01/07/2020

Natural Language Processing

[deeplearning.ai]

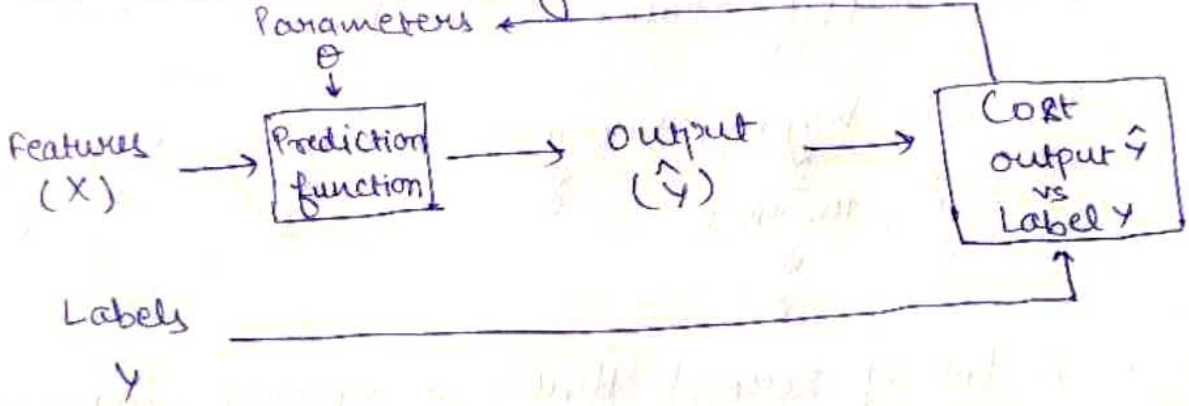
Natural Language Processing with classification and Vector Spaces :-

WEEK-1 : Vocabulary & Feature Extraction

1. Sentiment Analysis with logistic Regression

→ Supervised ML & Sentiment Analysis

→ Supervised ML (training)



→ Sentiment Analysis

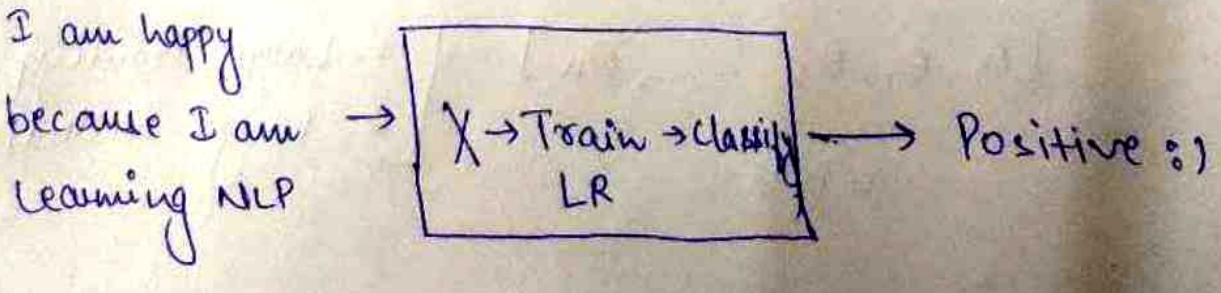
Tweet : I am happy because I am learning NLP.

→ Positive : 1

→ Negative : 0

↓

Logistic Regression



Vocabulary & Feature Extraction

→ 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.

$$[\underset{\substack{\downarrow\\1}}{I}, \underset{\substack{\downarrow\\1}}{am}, \underset{\substack{\downarrow\\1}}{happy}, \underset{\substack{\downarrow\\1}}{because}, \underset{\substack{\downarrow\\1}}{learning}, \underset{\substack{\downarrow\\1}}{NLP}, \underset{\substack{\downarrow\\ \cdots}}{\ldots}, \underset{\substack{\downarrow\\0}}{hated}, \underset{\substack{\downarrow\\0}}{the}, \underset{\substack{\downarrow\\0}}{movie}]$$

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

→ Problems with sparse representations

I am happy because I am learning NLP

$$\downarrow$$

$$[1,1,1,1,1,1,\overbrace{\ldots,0,\ldots,0,0,0}^{\text{All zeros!}}]$$

$$1 \longleftrightarrow |V|$$

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

1. Large Training time
2. Large Prediction time

$n = |V|$

Negative & Positive Frequencies

→ 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 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

Negative Tweets

I am sad, I am not learning NLP
I am sad

→

vocabulary	Negfreq (0)
I	3
am	3
happy	0
because	0
learning	1
NLP	1
sad	2
not	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