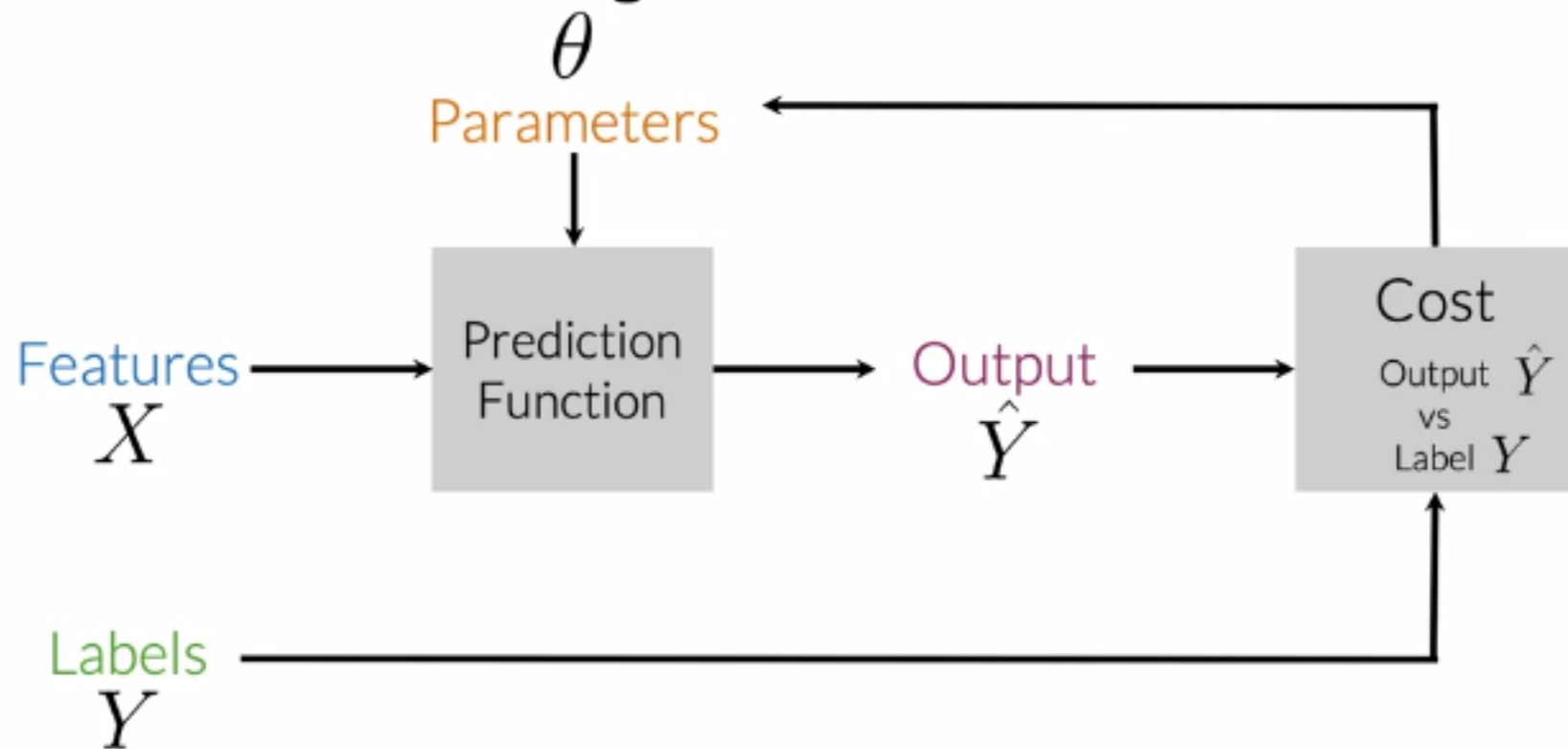# Supervised ML (training)

Sentiment analysis
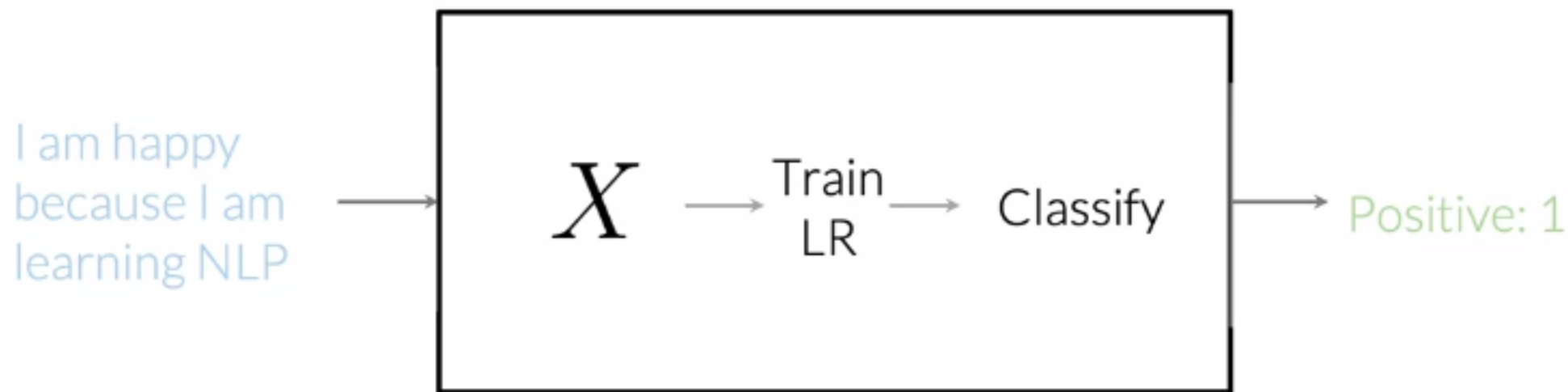
Tweet: I am happy because I am learning NLP

Positive: 1

Negative: 0

Logistic regression

Sentiment analysis

I am happy
because I am
learning NLP $\longrightarrow$ $\boxed{X \longrightarrow \begin{array}{c}\text{Train}\\\text{LR}\end{array} \longrightarrow \text{Classify}}$ $\longrightarrow$ Positive: 1

Vocabulary

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

$\longrightarrow$

I am happy because I am learning NLP

...
...
...

I hated the movie

$V =$

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

Feature extraction
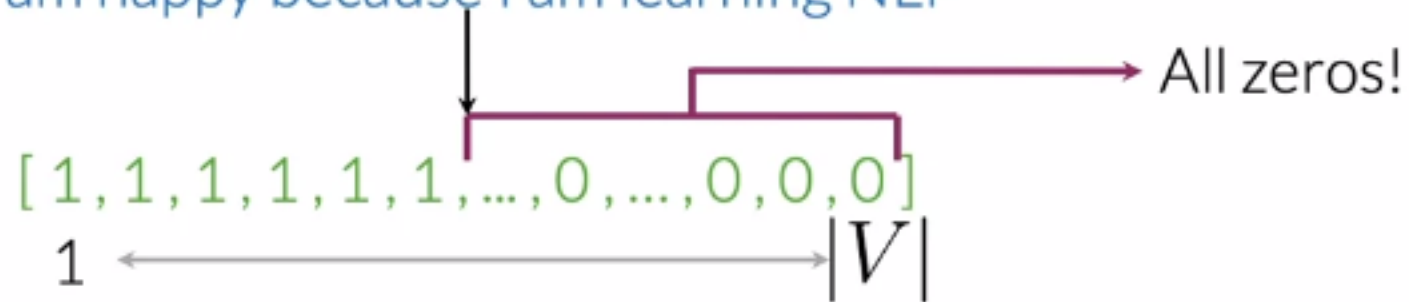
I am happy because I am learning NLP

[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

$$[\,1\,,1\,,1\,,1\,,1\,,1\,,\ldots,0\,,\ldots,0\,,0\,,0\,]$$

All zeros!

$1 \longleftrightarrow |V|$

Problems with sparse representations

I am happy because I am learning NLP

All zeros!

$$[\,1\,,\,1\,,\,1\,,\,1\,,\,1\,,\,1\,,\,\ldots\,,\,0\,,\,\ldots\,,\,0\,,\,0\,,\,0\,]$$

$1 \qquad\qquad\qquad\qquad |V|$

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

$$n = |V|$$

Problems with sparse representations

I am happy because I am learning NLP

All zeros!

$$[\,1\,,1\,,1\,,1\,,1\,,1\,,\ldots,0\,,\ldots,0\,,0\,,0\,]$$

$1 \longleftarrow\!\!\!\longrightarrow |V|$

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

$$n = |V|$$

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

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

I am happy because I am learning NLP

I am happy

Negative tweets

I am sad, I am not learning NLP

I am sad

Positive and negative counts

Positive tweets

I am happy because I am learning NLP

I am happy

Vocabulary

I

am

happy

because

learning

NLP

sad

not

Positive and negative counts

Positive tweets

I am happy because I am learning NLP
I am happy

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

Positive and negative counts

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

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	1
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	1
not	0	1

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

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

Feature extraction

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

$$X_m = [\qquad\qquad\qquad\qquad\qquad\qquad\qquad]$$

Features of
tweet m

Feature extraction

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

$$X_m = [1, \qquad\qquad\qquad\qquad]$$

Features of
tweet m

Bias

Feature extraction

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

$$X_m = \left[1, \sum_w freqs(w, 1), \qquad\qquad\right]$$

Features of
tweet m

Bias

Sum Pos.
Frequencies

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

I am sad, I am not learning NLP

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

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

$$X_m = [1, \sum_w \textbf{\textit{freqs}}(w, 1), \sum_w \textit{freqs}(w, 0)]$$

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

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

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

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

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

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

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

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

11

Feature extraction

I am sad, I am not learning NLP

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

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: stop words and punctuation

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

Preprocessing: stop words and punctuation

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

Stop words
and
is
are
at
has
for
a

Punctuation
,
.
:
!
"
'

Preprocessing: stop words and punctuation

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

Stop words
and
is
are
at
has
for
a

Punctuation
,
.
:
!
"
'

Preprocessing: stop words and punctuation

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

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

Stop words
and
is
are
at
has
for
a

Punctuation
,
.
:
!
"
'

Preprocessing: stop words and punctuation

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

Stop words
and
is
a
at
has
for
of

Punctuation
,
.
:
!
"
'

Preprocessing: stop words and punctuation

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

Stop words
and
is
a
at
has
for
of

Punctuation
,
.
:
!
"
'

Preprocessing: stop words and punctuation

@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: Handles and URLs

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

Preprocessing: Handles and URLs

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

Preprocessing: Handles and URLs

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

tuning GREAT AI model

Preprocessing: Stemming and lowercasing

tuning GREAT AI model

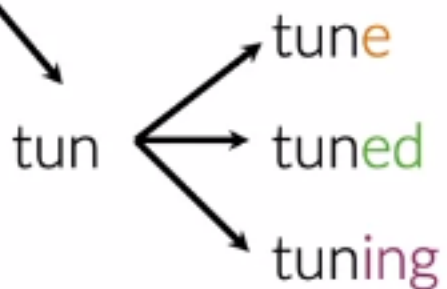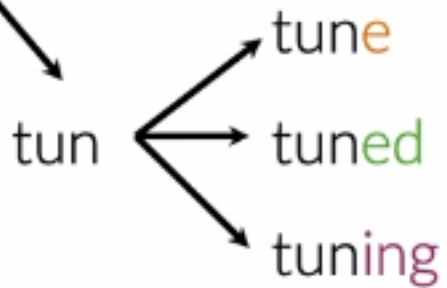# Preprocessing: Stemming and lowercasing

tuning GREAT AI model

tun

Preprocessing: Stemming and lowercasing

tuning GREAT AI model

tun → tune
tun → tuned
tun → tuning

Preprocessing: Stemming and lowercasing

tuning GREAT AI model

tun → tune
tun → tuned
tun → tuning

GREAT

Great

great

Preprocessing: Stemming and lowercasing

tuning GREAT AI model

tun → tune

tun → tuned

tun → tuning

GREAT

Great → great

great

Preprocessing: Stemming and lowercasing

tuning GREAT AI model

tun → tune
tun → tuned
tun → tuning

GREAT
Great → great
great

Preprocessed tweet:
[tun, great, ai, model]

General overview

I am Happy Because i am learning NLP @deeplearning

↓ Preprocessing

[happy, learn, nlp]

General overview

I am Happy Because i am learning NLP @deeplearning

↓ Preprocessing

[happy, learn, nlp]

↓ Feature Extraction

[1, 4, 2]

General overview

I am Happy Because i am learning NLP @deeplearning

Preprocessing

[happy, learn, nlp]

Feature Extraction

Bias ⟵ [1, 4, 2]

General overview

I am Happy Because i am learning NLP @deeplearning

$\downarrow$ Preprocessing

[happy, learn, nlp]

$\downarrow$ Feature Extraction

Bias $\longleftarrow$ [1, 4, 2]

$\downarrow$

Sum positive frequencies

General overview

I am Happy Because i am learning NLP @deeplearning

| Preprocessing

[happy, learn, nlp]

| Feature Extraction

Bias ←— [1, 4, 2] —→ Sum negative frequencies

Sum positive frequencies

General overview

I am Happy Because i am
learning NLP
@deeplearning

I am sad not learning NLP

...

I am sad :(

General overview

I am Happy Because i am
learning NLP
@deeplearning

I am sad not learning NLP $\longrightarrow$

...

I am sad :(

[happy, learn, nlp]

[sad, not, learn, nlp]

...

[sad]

General overview

I am Happy Because i am
learning NLP
@deeplearning

I am sad not learning NLP $\longrightarrow$

...

I am sad :(

[happy, learn, nlp]

[sad, not, learn, nlp]

...

[sad]

$\longrightarrow$

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

General overview

$$\begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(m)} & X_2^{(m)} \end{bmatrix} \longleftrightarrow \begin{matrix} [[1, 40, 20], \\ [1, 20, 50], \\ ... \\ [1, 5, 35]] \end{matrix}$$

General Implementation

```python
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
```