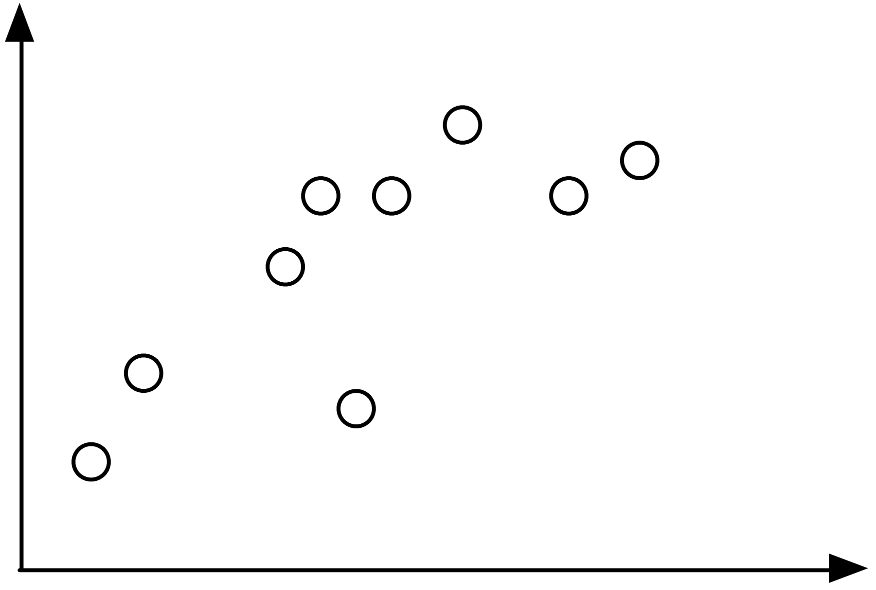
$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$



Predicting

- For a given x predict $\widehat{y}$ where

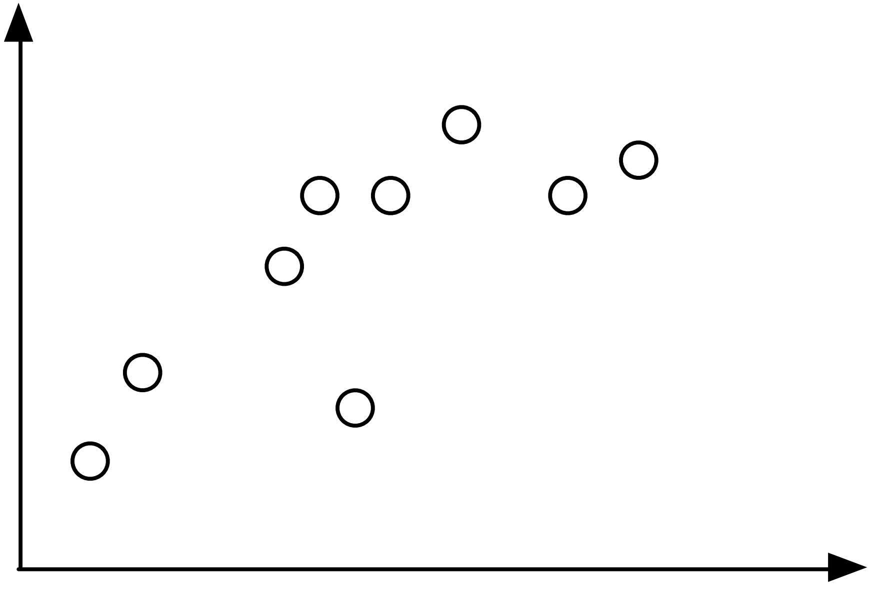
$$\hat{y} = \beta_0 + \beta_1 x$$



Predicting

- If x contains multiple features/covariates

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_j$$



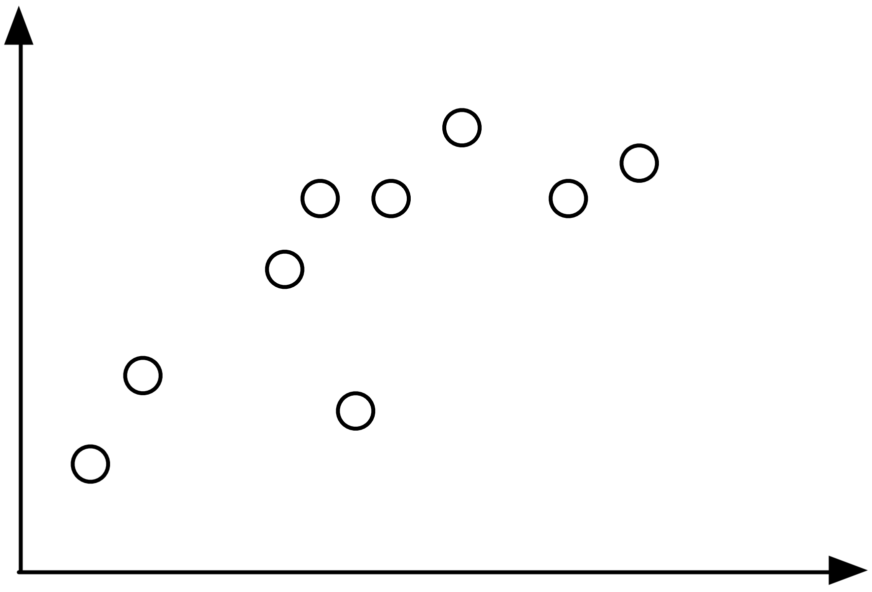
Predicting

- For a given x predict $\widehat{y}$ where

$$\hat{y} = \beta_0 + \beta_1 x$$

$\beta_0 = 1.0$
$\beta_1 = 0.5$

What line?



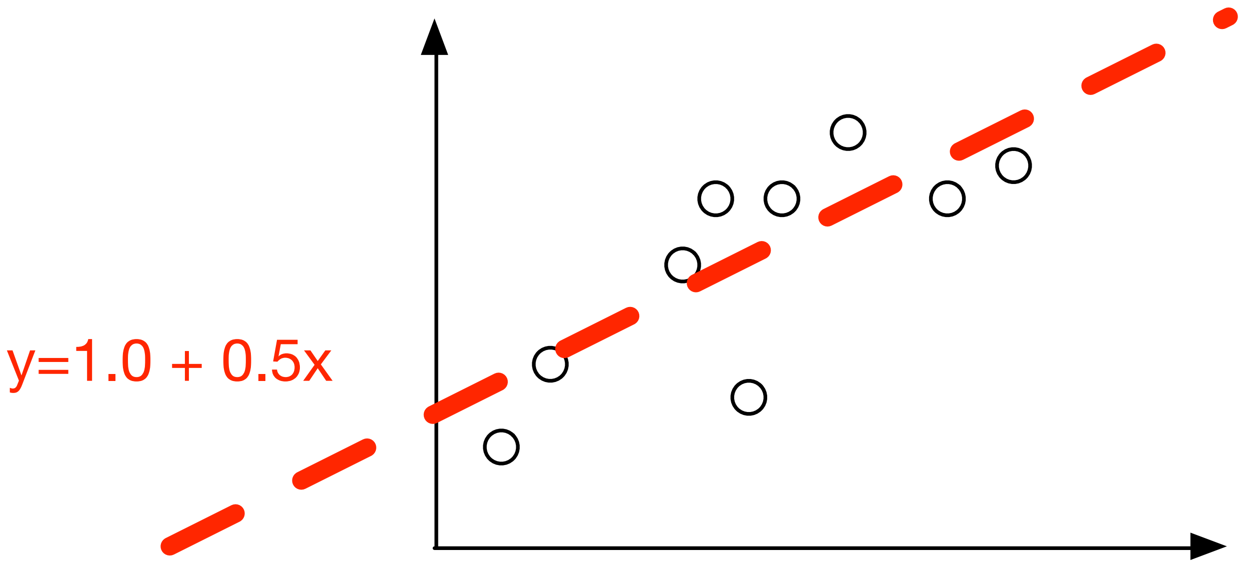
Predicting

- For a given x predict $\hat{y}$ where

$$\hat{y} = \beta_0 + \beta_1 x$$

$\beta_0 = 1.0$
$\beta_1 = 0.5$

What line?



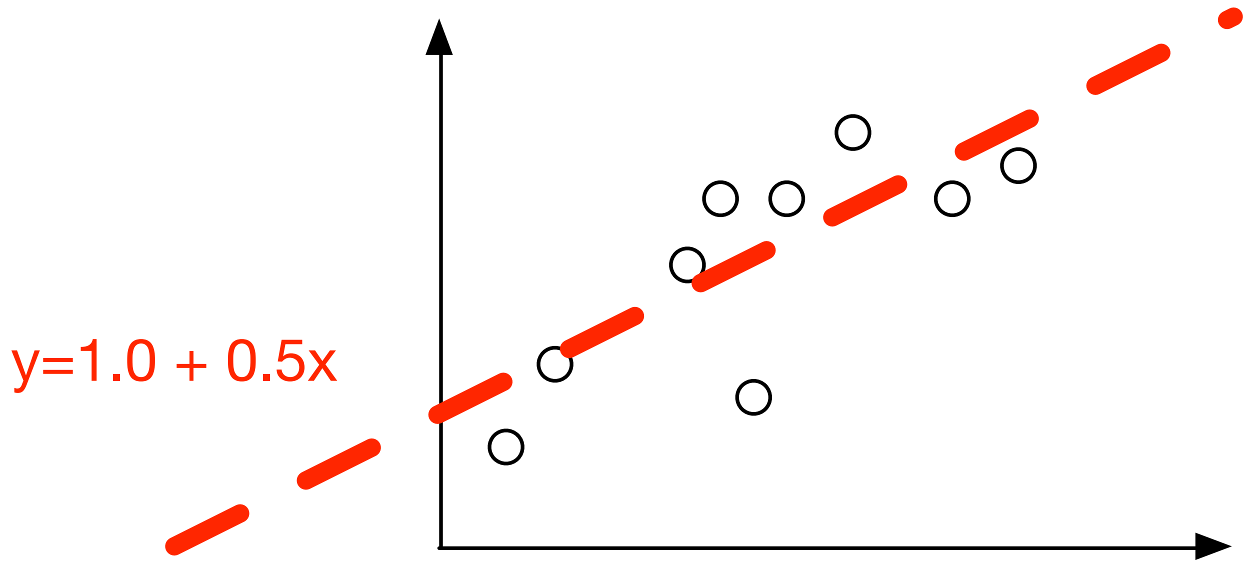
Predicting

- For a given x predict $\hat{y}$ where

$$\hat{y} = \beta_0 + \beta_1 x$$

$\beta_0 = 1.0$
$\beta_1 = 0.5$
$x = 5.0$

What's $\hat{y}$?



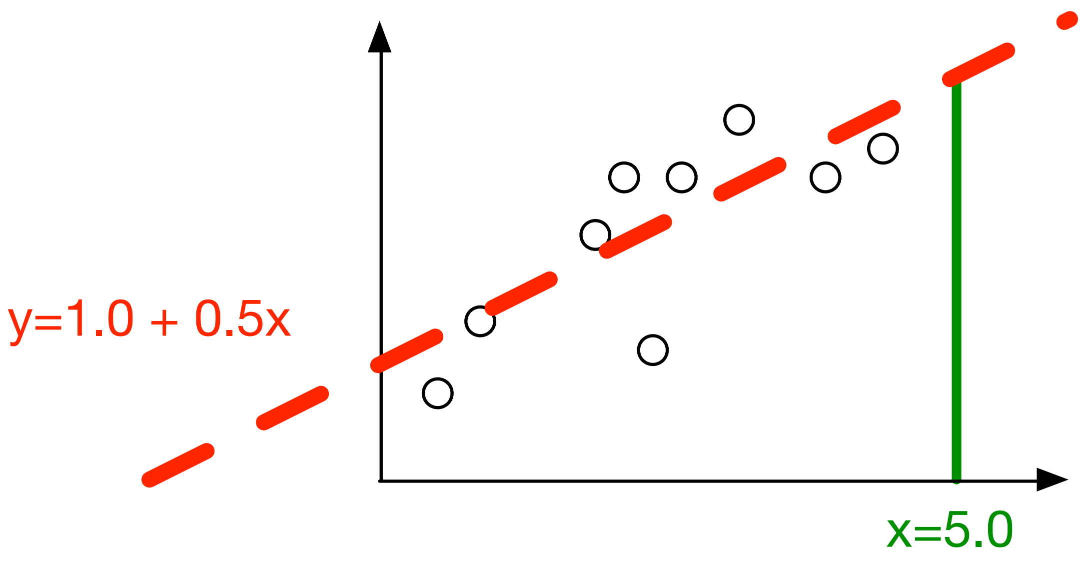
Predicting

- For a given x predict $\hat{y}$ where

$$\hat{y} = \beta_0 + \beta_1 x$$

$\beta_0 = 1.0$
$\beta_1 = 0.5$
$x = 5.0$

$\hat{y} = 3.5$



Probabilistic view

We are maximizing $P(Y_i | x_{i,} \beta)$

Minimizing RSS is equivalent to maximizing conditional likelihood

Unlike LDA this is a *discriminative* model because we are not modeling observed data

Recall LDA, we compute $P(x_i | topic, \beta, \alpha)$

Outline

Linear Regression

Evaluation

Logistic Regression

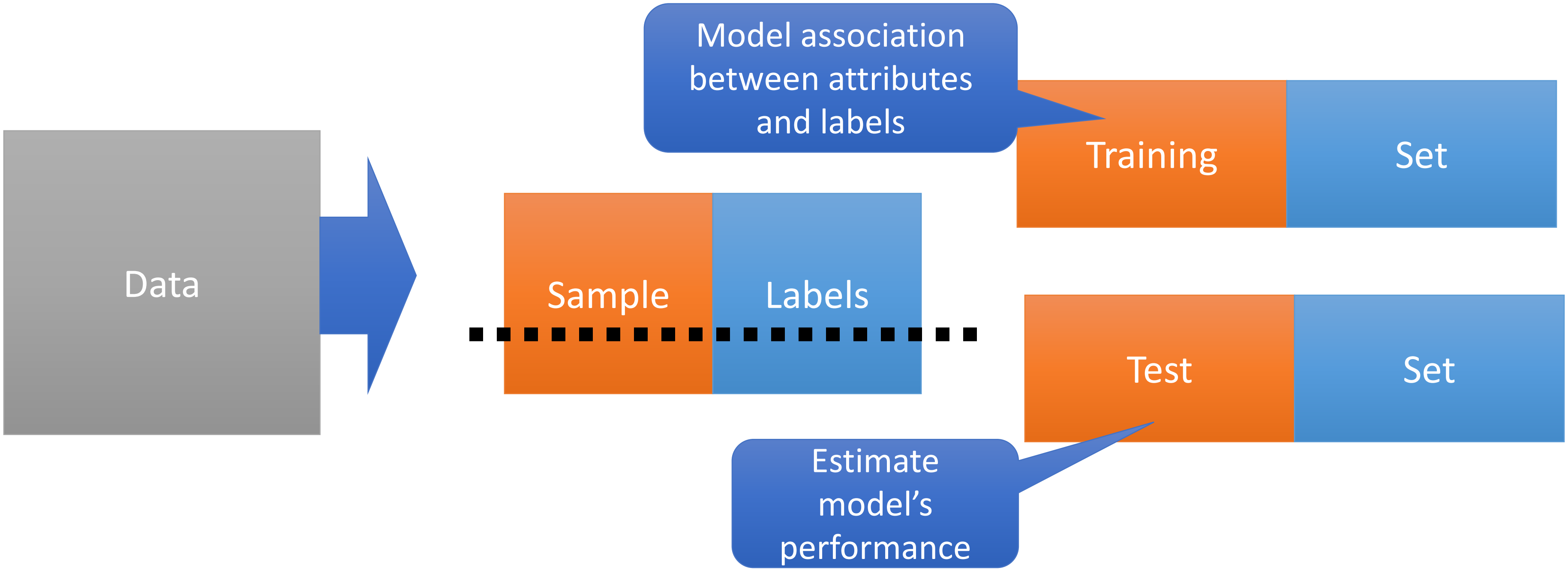
Learning weights

Training a predictor

Attributes (features) of an example

Predicted label of the example

Setup for training and evaluating a predictor



Copyright © 2016 Barnard College

Two types of predictions: Classification & Regression

Classification = Categorical

Regression = Numeric

Predicting sentiment:

- Classification

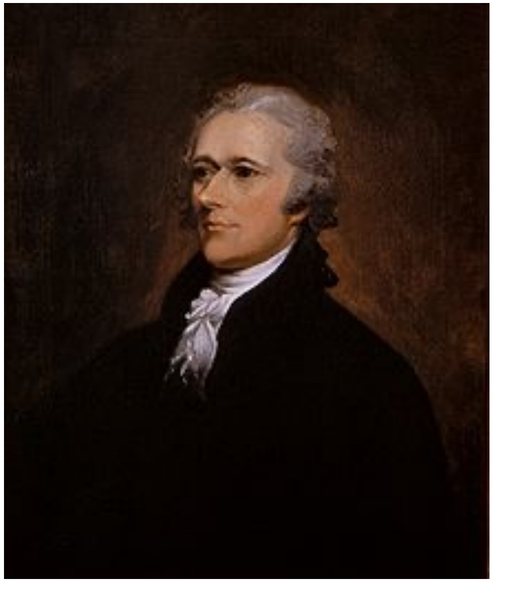
  {👍, 👎}

- Regression:

  [-1, …, 1]

Classification

- Positive/negative sentiment
- Spam/not spam
- Authorship attribution (Hamilton or Madison?)



Alexander Hamilton

Text Classification: definition

Input:

- a document x
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

Output: a predicted class $\hat{y} \in C$

Binary Classification: $\hat{y} \in \{0, 1\}$

Outline

Linear Regression

Evaluation

Logistic Regression

Learning weights

Logistic Regression

Like linear regression because we'll compute a dot product between

But we'll learn weights for each class

Logistic Regression Example

Document	Text	Author
X_1	the lady doth protest too much methinks	Shakespeare
X_2	it was the best of times it was the worst of times	Dickens

$f_7(x)$ is "the"

$f_{72}(x)$ is "the best"

$f_7(x_1) =$ 1

$f_{72}(x_1) =$ 0

$f_7(x_2) =$ 2

$f_{72}(x_2) =$ 1

Slide from Nate Chambers

Weights

Assume we have a document with the following features

$f_1(x)$ = 1

$f_2(x)$ = 2

$f_3(x)$ = 1

Slide from Nate Chambers

Weights

Assume we have a document with the following features. Goal is to classify the document as being written by Shakespeare or Dickens

$f_1(x) = 1$

$f_2(x) = 2$

$f_3(x) = 1$

Let's add weights to the features

Slide from Nate Chambers

Weights

- Now let's add *weights* to the features

	Shakespeare	Dickens
$f_1(x) = 1$	1.31	-0.23
$f_2(x) = 2$	0.49	0.72
$f_3(x) = 1$	-0.82	0.1

Slide from Nate Chambers

Weights

- Now let's add *weights* to the features
- We want a *score* for each class label

	Shakespeare	Dickens
$f_1(x) = 1$	1.31	-0.23
$f_2(x) = 2$	0.49	0.72
$f_3(x) = 1$	-0.82	0.1

Slide from Nate Chambers

Weights

- Now let's add *weights* to the features
- We want a *score* for each class label

	Shakespeare	Dickens
$f_1(x)$ = 1	1.31	-0.23
$f_2(x)$ = 2	0.49	0.72
$f_3(x)$ = 1	-0.82	0.1
	1.47	1.31

$$\text{score}(x, c) = \sum_i w_{i,c} f_i(x)$$

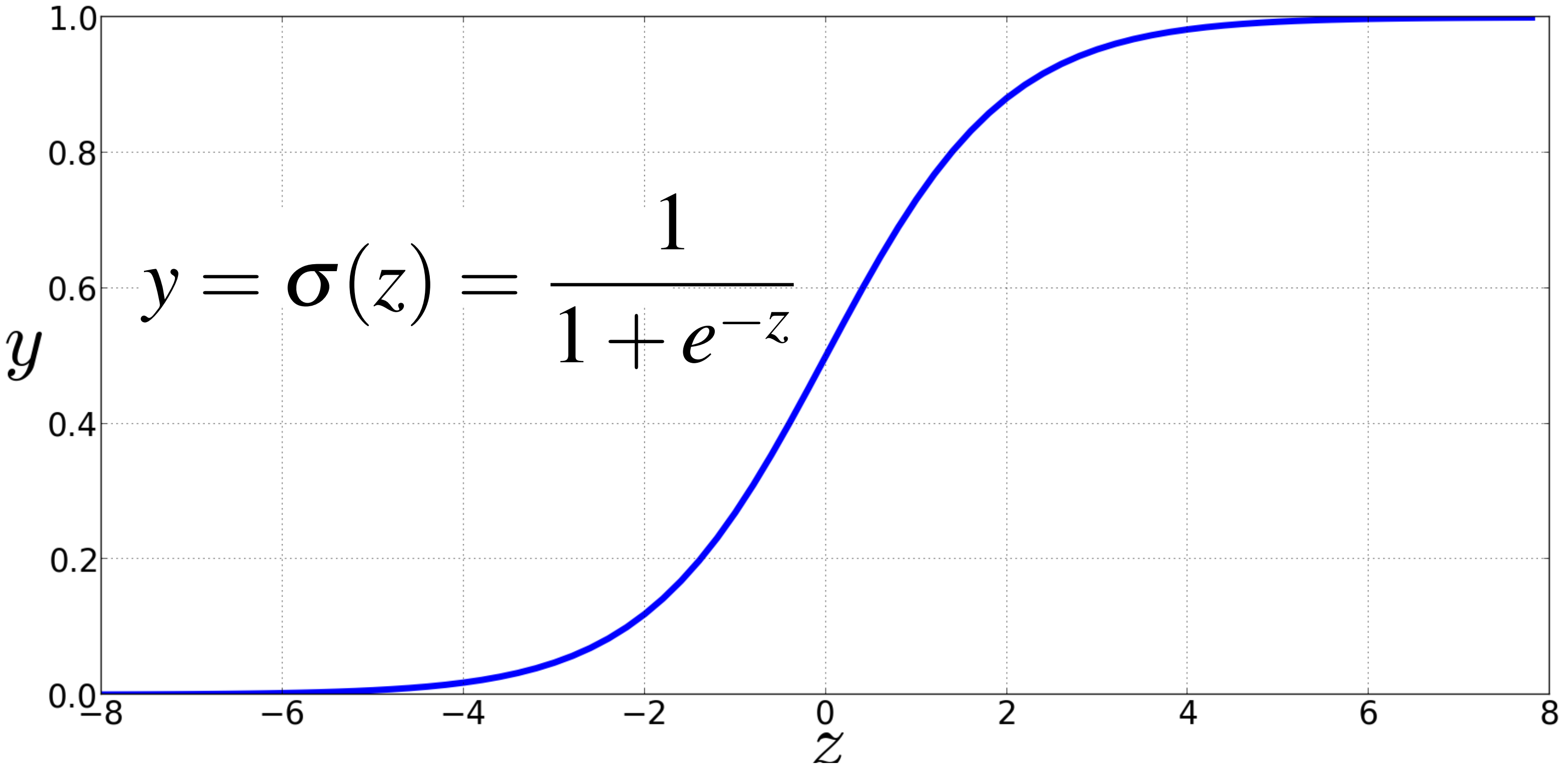
Slide from Nate Chambers

Converting scores to probabilities

Use the logit function!

$$y = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)}$$

Sigmoid/logistic function



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

Idea of logistic regression

- We’ll compute $w \cdot x + b$
- And then we’ll pass it through the sigmoid function:

$$\sigma(w \cdot x + b)$$

- And we'll just treat it as a probability

Making probabilities with sigmoids

$$
\begin{aligned}
P(y=1) &= \sigma(w \cdot x + b) \\
&= \frac{1}{1+\exp\left(-(w \cdot x + b)\right)}
\end{aligned}
$$

$$
\begin{aligned}
P(y=0) &= 1-\sigma(w \cdot x + b) \\
&= 1-\frac{1}{1+\exp\left(-(w \cdot x + b)\right)} \\
&= \frac{\exp\left(-(w \cdot x + b)\right)}{1+\exp\left(-(w \cdot x + b)\right)}
\end{aligned}
$$

By the way:

$$
\begin{aligned}
P(y=0) &= 1-\sigma(w\cdot x+b) \\
&= 1-\frac{1}{1+\exp\left(-(w\cdot x+b)\right)} \\
&= \frac{\exp\left(-(w\cdot x+b)\right)}{1+\exp\left(-(w\cdot x+b)\right)}
\end{aligned}
$$

$= \quad \sigma(-(w\cdot x+b))$

Because

$$1-\sigma(x)=\sigma(-x)$$

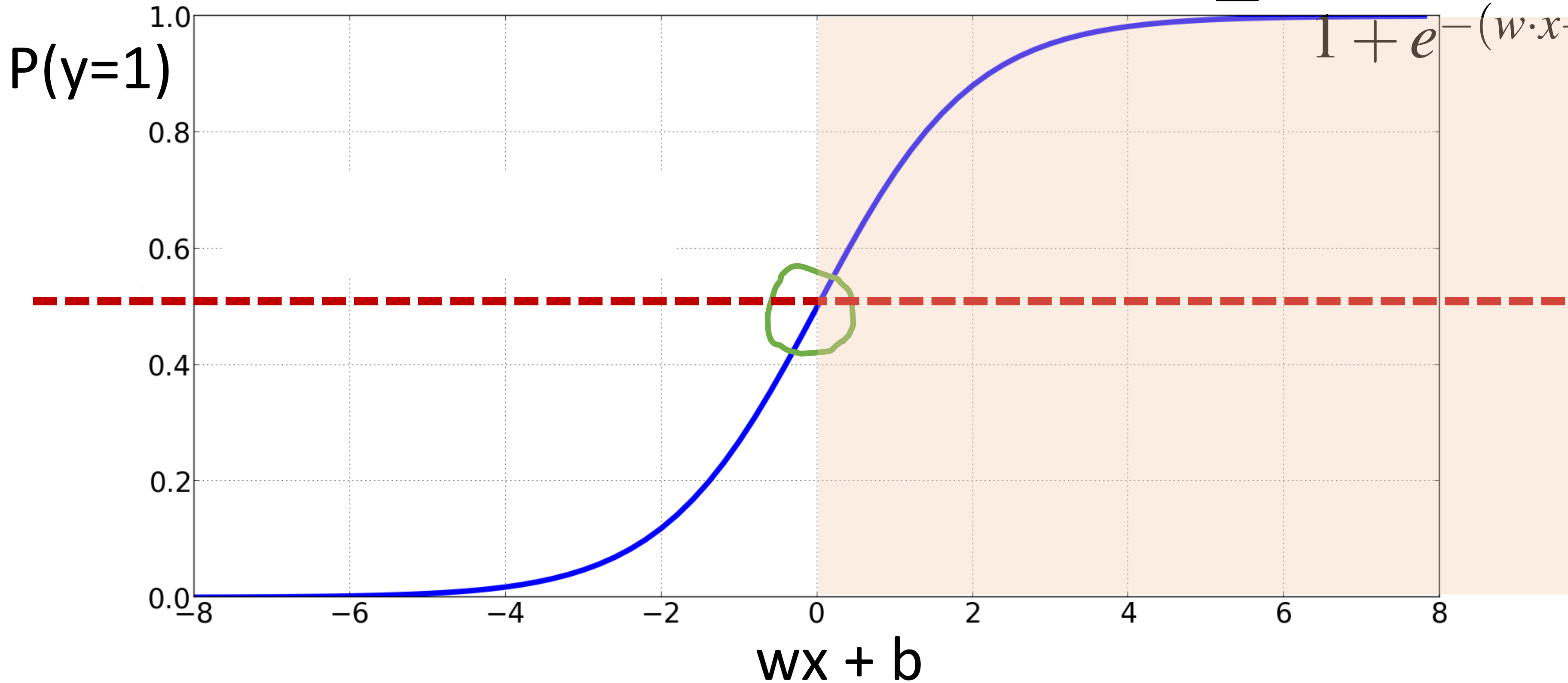
Turning a probability into a classifier

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

0.5 here is called the **decision boundary**

The probabilistic classifier

$$P(y=1) = \sigma(w \cdot x + b)$$
$$= \frac{1}{1+e^{-(w \cdot x + b)}}$$



Turning a probability into a classifier

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

if $w \cdot x + b > 0$

if $w \cdot x + b \leq 0$

Examples

Feature	Coefficient	Weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

Example 1: Empty Document

$X = \{ \}$

$$P(Y=0) = \frac{1}{1+\exp(0.1)} = 0.48$$

$$P(Y=1) = \frac{\exp(0.1)}{1+\exp(0.1)} = 0.52$$

Bias β_0 represents the class priors

Examples

Feature	Coefficient	Weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

Example 2:
$X = \{ Mother, Nigeria\}$

Examples

Feature	Coefficient	Weight
bias	β_0	0.1
“viagra”	β_1	2.0
“mother”	β_2	-1.0
“work”	β_3	-0.5
“nigeria”	β_4	3.0

Example 2:
$X = \{ Mother, Nigeria \}$

$$P(Y = 0) = \frac{1}{1 + \exp(0.1 \ - 1.0 + 3.0)} = 0.11$$

$$P(Y = 1) = \frac{\exp(0.1 \ - 1.0 + 3.0)}{1 + \exp(0.1 \ - 1.0 + 3.0)} = 0.88$$

Examples

Feature	Coefficient	Weight
bias	β_0	0.1
“viagra”	β_1	2.0
“mother”	β_2	-1.0
“work”	β_3	-0.5
“nigeria”	β_4	3.0

Example 3:
$X = \{\, Mother, Work, Nigeria, Mother\}$

Examples

Feature	Coefficient	Weight
bias	β_0	0.1
“viagra”	β_1	2.0
“mother”	β_2	-1.0
“work”	β_3	-0.5
“nigeria”	β_4	3.0

Example 3:
$X = \{ Mother, Work, Nigeria, Mother \}$

$$P(Y = 0) = \frac{1}{1 + \exp(0.1 \ - 1.0 + 2.0 + 3.0 \ - 1.0)} \quad = 0.60$$

$$P(Y = 1) = \frac{\exp(0.1 \ - 1.0 + 2.0 + 3.0 \ - 1.0)}{1 + \exp(0.1 \ - 1.0 + 2.0 + 3.0 \ - 1.0)} \quad = 0.30$$

Logistic Regression

- Given a set of weights, β, compute conditional likelihood $P(y\ |\beta, x)$
- Find the weights that maximize the conditional likelihood on training data
- **Intuition:** higher weights implies corresponding feature is strongly indicative of the class for the observation

Outline

Linear Regression

Evaluation

Logistic Regression

Learning weights

Process Learning Weights

1. Randomly initialize weights

2. Make predictions $\hat{y}$

3. Quantify how close $\hat{y}$ *and* y are

    We call this the ***distance*** aka Loss function

4. Update weights accordingly

    aka Optimization

5. Repeat 2-4

Distance between $\hat{y}$ and y

We want to know how far is the classifier output:

$\hat{y} = \sigma(w \cdot x + b)$

from the true output:

y [= either 0 or 1]

We'll call this difference:

$L(\hat{y}, y)$ = how much $\hat{y}$ differs from the true y

Intuition of negative log likelihood loss = cross-entropy loss

- A case of conditional maximum likelihood estimation
- We choose the parameters *w*,*b* that maximize
- the log probability
- of the true *y* labels in the training data
- given the observations *x*