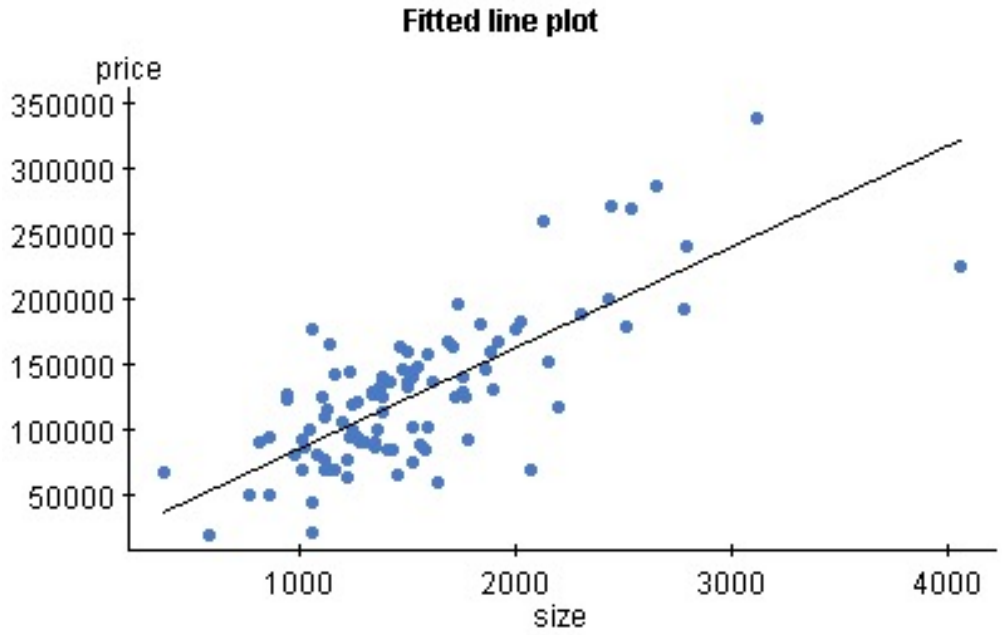
Linear & Logistic Regression

David Quigley
CSCI 5622
2021 Fall

Course Logistics

- Project: Groups due 9/9!
  - Everyone submits an identical document
  - Project 2: Pitch and Project 2.1: Pitch Feedback instructions are now viewable on Canvas
- Problem Sets: Problem Set 1 Due 9/16
  - Piazza is pretty active
    - If you have a question, it may have already been answered there!

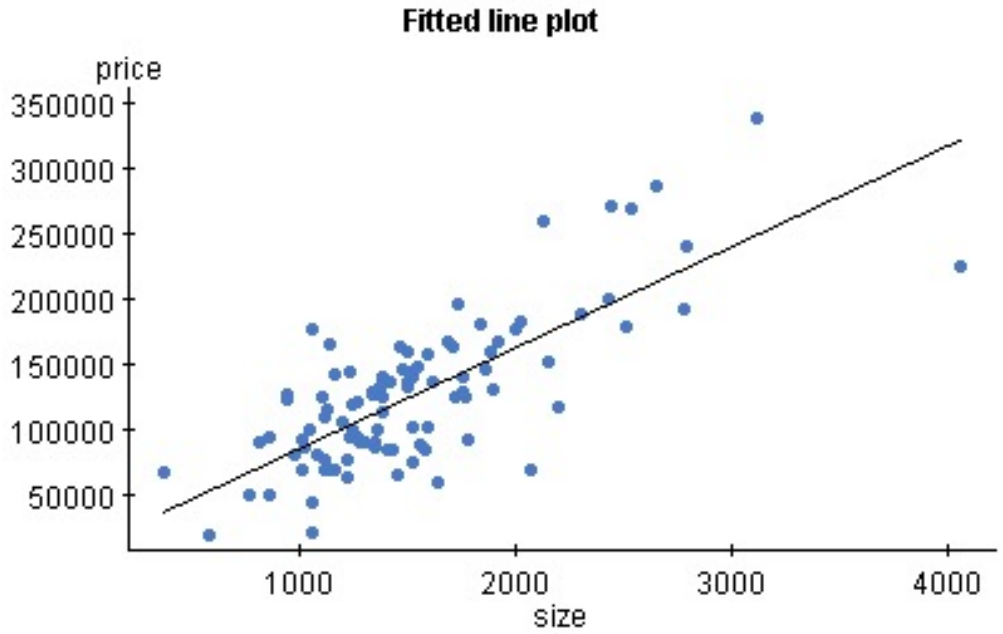
Linear Regression – The Basics



Linear Regression – The Basics

$Y \approx \beta_0 + \beta_1 X.$

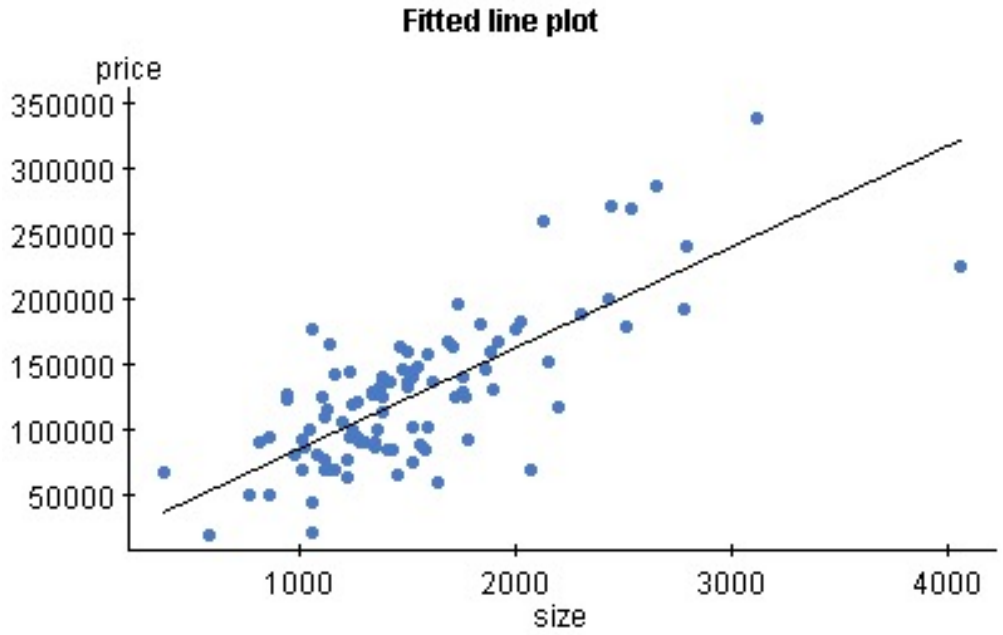
$y = w_0 + w_1 * x$



Linear Regression – The Basics

$Y \approx \beta_0 + \beta_1 X.$

$y = w_0 + w_1 * x$



How do I find (or generate) this line of best fit?

Linear Regression – Multiple Dimensions

$X = [x_1]$, we care about w_0 (bias, intercept) + w_1(weight, slope) * x_1

What about multiple dimensions?

Linear Regression – Multiple Dimensions

$X = [x_1]$, we care about w_0 (bias, intercept) + w_1(weight, slope) * x_1

What about multiple dimensions? *Each Dimension Gets a Weight!*

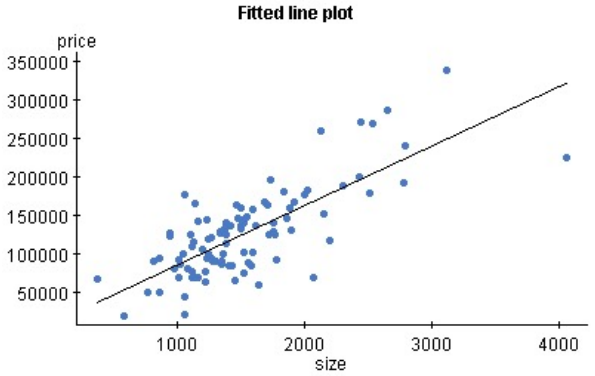
$X = [x_1, x_2, x_3, \ldots, x_n]$

$w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + \ldots + w_n * x_n = W \cdot X$

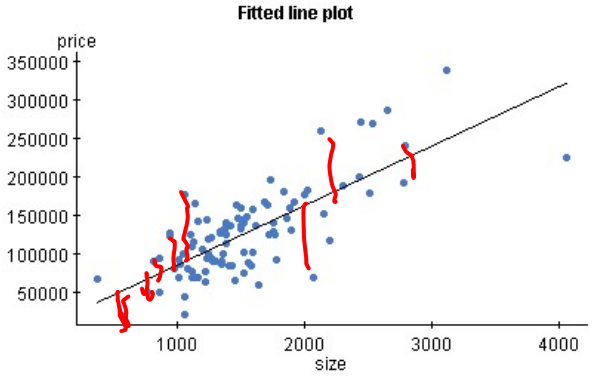
(NOTE: For dot product to work, X needs to be prepended, $X = [1, x_1, \ldots]$

Linear Regression – The Details

First, let's define "how good" a line is.
- How well does it define our points?



Linear Regression – The Details



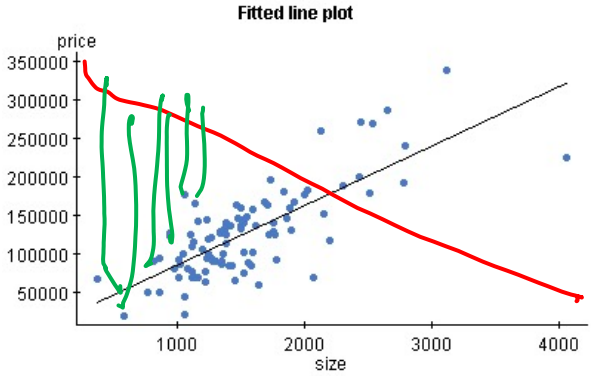
First, let's define "how good" a line is.
- How well does it define our points?

How far away, on average, is each point from the line?

The *Mean Squared Error*

1) The difference between each point and the line along the output axis
    1) (square this value to get it to always be a positive value)
2) Take the mean of this value

Linear Regression – The Details



First, let's define "how good" a line is.
- How well does it define our points?

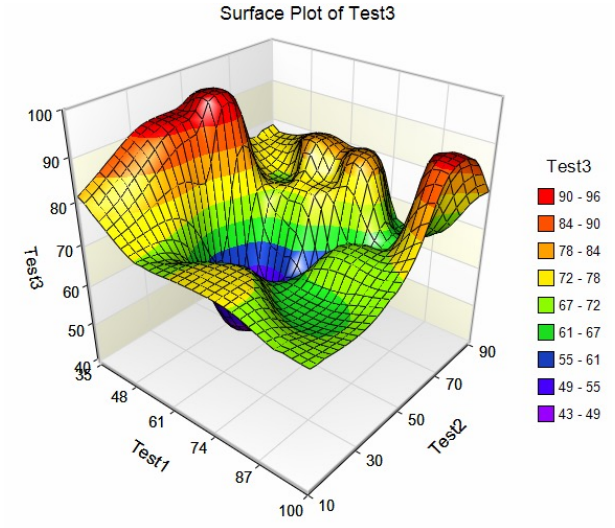
How far away, on average, is each point from the line?
The *Mean Squared Error*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

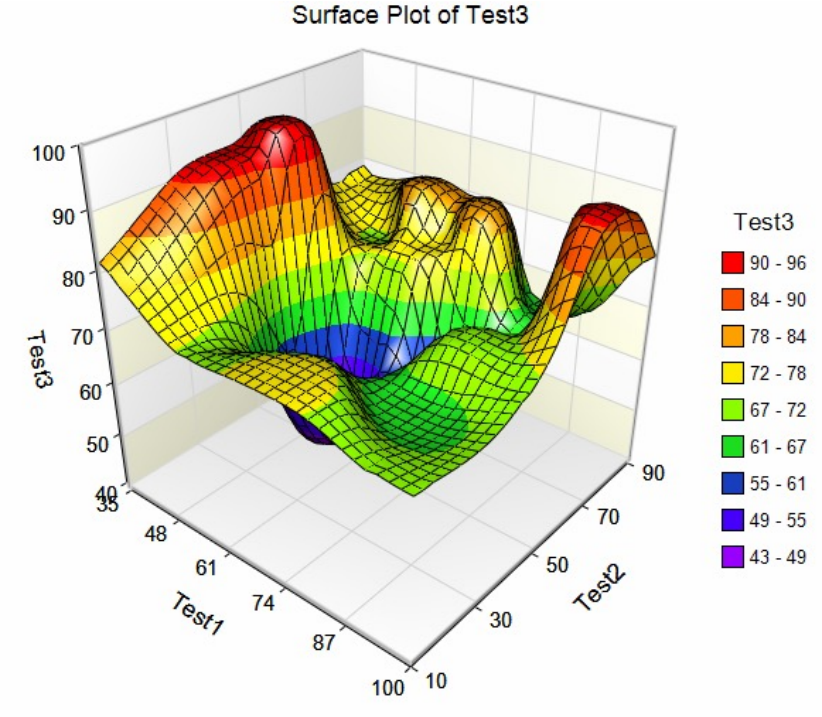
We want the minimum distance from the points to the line.

Minimizing a function...

Assume Test1 and Test2 are our w values, we want to minimize Test3. How would you do it?



Optimize by checking all points, choosing min

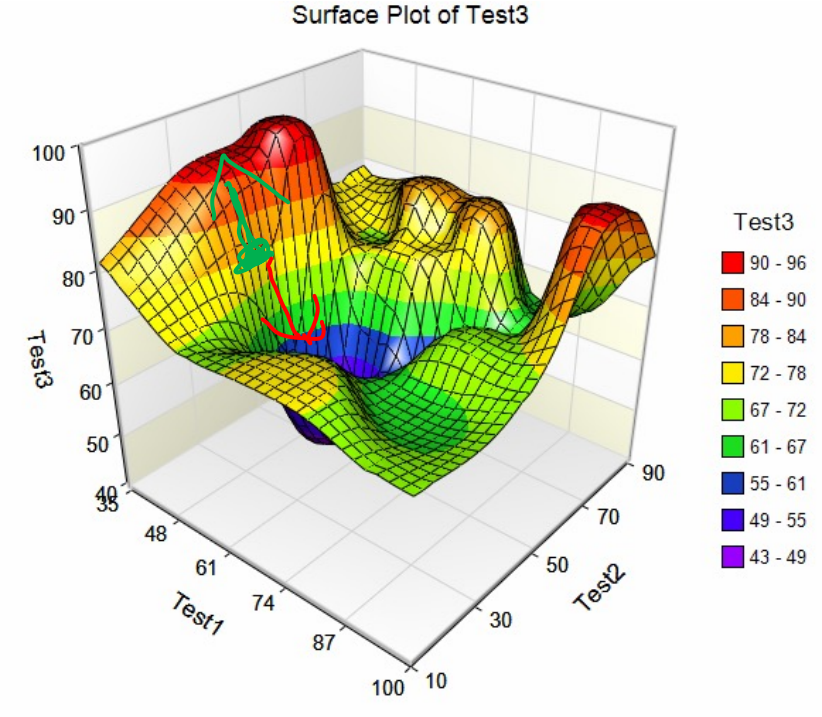


Optimizing using the gradient

For any function f, there exists a gradient.

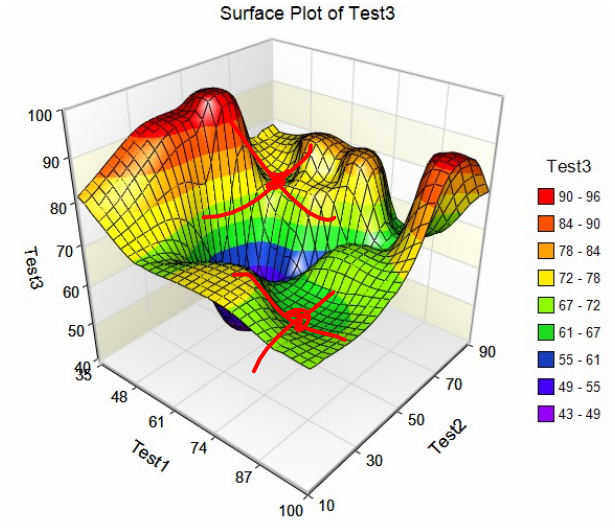
The gradient is the vector that points in the direction in which f is increasing the fastest.

Optimize using the gradient



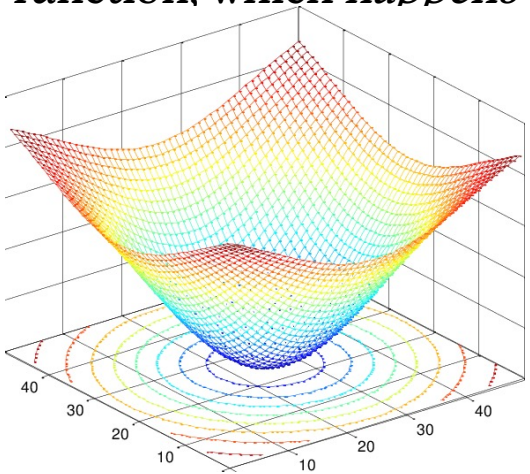
Optimizing argmin(MSE(w))

Gradient descent would not work for any given function...



Optimizing argmin(MSE(w))

But our function is a *linear* function, which *happens* to be concave!



i.e. If we're at a *local* minimum, we're at the *global* minimum

https://homes.cs.washington.edu/~marcotcr/blog/concavity/

Optimizing using Gradient Descent

We can think of this in terms of *updating our weights*:

$w \leftarrow w - \nabla_w \text{MSE}(w)$

$\nabla_w \text{MSE}(w) = [\partial \text{MSE}(w) / \partial w_0, \partial \text{MSE}(w) / \partial w_1, \dots, \partial \text{MSE}(w) / \partial w_n]$

The weights are updated by removing the gradient
Each weight is updated by subtracting its partial derivative

Optimizing using Gradient Descent

We can think of this in terms of *a rate of change*:

$$w \leftarrow w - \lambda * \nabla_w \text{MSE}(w)$$

Optimizing using Gradient Descent

We can think of this in terms of *each feature weight's update*:

$w_k \leftarrow w_k - \lambda * \partial MSE(w) / \partial w_k$ for each $k = 0 \rightarrow D$ (dimensions)

Optimizing using Gradient Descent

We can think of this in terms of *each feature weight's update*:

$w_k \leftarrow w_k - \lambda * \partial MSE(w) / \partial w_k$ for each $k = 0 \rightarrow D$ (dimensions)

We can think of the partial derivative* of w_k in terms of the k^{th} feature of the i^{th} training example

*The proof of this fact can be found in readings... (i.e. it's really hard math!)

Optimizing using gradient descent

```
w = //initial weights
x = //training examples (2d matrix)
y = //training answers (1d vector)
Lamda = //rate
While not converged():
        For k in range(0,D): //for each dimension
                update[k] = 0
                for i in range(0,n): //for each training example
                        update[k] = (sigm(dot(w, x[i,:])) – y[i]) *
                                x[i,k]
w = w – lamda * update
```

Gradient Descent – Real World Examples

Real world example: predicting the weather

Gradient Descent – Predicting the weather

We have relevant features (precipitation, humidity, wind, latitude, longitude, time of day, day of year...

- say 99 features, across 19,354 “incorporated places” in us
- current temperature regression

We have sampled data points

- 1 sample per site per minute for 30 years

Gradient Descent – Predicting the weather

We have relevant features (precipitation, humidity, wind, latitude, longitude, time of day, day of year…

- say 99 features, across 19,354 “incorporated places” in us
- plus current temperature

We have sampled data points

- 1 sample per site per minute for 30 years

3.05e+13 features

There are bigger problems out there!

We’ll solve that problem in a moment…

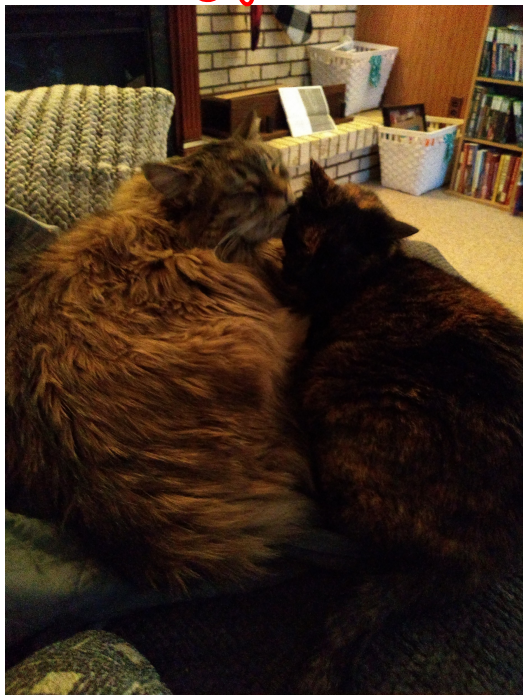
Logistic Regression

Classification – A New Problem Space

?

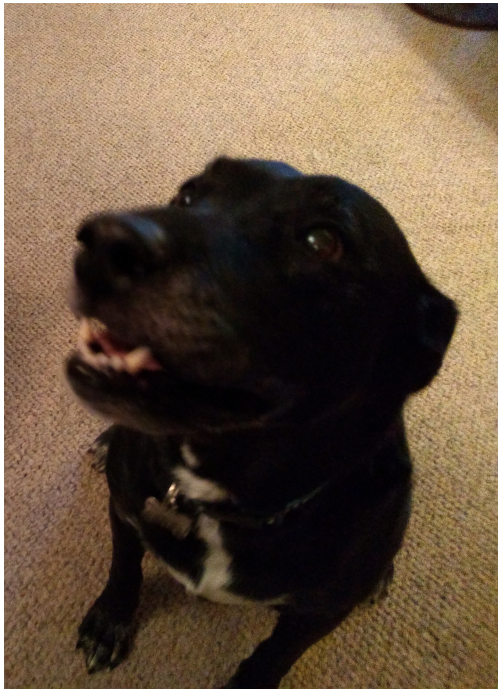
Classification – A New Problem Space

Barney Grotes?

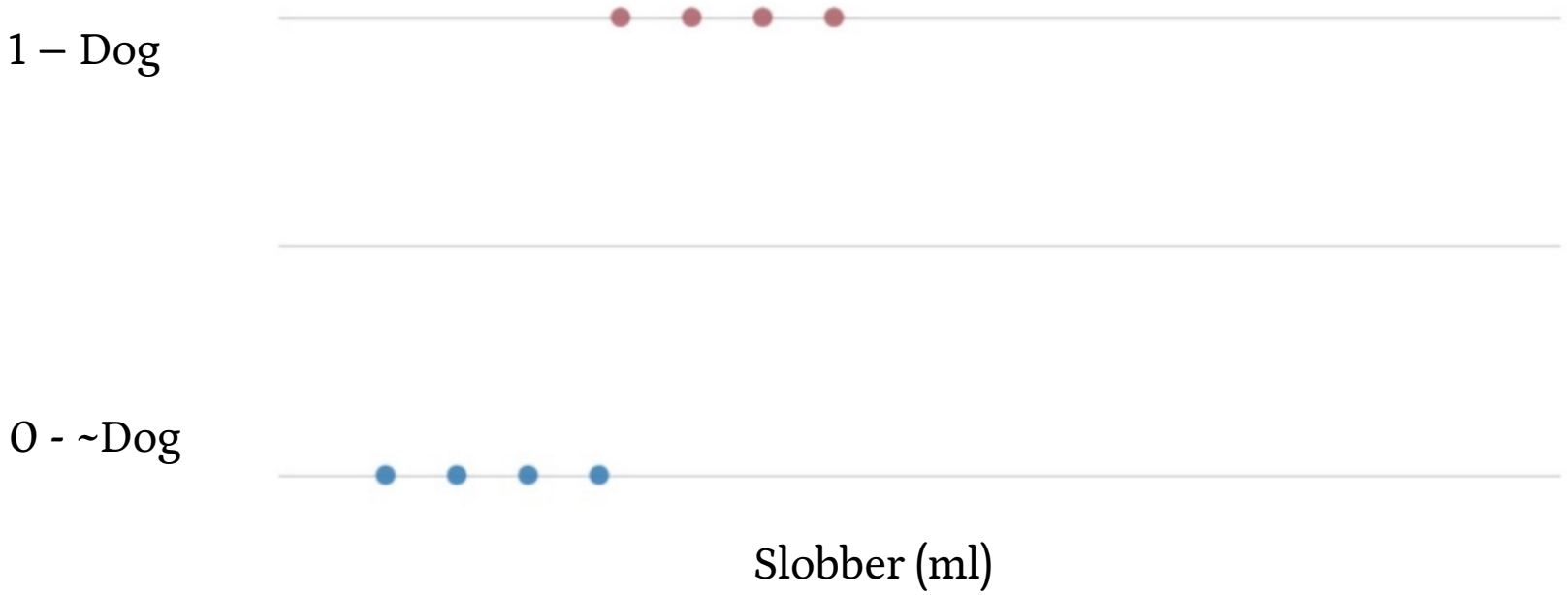


?

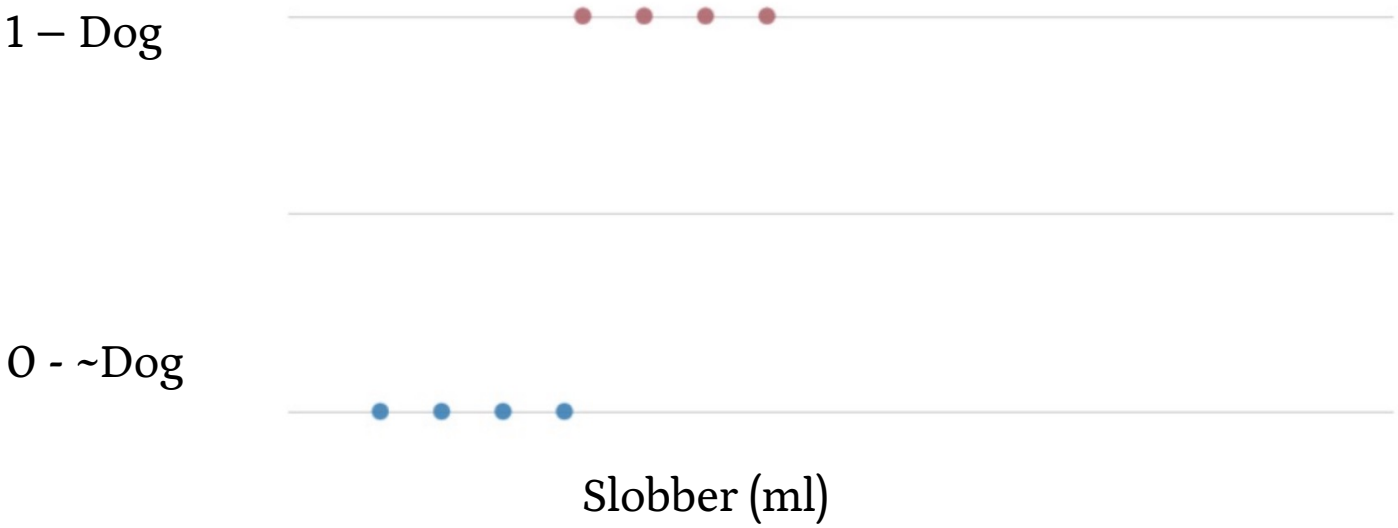
Lulu?



Our Problem Space – Dog Slobber

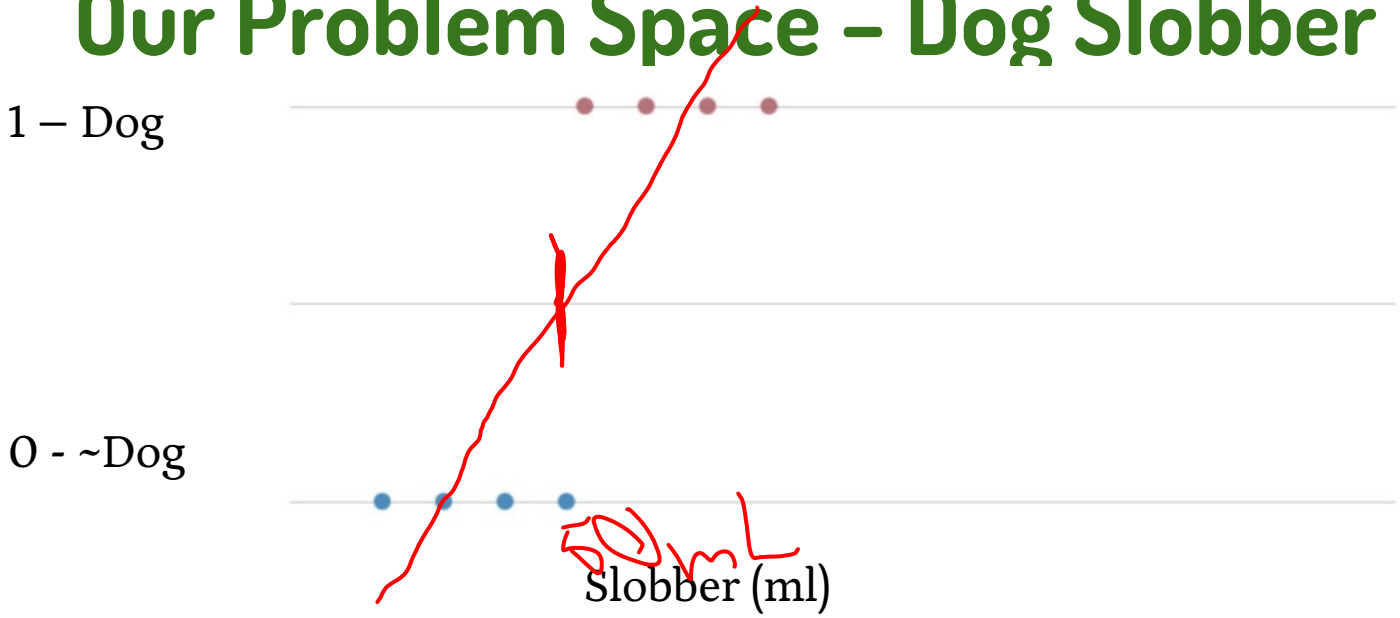


Our Problem Space – Dog Slobber



Modeling p(y | x, D) given one feature (slobber)?

Our Problem Space – Dog Slobber



Linear Regression

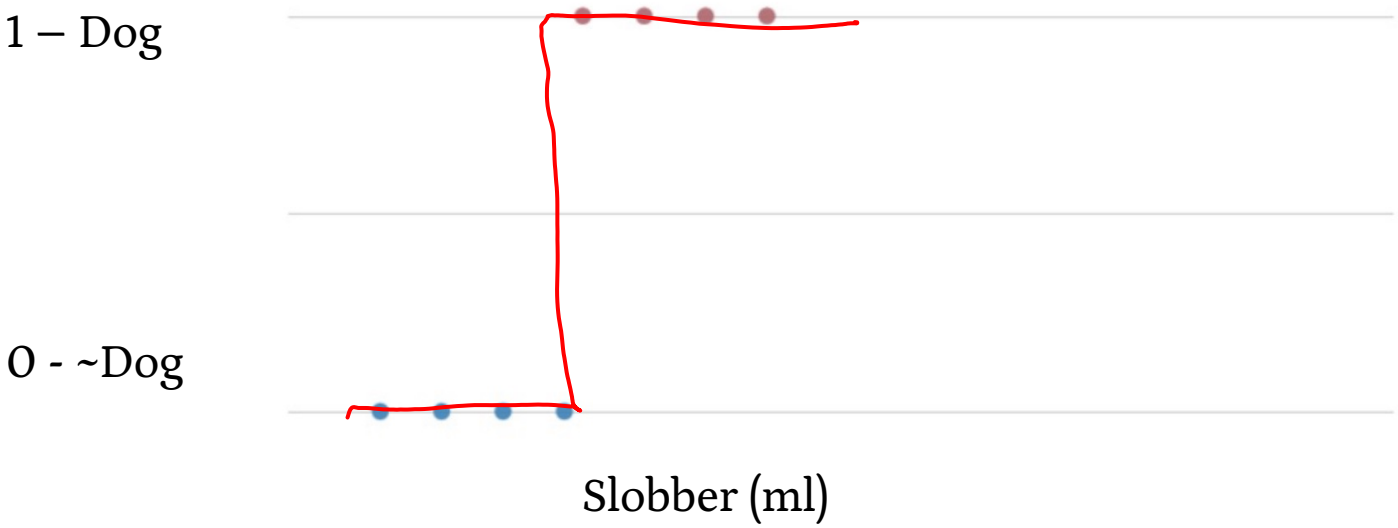
$p(y \mid x, D) = \beta_0 + \beta_1 X.$

Our Problem Space – Dog Slobber

Linear Regression

$p(y \mid x, D) = \beta_0 + \beta_1 X.$

Our Problem Space – Dog Slobber

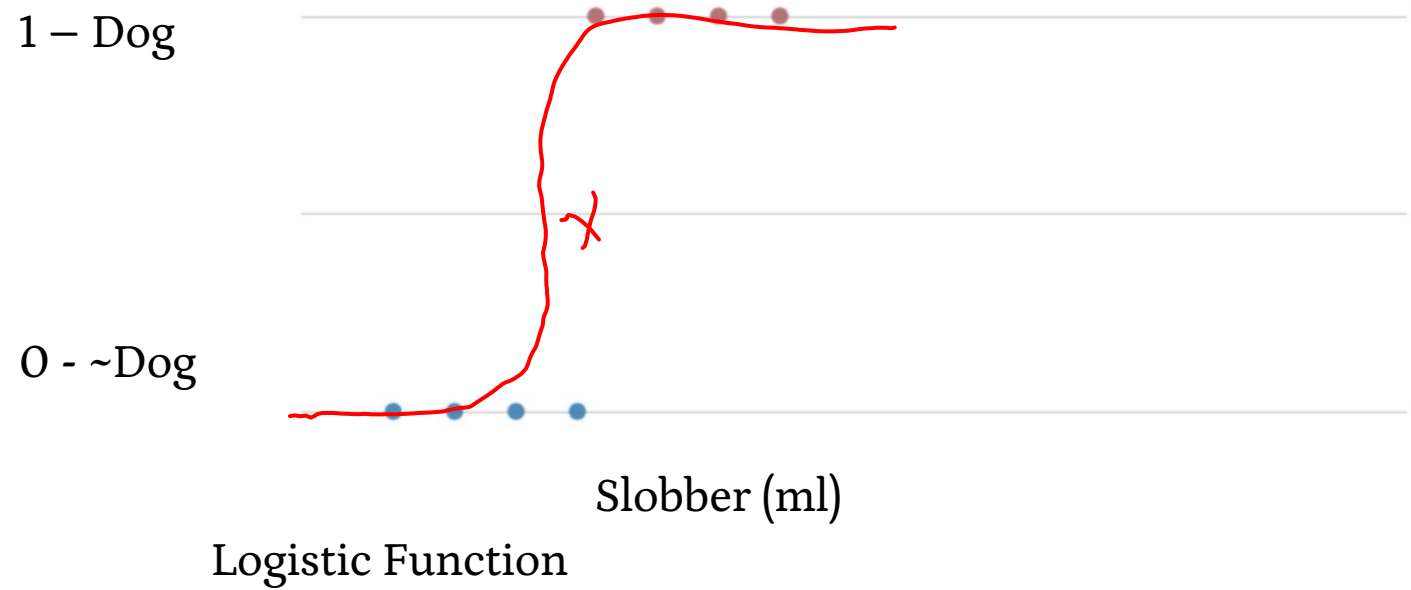


Perceptron

if x > 50 ml, y = 1

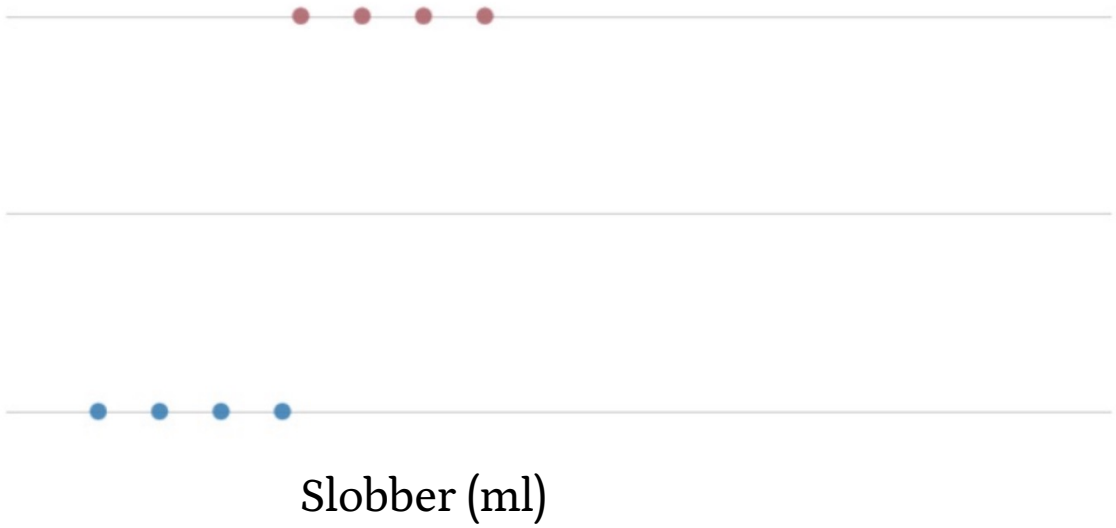
else, y = 0

Our Problem Space – Dog Slobber



Logistic Function

Our Problem Space – Dog Slobber

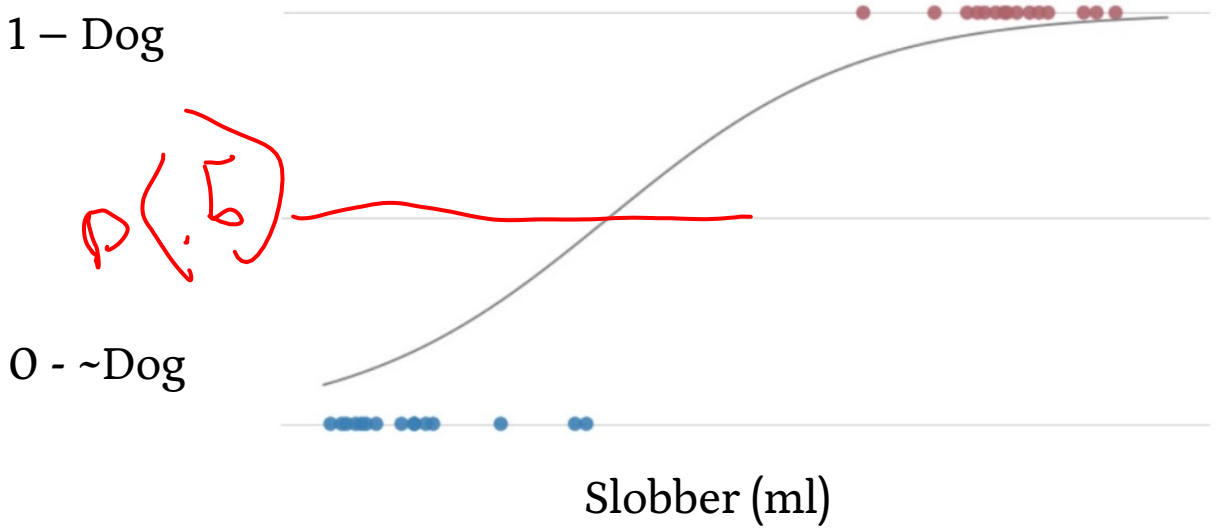


Logistic Function $p(X) = \frac{e^{w0+w1*X}}{1 + e^{w0+w1*X}}$

Logistic Function

$$\text{Logistic Function } p(X) = \frac{e^{w0+w1*X}}{1 + e^{w0+w1*X}} = \frac{1}{1 + e^{-(w0+w1*X)}}$$

Our Problem Space – Dog Slobber



if p(X) >= .5, y = 1

else, y = 0

Logistic Regression

- Simple *discriminative* model
- Easy to train
- Does not make data assumptions
- Works well on *medium* data sets

Logistic Function

- Probabilistic (values 0 – 1)
- Very Smooth

My New Problem

- Manager for Roxxcart
- Setting up a Winter Holiday Playlist
- You need to *only play "Holiday Songs"* for the next 3 months.

Y = "Holiday Song" or "Not Holiday Song"

What features exist in a song?

Timbre Title

Lyrics Release Instruments

Freq BPM

Song Features

- A lot of basic features
    - Key signature, tempo, etc.
- A lot of *Band / Song based features*
    - Which band is playing / which musicians are performing?
    - Which instruments are used / featured?
    - These could be helpful, but they get hard to work with, we'll look at adding them later.
- A lot of features *based on moments / occurrence through the song*
    - Rhythmic patterns, note sequences, etc.
    - These temporal features are very tricky to implement, we'll discuss them at a later date. For now we'll discard them

Song Features (Continued)

- Lyrics
  - One of the primary components of many songs is the lyrics!
    - They’re especially commonly useful in determining if a song is considered “Holiday” themed

Everything You Need to Know About Natural Language Processing in 5 Minutes*

*This is impossible. If you actually care about doing NLP in the real world, consider taking the NLP Class, CSCI 5832 (cross listed in LING).

What's in Lyrics?

Iridescent All Over by Let's Make a Music

You know it's hard to look at you
The way you're shining
Cause your hair and your smile and your eyes
Are sparkling, babe

Going blind every time
You look my way
See you like starlings do
Iridescent all over babe

What's in Lyrics?

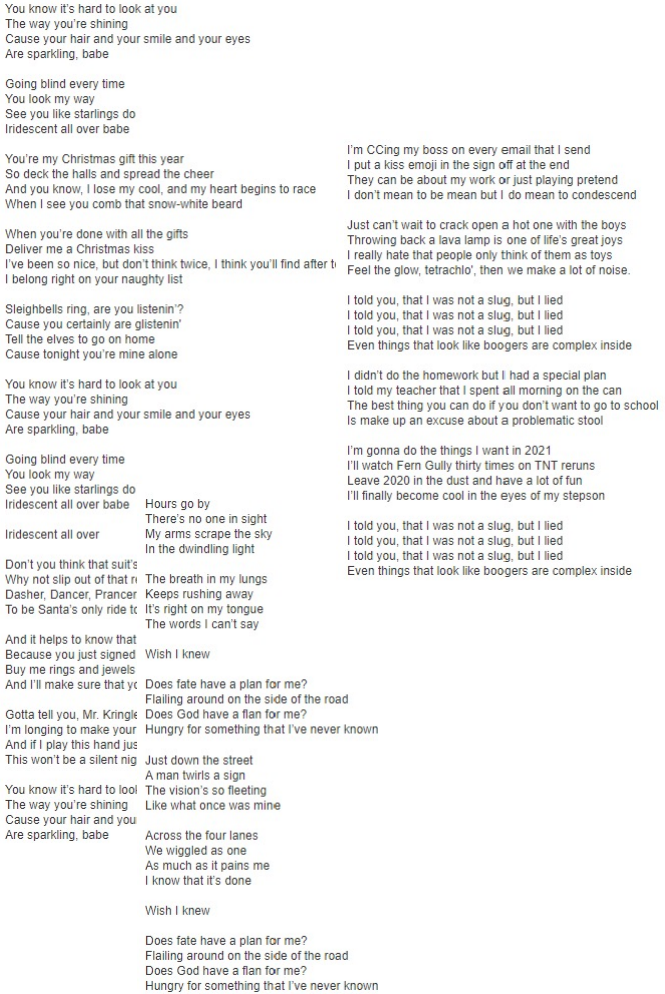
Iridescent All Over by Let's Make a Music

You're my Christmas gift this year
So deck the halls and spread the cheer
And you know, I lose my cool, and my heart begins to race
When I see you comb that snow-white beard

Song as Document

Each song is a sample

Draw features from each



You know it's hard to look at you
The way you're shining
Cause your hair and your smile and your eyes
Are sparkling, babe

Going blind every time
You look my way
See you like starlings do
Iridescent all over babe

You're my Christmas gift this year
So deck the halls and spread the cheer
And you know, I lose my cool, and my heart begins to race
When I see you comb that snow-white beard

When you're done with all the gifts
Deliver me a Christmas kiss
I've been so nice, but don't think twice, I think you'll find after t
I belong right on your naughty list

Sleighbells ring, are you listenin'?
Cause you certainly are glistenin'
Tell the elves to go on home
Cause tonight you're mine alone

You know it's hard to look at you
The way you're shining
Cause your hair and your smile and your eyes
Are sparkling, babe

Going blind every time
You look my way
See you like starlings do
Iridescent all over babe

Iridescent all over

Don't you think that suit's
Why not slip out of that r
Dasher, Dancer, Prancer
To be Santa's only ride t

And it helps to know that
Because you just signed
Buy me rings and jewels
And I'll make sure that y

Gotta tell you, Mr. Kringl
I'm longing to make your
And if I play this hand jus
This won't be a silent nig

You know it's hard to lool
The way you're shining
Cause your hair and you
Are sparkling, babe

Hours go by
There's no one in sight
My arms scrape the sky
In the dwindling light

The breath in my lungs
Keeps rushing away
It's right on my tongue
The words I can't say

Wish I knew

Does fate have a plan for me?
Flailing around on the side of the road
Does God have a flan for me?
Hungry for something that I've never known

Just down the street
A man twirls a sign
The vision's so fleeting
Like what once was mine

Across the four lanes
We wiggled as one
As much as it pains me
I know that it's done

Wish I knew

Does fate have a plan for me?
Flailing around on the side of the road
Does God have a flan for me?
Hungry for something that I've never known

I'm CCing my boss on every email that I send
I put a kiss emoji in the sign off at the end
They can be about my work or just playing pretend
I don't mean to be mean but I do mean to condescend

Just can't wait to crack open a hot one with the boys
Throwing back a lava lamp is one of life's great joys
I really hate that people only think of them as toys
Feel the glow, tetrachlo', then we make a lot of noise.

I told you, that I was not a slug, but I lied
I told you, that I was not a slug, but I lied
I told you, that I was not a slug, but I lied
Even things that look like boogers are complex inside

I didn't do the homework but I had a special plan
I told my teacher that I spent all morning on the can
The best thing you can do if you don't want to go to school
Is make up an excuse about a problematic stool

I'm gonna do the things I want in 2021
I'll watch Fern Gully thirty times on TNT reruns
Leave 2020 in the dust and have a lot of fun
I'll finally become cool in the eyes of my stepson

I told you, that I was not a slug, but I lied
I told you, that I was not a slug, but I lied
I told you, that I was not a slug, but I lied
Even things that look like boogers are complex inside

Words as Features

Look for the presence or absence of any given word, and count that as that feature

It's beginning to look a lot like Christmas...

Words as Features

Look for the presence or absence of any given word, and count that as that feature

It's beginning to look a lot like Christmas...

Feature	X_1 = "Santa"	X_2 = "Dreidel"	X_3 = "Christmas"	X_3 = "Bad"	X_4 = "Hate"
Is In Song	0	0	0	0	0

Choosing our Feature Words

How will we pick our words?

Choosing our Feature Words

How will we pick our words?

Trust the Experts!
Looking at datasets like AFINN, or SentiWordNet, can give us an idea of words we care about if we're looking for sentiment.

https://github.com/fnielsen/afinn/tree/master/afinn/data
https://github.com/aesuli/sentiwordnet

Choosing our Feature Words

How will we pick our words?

Use all the words in your dictionary?
Use all the words in your dataset?

https://imgflip.com/i/3nld61

Logistic Regression – Multiple Dimensions

$X = [x_1]$, we care about w_0 (bias, intercept) + w_1(weight, slope) * x_1

What about multiple dimensions? *Each Dimension Gets a Weight!*

$X = [x_1, x_2, x_3, \ldots, x_n]$

$w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + \ldots + w_n * x_n = W \cdot X$

(NOTE: For dot product to work, X needs to be prepended, $X = [1, x_1, \ldots]$)

How do we choose a weight? *Trust the experts? Good guesswork? At random? We'll explore this later!*

Logistic Regression – Rihanna Example

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

Logistic Regression – Rihanna Example

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

What classification do we give x=[“Santa”, “Bad”]

$x = [1, 1, 0, 0, 1, 0]$

$w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + w_5 * x_5$

Logistic Regression – Rihanna Example

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

What classification do we give x=[“Santa”, “Bad”]

$x = [1, 1, 0, 0, 1, 0]$

$w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + w_5 * x_5$

$-0.1 + 10.0 * 1 + 15.0 * 0 + 12.0 * 0 + -2.0 * 1 + -4.0 * 0 = +12.9$

Logistic Regression – Rihanna Example

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

What classification do we give x=[“Santa”, “Bad”]

$x = [1, 1, 0, 0, 1, 0]$

$w_0 + w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + w_5 * x_5$

$-0.1 + 10.0 * 1 + 15.0 * 0 + 12.0 * 0 + -2.0 * 1 + -4.0 * 0 = +12.9$

$$\frac{1}{1 + e^{-(12.9)}} = 0.999... > 0.5 = \text{Holiday Song!}$$

Logistic Regression – Working with Weights

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

What sort of relationship do we see from weights and classification?

Logistic Regression – Weights Fall short

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

Take the song line “Oh Dreidel, Dreidel, Dreidel, I made it out of clay…”

X = [“Dreidel”, “Dreidel”, “Dreidel”]

Logistic Regression – Weights Fall short

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

Take the song line “Oh Dreidel, Dreidel, Dreidel, I made it out of clay…”

X = [“Dreidel”, “Dreidel”, “Dreidel”]

The way we’ve been doing them, that’s x = [1,0,1,0,0,0]

Does this accurately reflect the amount of Holiday representation in our song lyric?

Logistic Regression – Bag of Words approach

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

Take the song line “Oh Dreidel, Dreidel, Dreidel, I made it out of clay…”
X = [“Dreidel”, “Dreidel”, “Dreidel”]
The way we’ve been doing them, that’s x = [1,0,1,0,0,0]
Does this accurately reflect the amount of Holiday representation in our song lyric?
Can you think of another way to use our accounting of features?

What if we count the number of instances of each word x = [1,0,3,0,0,0]

Logistic Regression – Binary vs. Bag of Words

Feature	*Bias*	X_1 = “Santa”	X_2 = “Dreidel”	X_3 = “Christmas”	X_3 = “Bad”	X_4 = “Hate”
Weight	-0.1	10.0	15.0	12.0	-2.0	-4.0

Take the song “This is a song”

X = [1, 0, 0, 0, 0, 0]

Logistic Regression – What are the Odds?

Probability	vs. Odds	
p(y)	p(y) / (1 - p(y))	= p(y) / p(~y)
p(y) = .1	.1 / .9	= 1/9
p(y) = .75	.75 / .25	= 3

(if the odds are less than 1, we often flip it and think in terms of "against" the positive)

p(y) = .1	.1 / .9	= 9 to 1 against
p(y) = .75	.75 / .25	= 3 to 1 for

Logistic Regression – Odds

$$p(x) = \frac{1}{1 + e^{-(\text{score})}}$$

$$\text{Odds} = \frac{\frac{1}{1 + e^{-(\text{score})}}}{1 - \frac{1}{1 + e^{-(\text{score})}}}$$

$$\text{Odds} = e^{\text{score}}$$

Logistic Regression – log-odds *or* logit

$\text{Odds} = e^{\text{score}}$

$\ln(\text{odds}) = \text{score}$