California State University, Dominguez Hills
Department of Computer Science
CSC 595

Professor: Dr. Benyamin Ahmadnia
bahmadniayebosari@csudh.edu

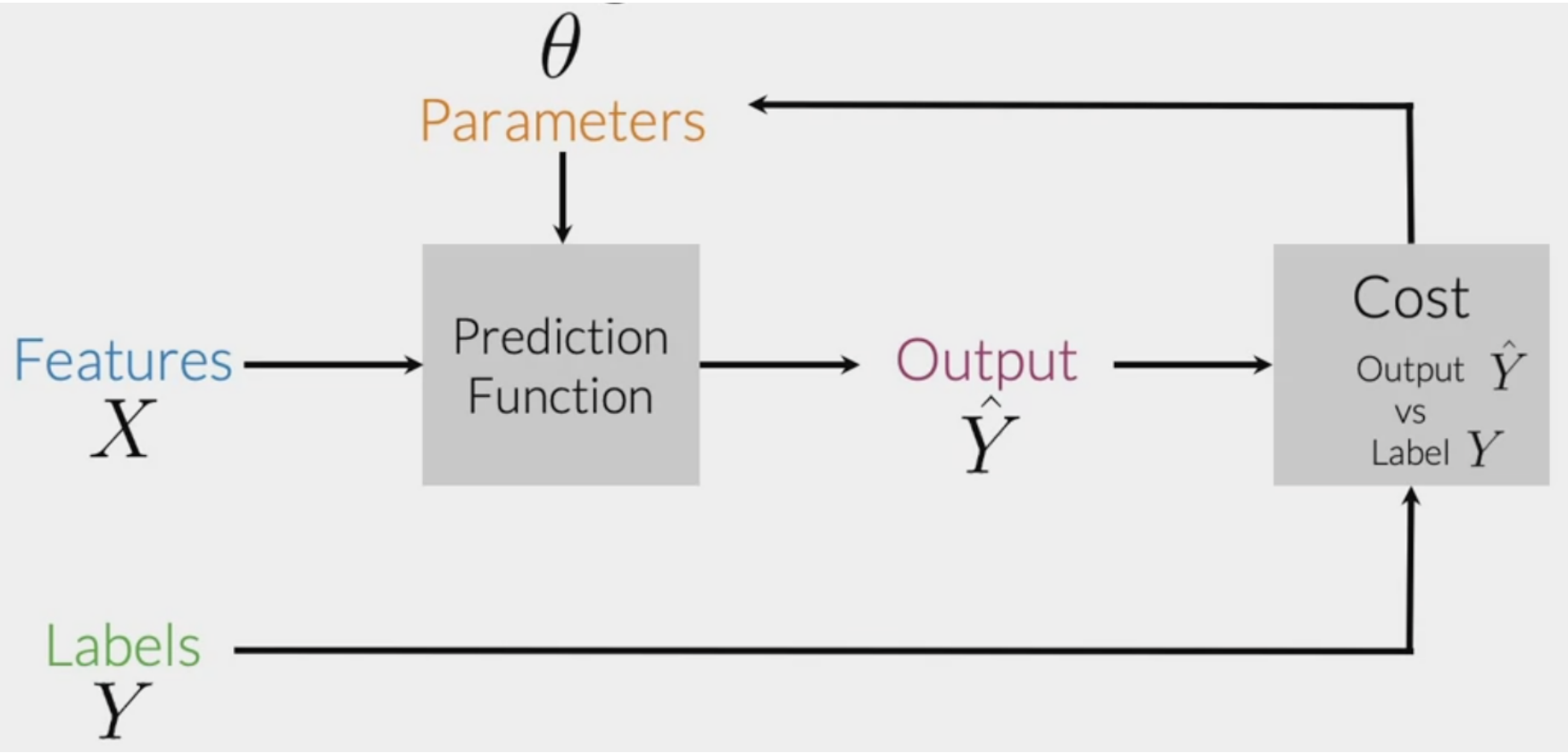
Spring 2025

Copyright Notice

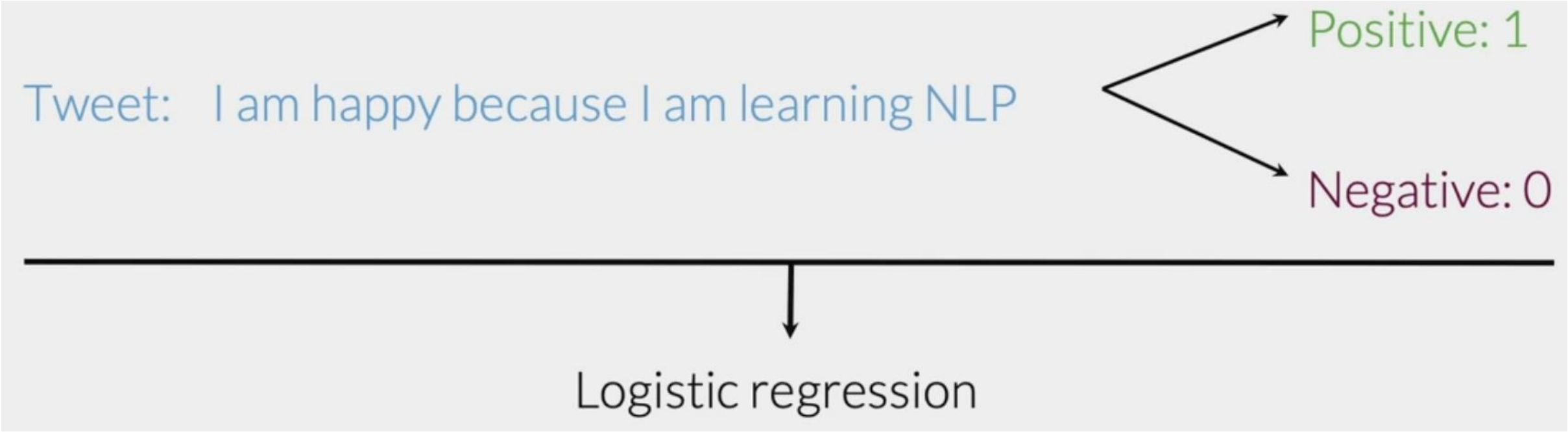
Table of Contents

Sentiment Analysis with Logistic Regression

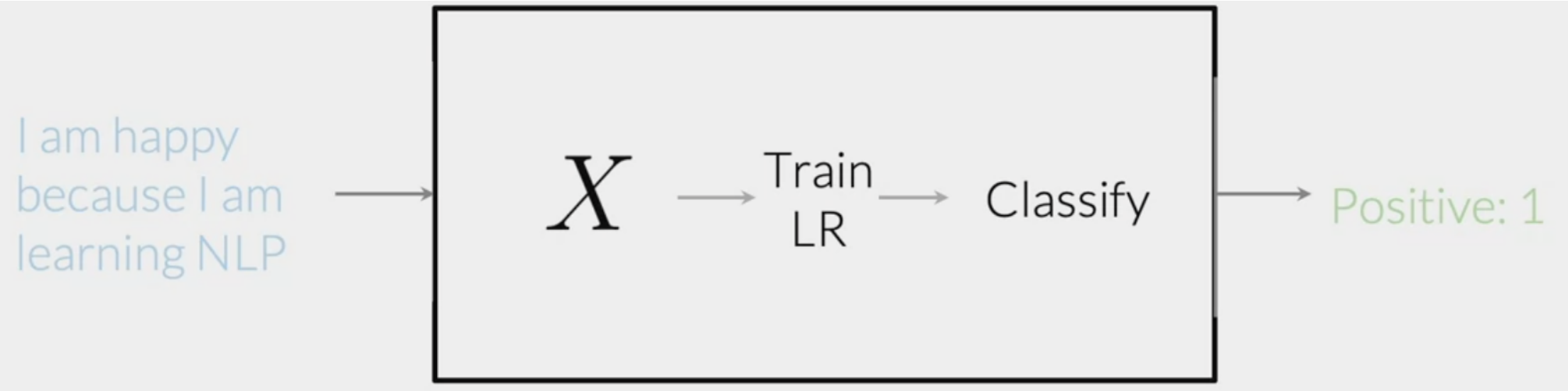
Supervised ML (training)



Sentiment Analysis



Sentiment Analysis



To Classify a Tweet …

- Positive or negative
- Extract the features
- Train the model
- Classify the tweet based on the trained model

Vocabulary

Tweets:
[tweet_1, tweet_2, ..., tweet_m]

I am happy because I am learning NLP
...
...
...
I hated the movie

$V =$

[I, am, happy, because, learning, NLP, ... hated, the, movie]

Feature Extraction

I am happy because I am learning NLP

I	am	happy	because	learning	NLP	...	hated	the	movie
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
1	1	1	1	1	1	...	0	0	0

A lot of zeros! That's a sparse representation.

Problems with Sparse Representations

I am happy because I am learning NLP

All zeros!

[1, 1, 1, 1, 1, 1, ..., 0, ..., 0, 0, 0]

1 ⟷ $|V|$

$$[\theta_0, \theta_1, \theta_2, \ldots, \theta_n]$$

$$n = |V|$$

1. Large training time
2. Large prediction time

Positive and Negative Counts

Corpus

I am happy because I am learning NLP

I am happy

I am sad, I am not learning NLP

I am sad

Vocabulary
I
am
happy
because
learning
NLP
sad
not

Positive and Negative Counts

Positive tweets	Negative tweets
I am happy because I am learning NLP	I am sad, I am not learning NLP
I am happy	I am sad

Positive Counts

Positive tweets

I am happy because I am learning NLP

I am happy

Vocabulary	PosFreq (1)
I	
am	
happy	2
because	
learning	
NLP	
sad	
not	

Negative Counts

Vocabulary	NegFreq (0)
I	
am	3
happy	
because	
learning	
NLP	
sad	
not	

Negative tweets

I am sad, I am not learning NLP

I am sad

Word Frequency in Classes

Vocabulary	PosFreq (1)	NegFreq (0)
I	3	3
am	3	3
happy	2	0
because	1	0
learning	1	1
NLP	1	1
sad	0	2
not	0	1

Word Frequency in Classes

Vocabulary	PosFreq (1)	NegFreq (0)
I	3	3
am	3	3
happy	2	0
because	1	0
learning	1	1
NLP	1	1
sad	0	2
not	0	1

freqs: dictionary mapping from (word, class) to frequency

Feature Extraction

freqs: dictionary mapping from (word, class) to frequency

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

Features of tweet m

Bias

Sum Pos. Frequencies

Sum Neg. Frequencies

Feature Extraction

Vocabulary	PosFreq (1)
I	3
am	3
happy	2
because	1
learning	1
NLP	1
sad	0
not	0

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

$$\boxed{8}$$

Feature Extraction

Vocabulary	NegFreq (0)
I	3
am	3
happy	0
because	0
learning	1
NLP	1
sad	2
not	1

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

$$\boxed{11}$$

Feature Extraction

I am sad, I am not learning NLP

$$X_m = [1, \sum_w freqs(w, 1), \sum_w freqs(w, 0)]$$

$$\downarrow$$

$$X_m = [1, 8, 11]$$

Preprocessing

@YMourri and @AndrewYNg are tuning a GREAT AI model at https://deeplearning.ai!!!

Stop words	Punctuation
and	,
is	.
are	:
at	!
has	“
for	‘
a	

Preprocessing

@YMourri ~~and~~ @AndrewYNg ~~are~~ tuning ~~a~~ GREAT AI model ~~at~~ https://deeplearning.ai!!!

@YMourri @AndrewYNg tuning GREAT AI model https://deeplearning.ai!!!

Stop words	Punctuation
and	,
is	.
are	:
at	!
has	“
for	‘
a	

Preprocessing

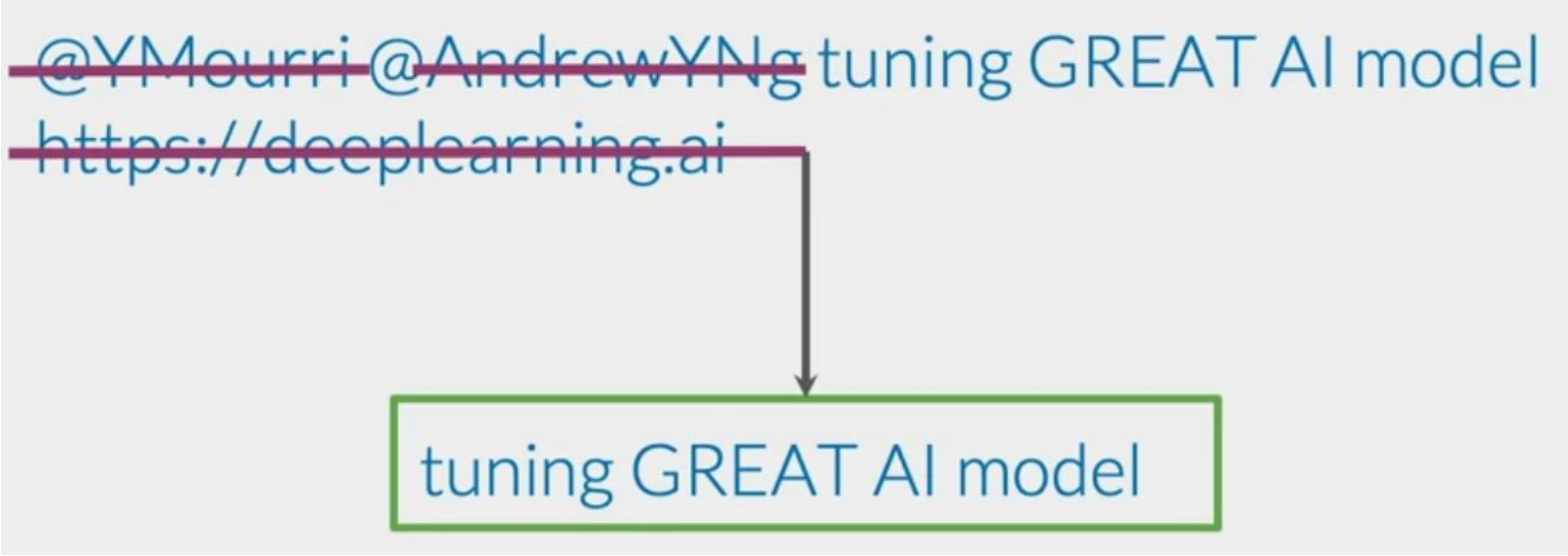
@YMourri @AndrewYNg tuning
GREAT AI model
https://deeplearning.ai!!!

@YMourri @AndrewYNg tuning
GREAT AI model
https://deeplearning.ai

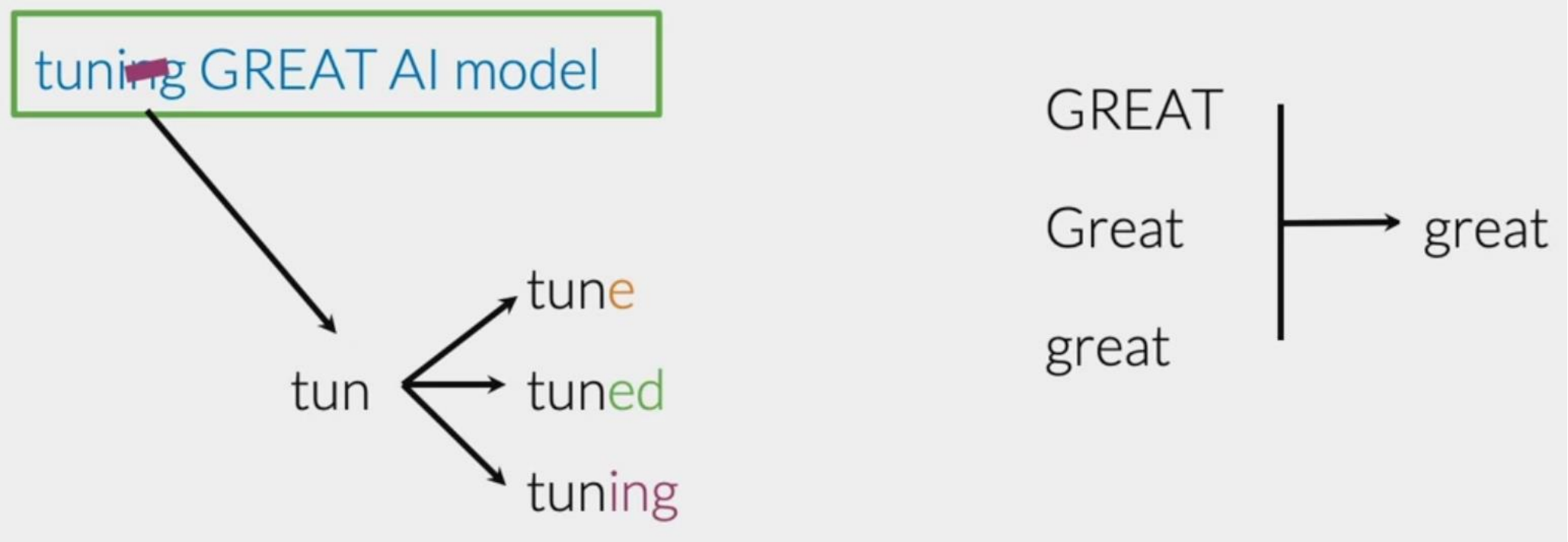
Stop words
and
is
a
at
has
for
of

Punctuation
,
.
:
!
“
‘

Preprocessing

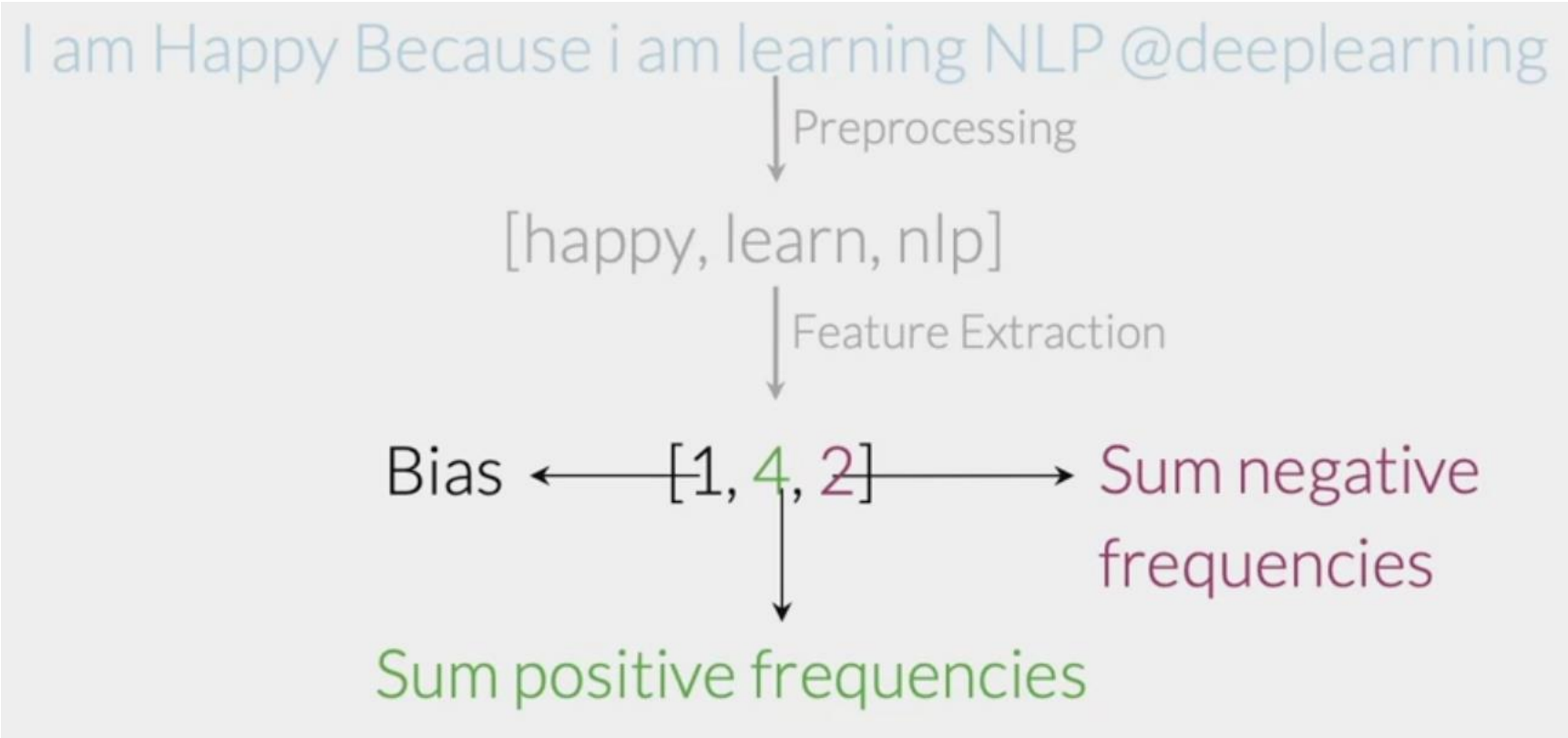


Preprocessing



Preprocessing

Putting it all Together



Putting it all Together

I am Happy Because i am learning NLP @deeplearning

I am sad not learning NLP

...

I am sad :(

→

[happy, learn, nlp]

[sad, not, learn, nlp]

...

[sad]

→

[[1, 40, 20],
[1, 20, 50],
...
[1, 5, 35]]

Putting it all Together

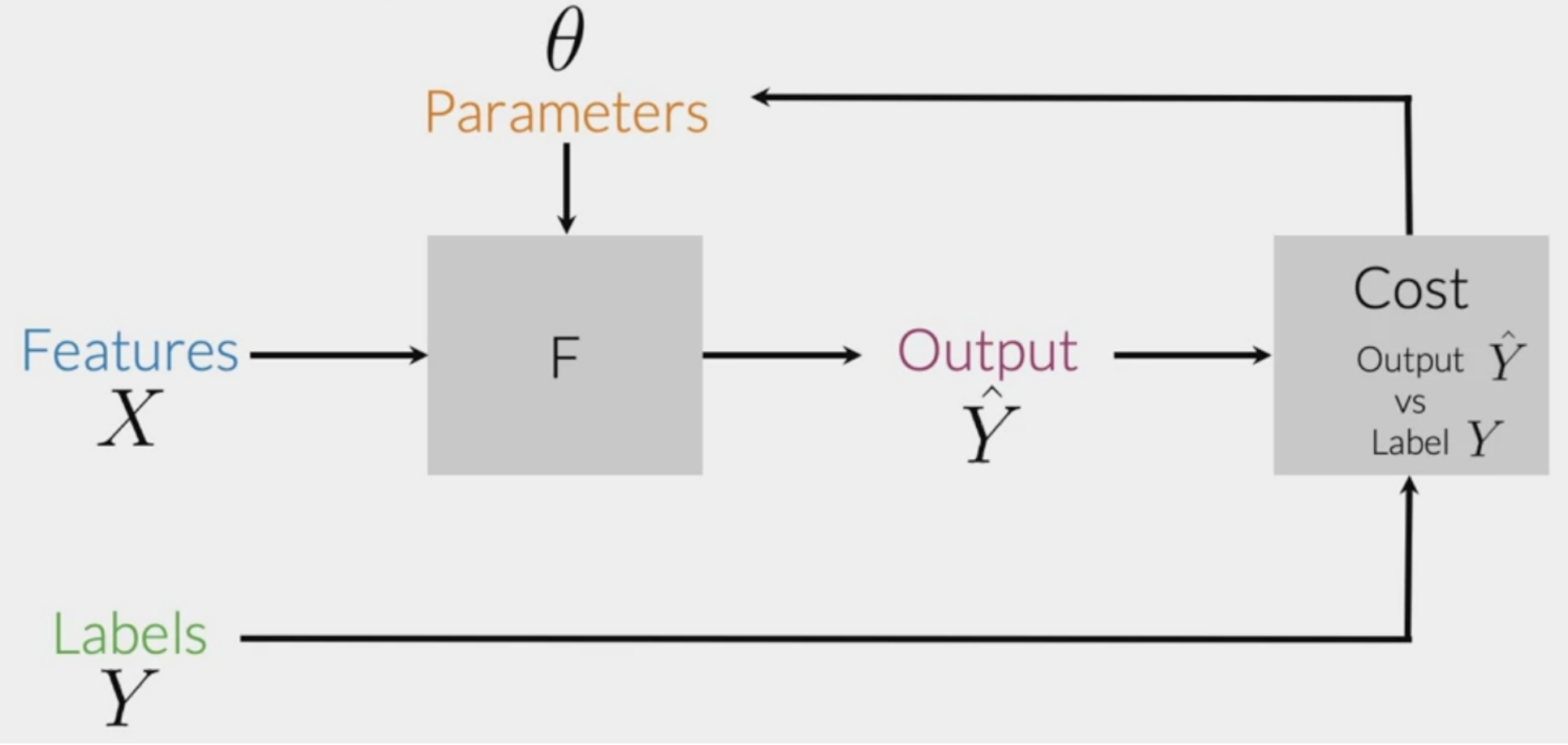
$$\begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(m)} & X_2^{(m)} \end{bmatrix} \longleftrightarrow$$

[[1, 40, 20]
[1, 20, 50],
...
[1, 5, 35]]

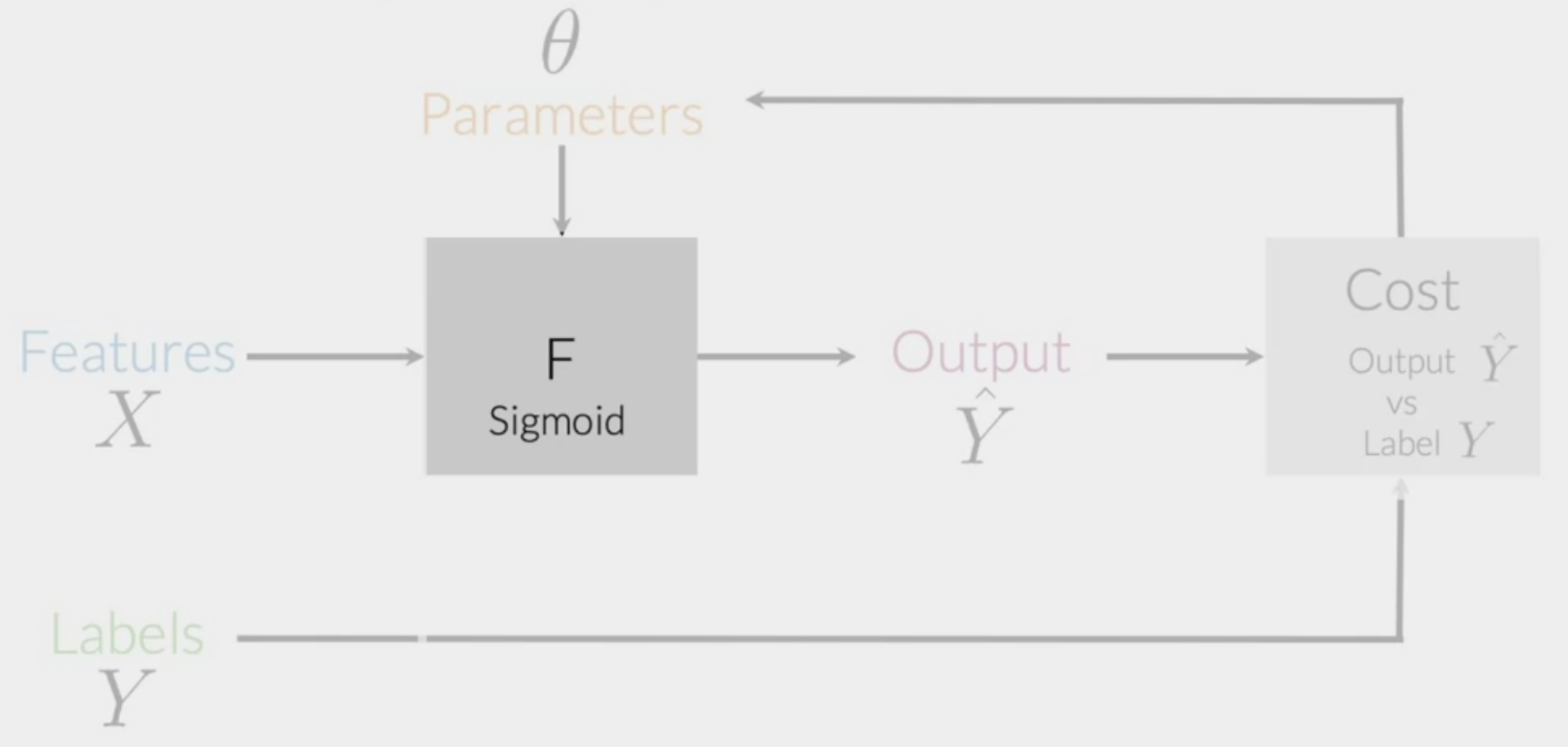
Putting it all Together

```
freqs = build_freqs(tweets,labels) #Build frequencies dictionary

X = np.zeros((m,3)) #Initialize matrix X

for i in range(m): #For every tweet

        p_tweet = process_tweet(tweets[i]) #Process tweet

    X[i,:] = extract_features(p_tweet,freqs) #Extract Features
```

Overview of Logistic Regression

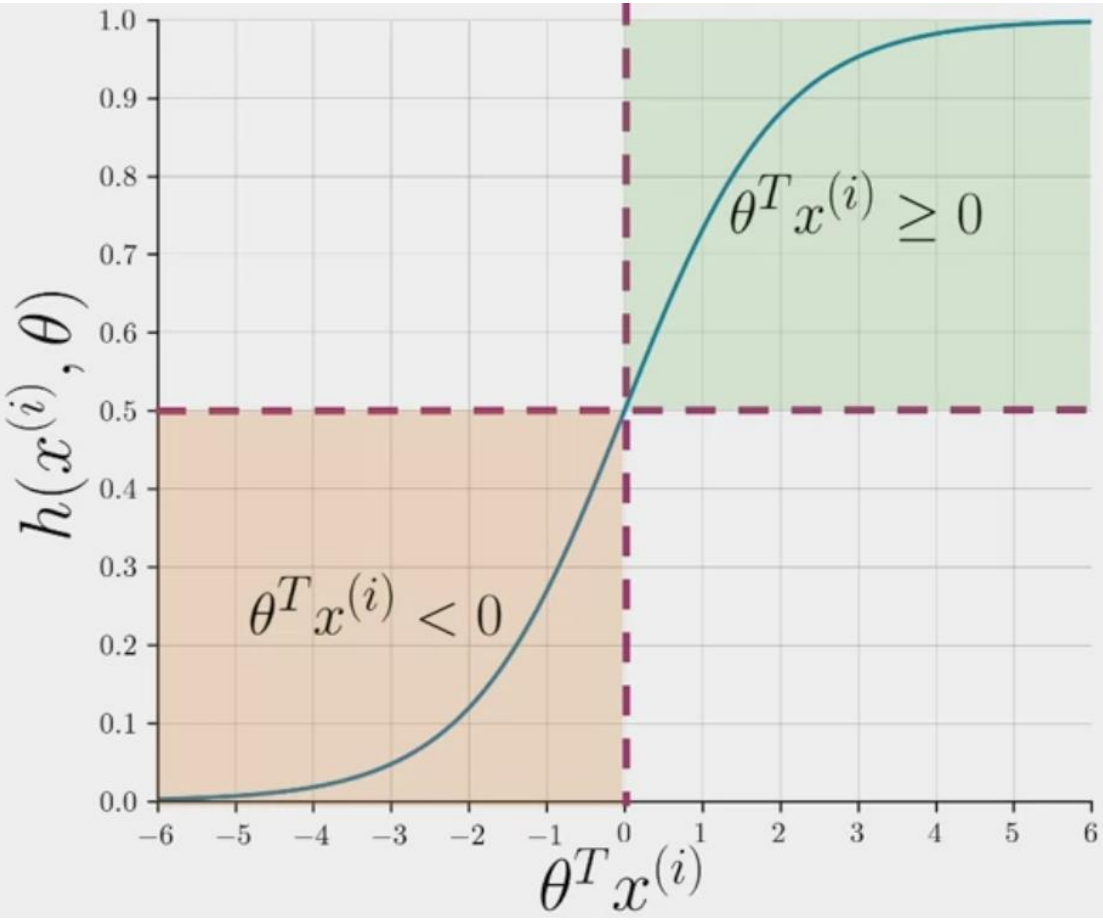


Overview of Logistic Regression



Overview of Logistic Regression

$$h(x^{(i)}, \theta) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$



Overview of Logistic Regression

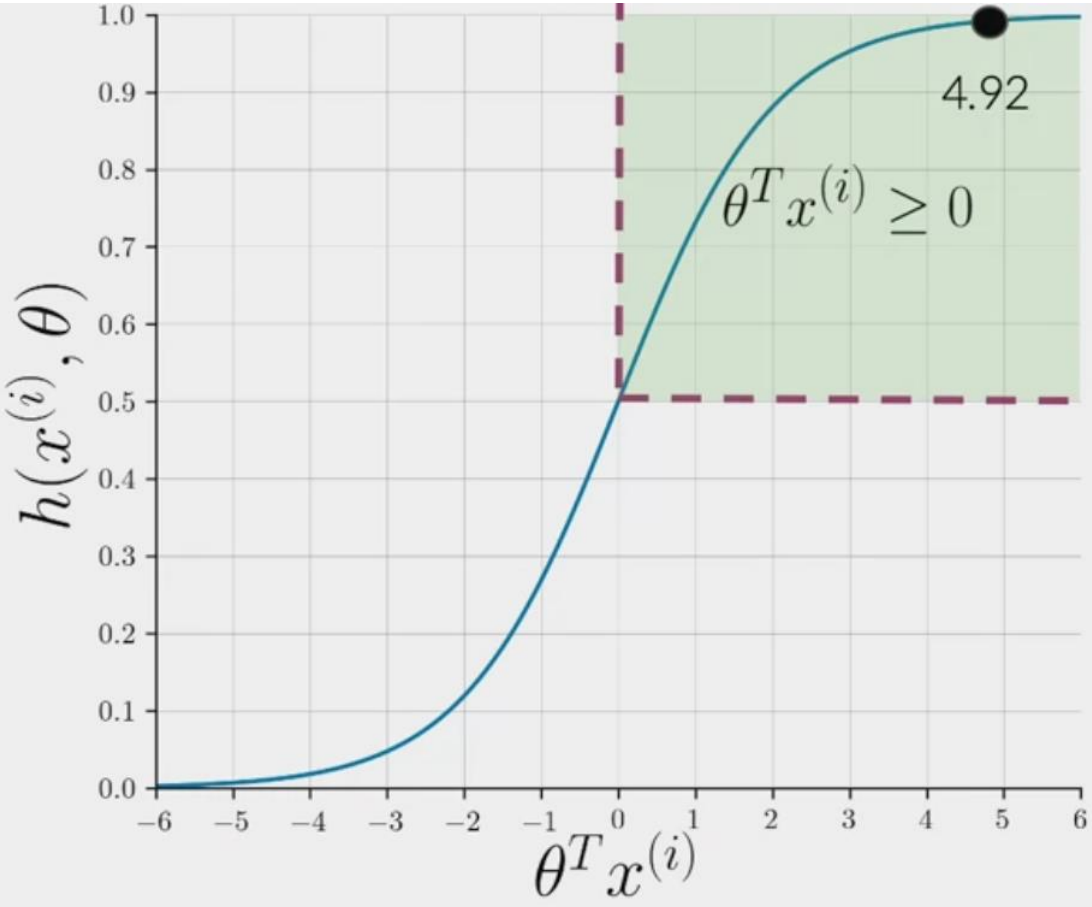
@YMourri and
@AndrewYNg are tuning a
GREAT AI model

↓

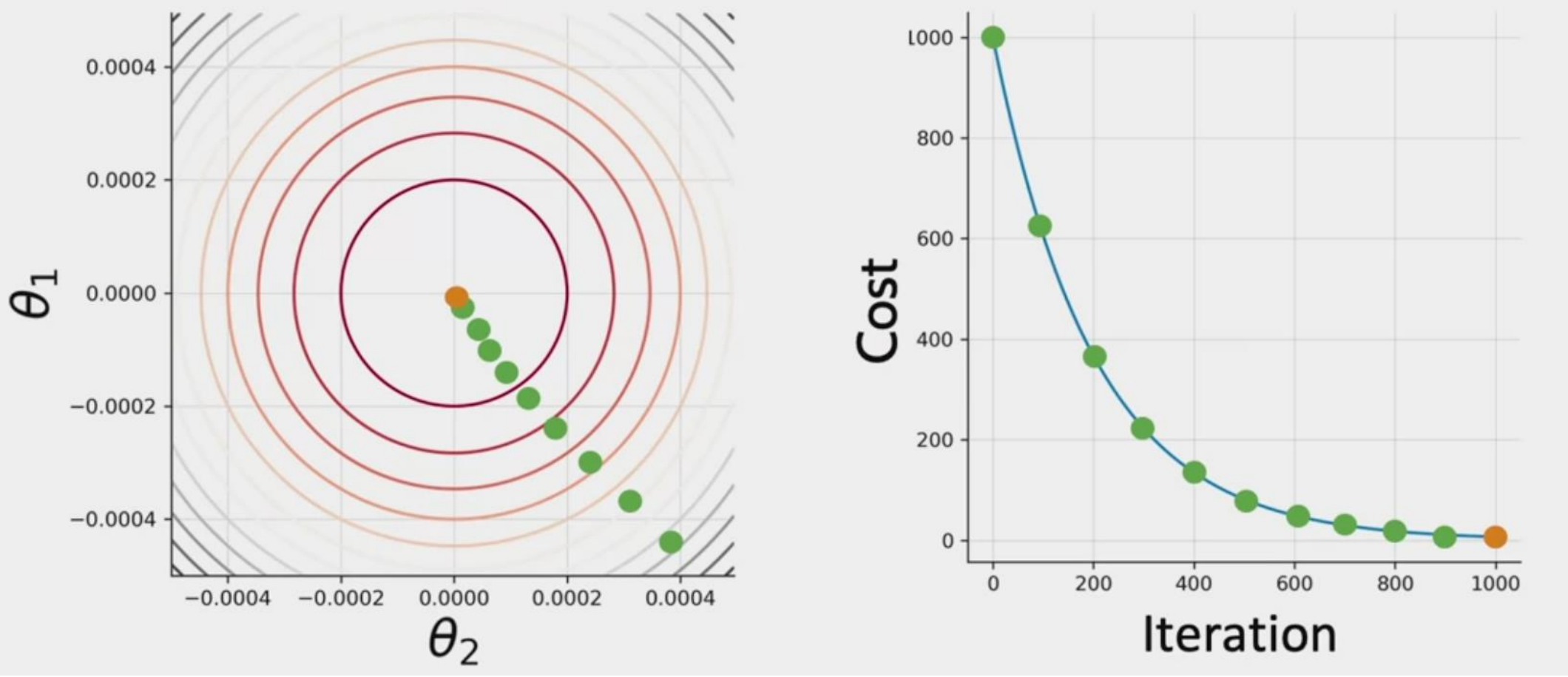
[tun, ai, great, model]

↓

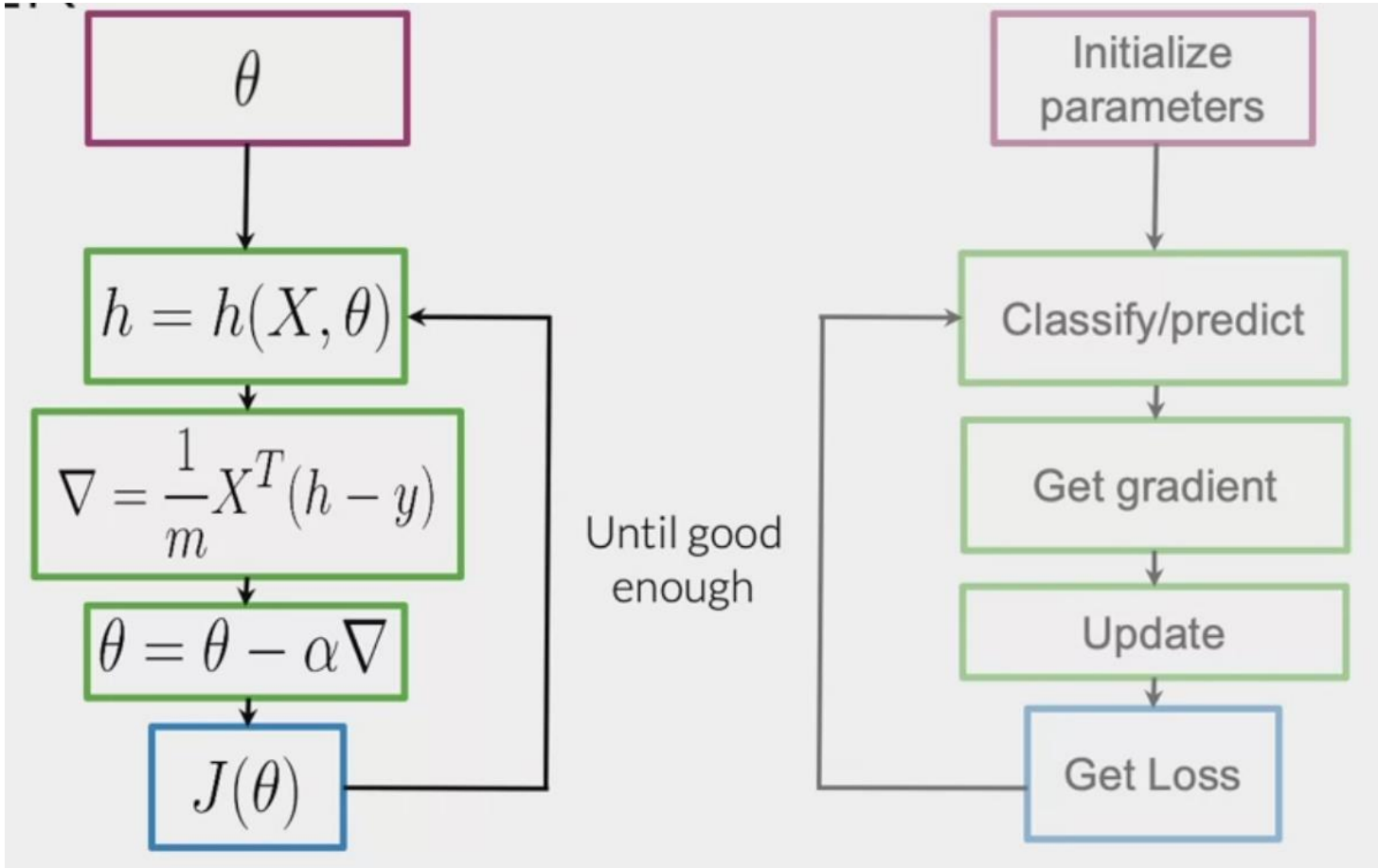
$$x^{(i)} = \begin{bmatrix} 1 \\ 3476 \\ 245 \end{bmatrix} \quad \theta = \begin{bmatrix} 0.00003 \\ 0.00150 \\ -0.00120 \end{bmatrix}$$



Training Logistic Regression



Training Logistic Regression

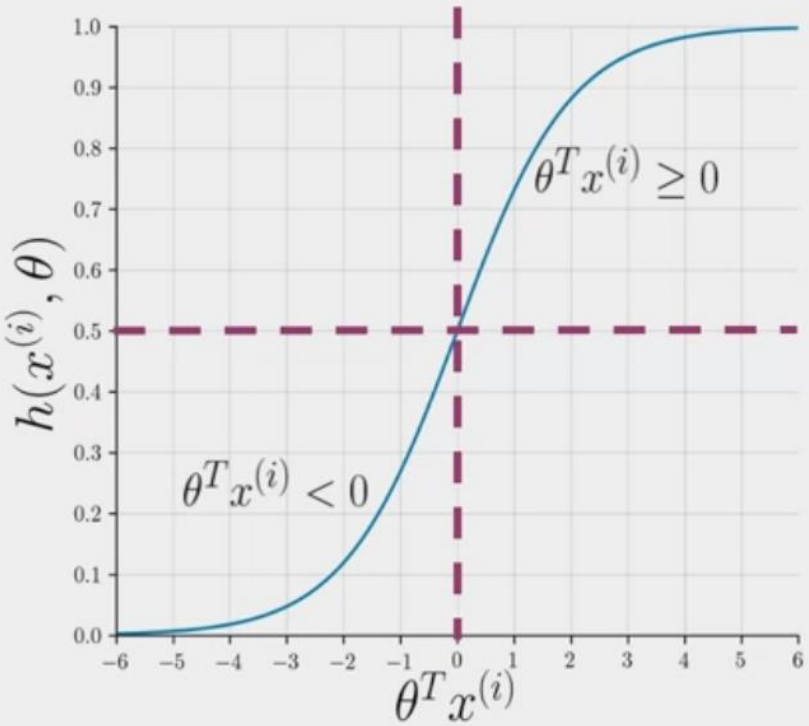


Testing Logistic Regression

- X_{val} Y_{val} θ

$$h(X_{val}, \theta)$$

$$pred = h(X_{val}, \theta) \geq 0.5$$



Testing Logistic Regression

- X_{val} Y_{val} θ

$$h(X_{val}, \theta)$$

$$pred = h(X_{val}, \theta) \geq 0.5$$

$$\begin{bmatrix} 0.3 \\ 0.8 \\ 0.5 \\ \vdots \\ h_m \end{bmatrix} \geq 0.5 = \begin{bmatrix} 0.3 \geq 0.5 \\ 0.8 \geq 0.5 \\ 0.5 \geq 0.5 \\ \vdots \\ pred_m \geq 0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ pred_m \end{bmatrix}$$

Testing Logistic Regression

- X_{val} Y_{val} θ

$$h(X_{val}, \theta)$$

$$pred = h(X_{val}, \theta) \geq 0.5$$

$$\sum_{i=1}^{m} \frac{(pred^{(i)} == y_{val}^{(i)})}{m}$$

Testing Logistic Regression

- X_{val} Y_{val} θ

$$h(X_{val}, \theta)$$

$$pred = h(X_{val}, \theta) \geq 0.5$$

$$\sum_{i=1}^{m} \frac{(pred^{(i)} == y_{val}^{(i)})}{m}$$

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ pred_m \end{bmatrix} == \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ Y_{val_m} \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ pred_m == Y_{val_m} \end{bmatrix}$$

Testing Logistic Regression

$$Y_{val} = \begin{bmatrix} 0 \\ 1 \\ \underline{1} \\ 0 \\ 1 \end{bmatrix} \quad pred = \begin{bmatrix} 0 \\ 1 \\ \underline{0} \\ 0 \\ 1 \end{bmatrix} \quad (Y_{val} == pred) = \begin{bmatrix} 1 \\ 1 \\ \underline{0} \\ 1 \\ 1 \end{bmatrix}$$

$$\boxed{\text{accuracy} = \frac{4}{5} = 0.8}$$