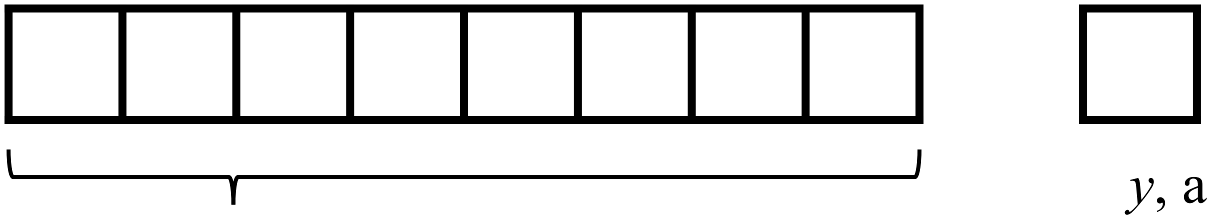
Natural Language Processing with Bag of Words Models

Matthew Engelhard

Lecture 1: what is a predictive model?

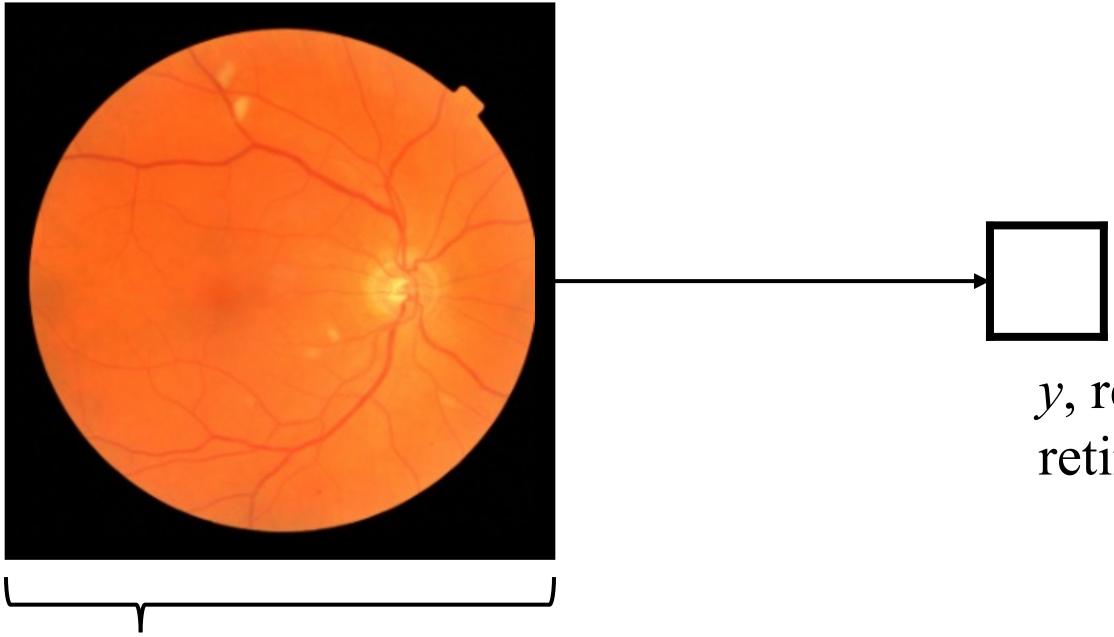


x, data/features for
a subject or patient

y, associated
value or label

End goal: predict *y* from *x*

CNN: a predictive model for image data

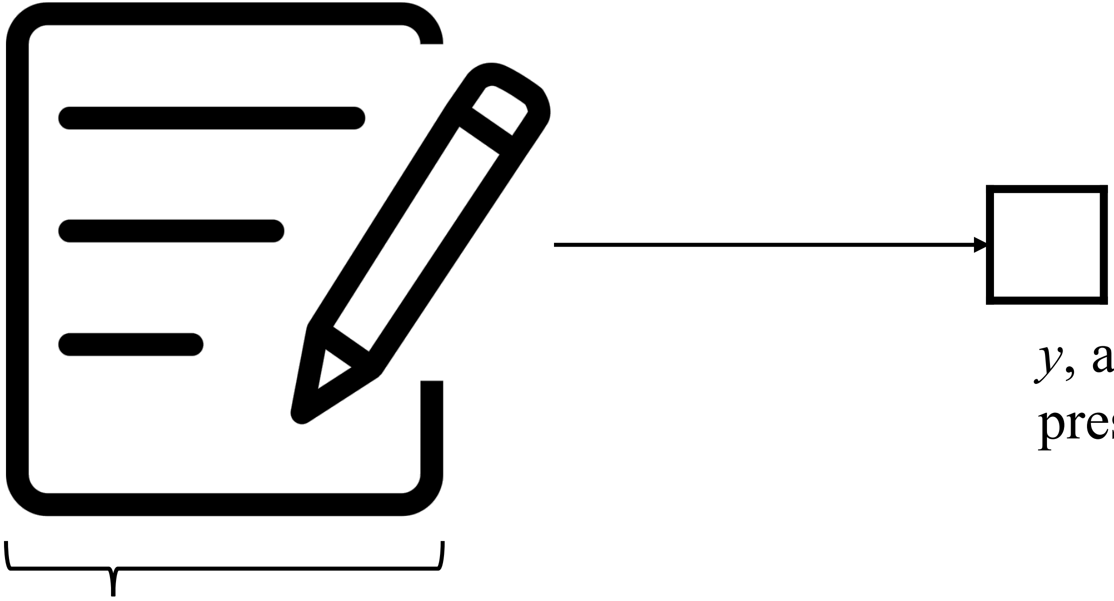


x, retinal image

y, referrable diabetic retinopathy

End goal: predict y from x

NLP: a predictive model for text data



y, abnormality present

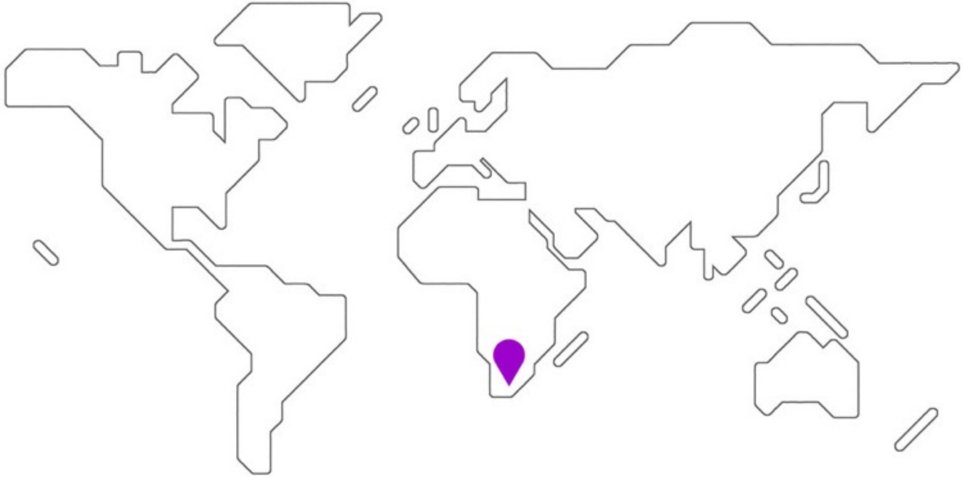
x, radiology report

End goal: predict y from x

Case Study: SMS Triage for Global Maternal Health

Maternal Health HelpDesk:

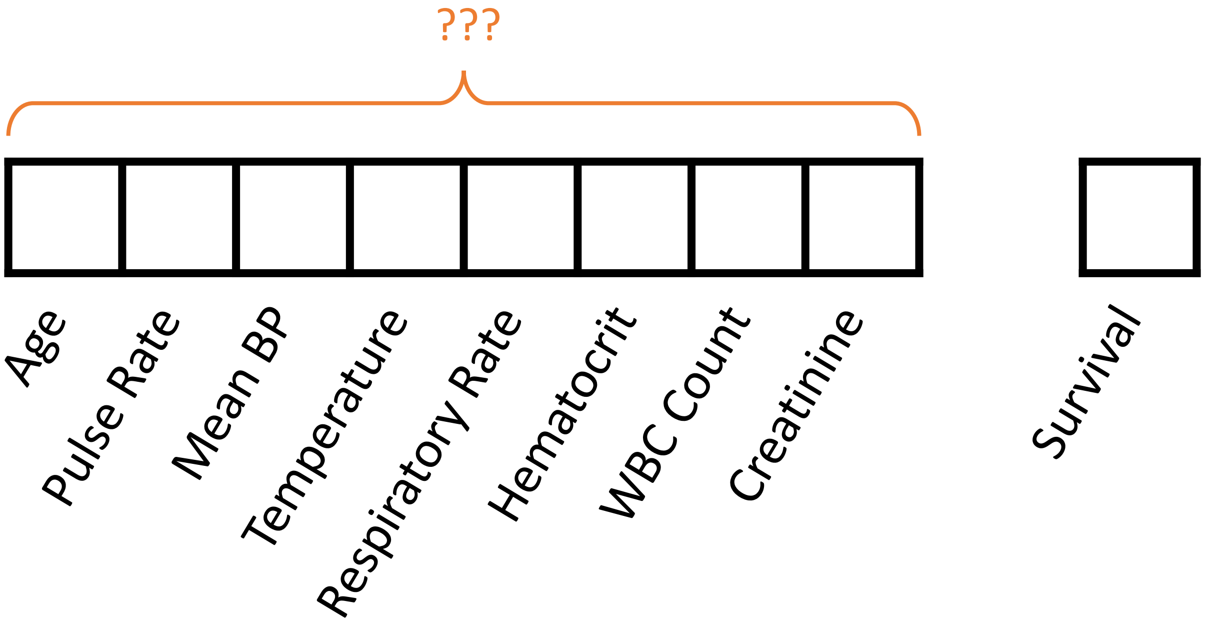
2 million women connected to NDoH staff via SMS



https://www.praekelt.org

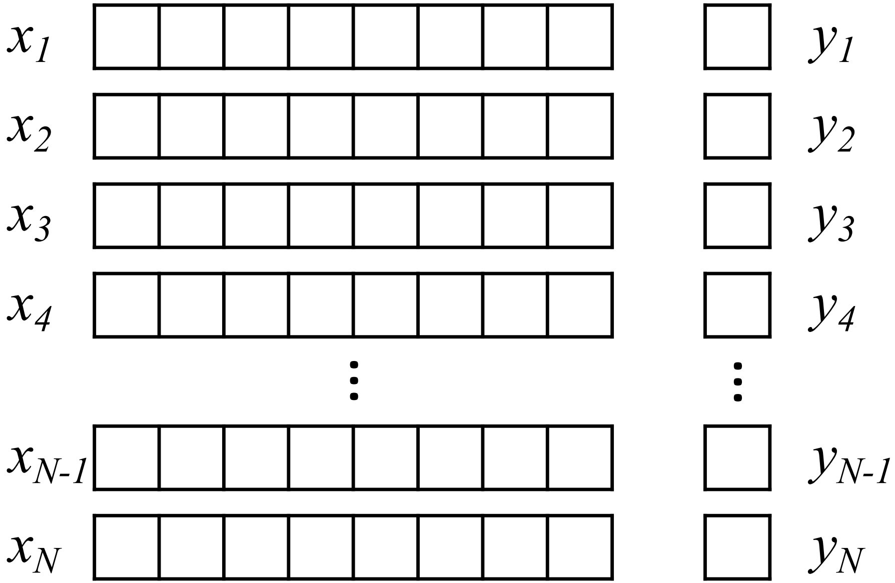
Binary Classification: Urgent Message? (Yes/No)

A Simple Predictive Model: ICU Mortality



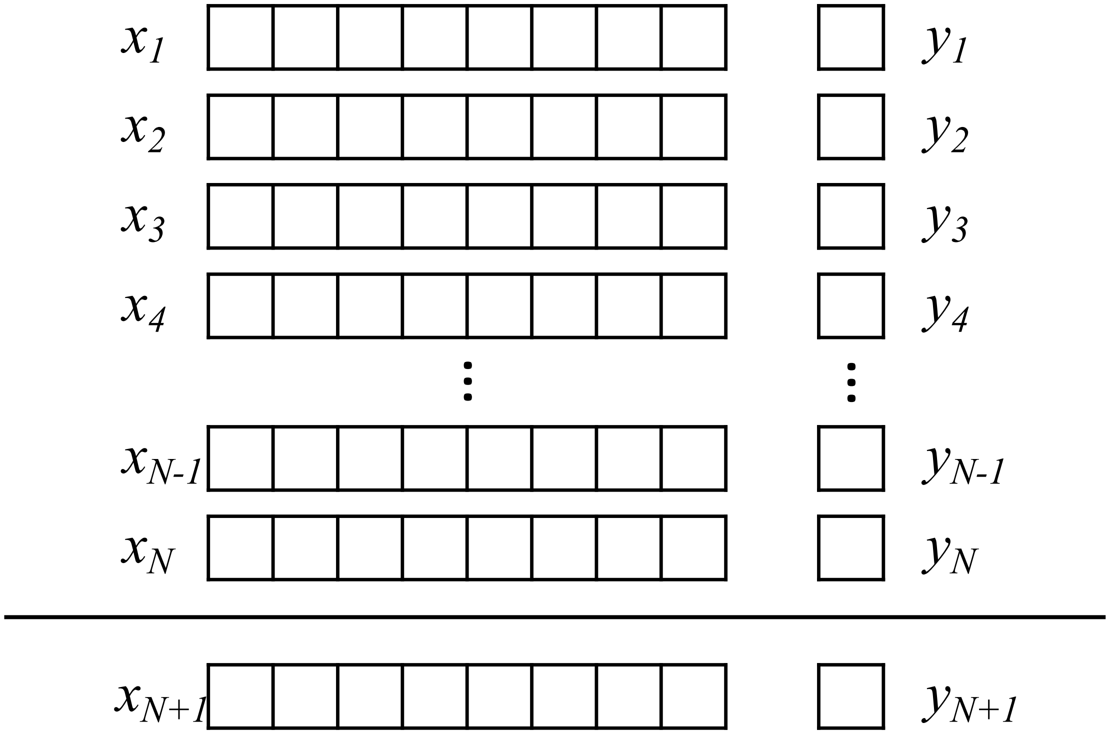
End goal: predict odds of hospital mortality

Training Set (Historical Data)



Find an equation that predicts y based on x across the training set

Making Predictions for New x



Find an equation that predicts y based on x across the training set

<- Learn to predict new y

This time, our training data is text

x_1	What helps with morning sickness?	□	y_1
x_2	How many months should I breastfeed?	□	y_2
x_3	I passed out and Mom said I was shaking	□	y_3
x_4	Where is the nearest clinic?	□	y_4
	⋮	⋮	
x_{N-1}	I am having heavy bleeding, what should I do?	□	y_{N-1}
x_N	What foods should I eat while pregnant?	□	y_N
x_{N+1}	My heart is racing and I can’t catch my breath	□	y_{N+1}

y_i: Urgent or Not Urgent?

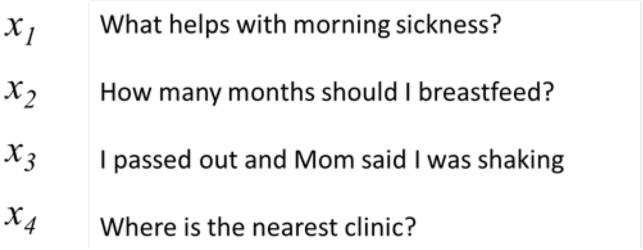
<- Learn to predict new y

We need numbers, not words

- **Can we convert our text to a vector or sequence of numbers?**

- If yes, we can use logistic regression (or any other predictive model)!

First try: count words in each SMS
Step 1: Define a **vocabulary** of words



x_1	What helps with morning sickness?
x_2	How many months should I breastfeed?
x_3	I passed out and Mom said I was shaking
x_4	Where is the nearest clinic?

list of all words
(in no particular order)

shaking
what
clinic
how
helps
was
nearest
many

with
said
months
the
morning
mom
should
sickness

and
I
is
how
out
breastfeed
passed
where

Step 2: count how many times each vocabulary word appears in a given SMS

What helps with morning sickness?

x_1	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
	shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Step 2: count how many times each vocabulary word appears in a given SMS

I passed out and Mom said I was shaking

x_3	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	2	0	0	1	0	1	0
	shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Step 2: count how many times each vocabulary word appears in a given SMS

Where is the nearest clinic?

x_4	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
	shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

Note that word order does not matter!

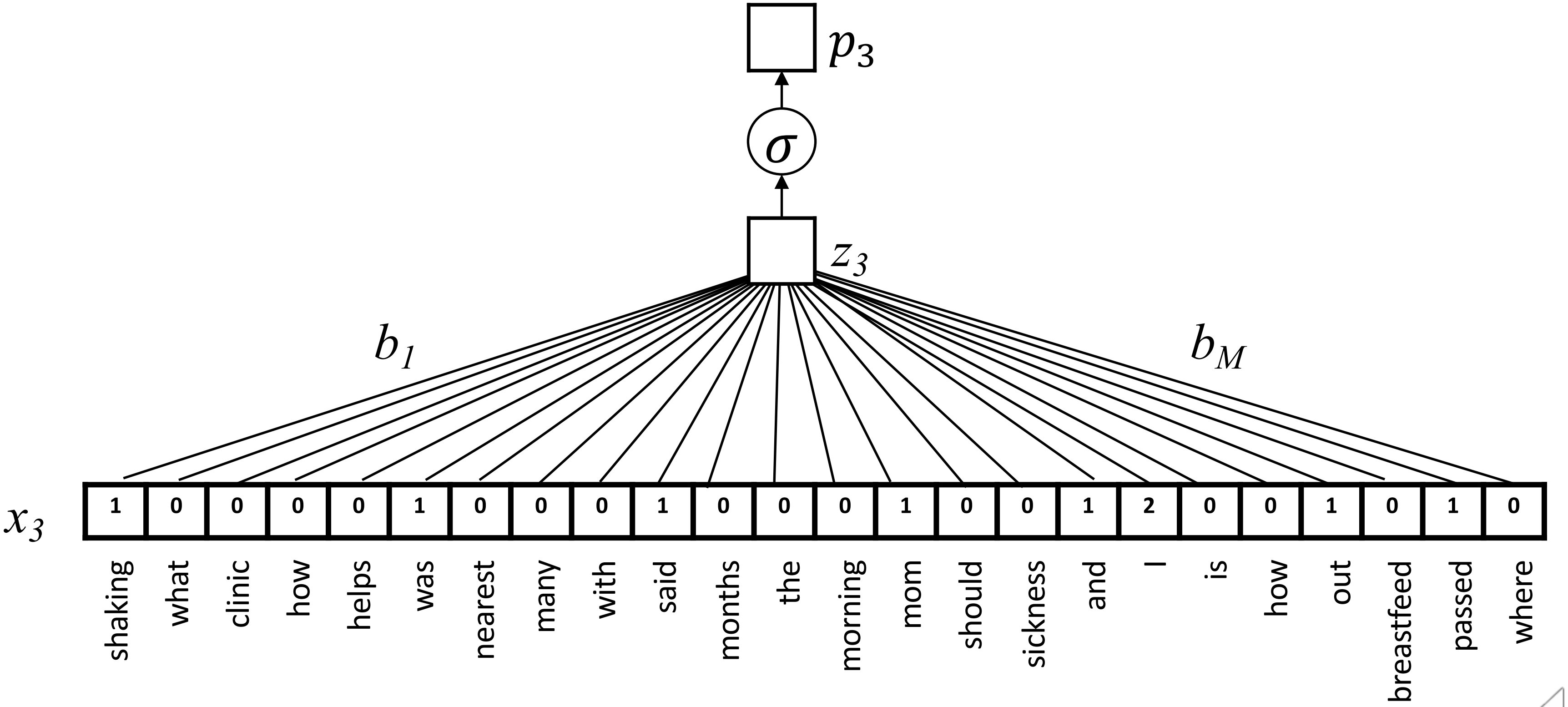
clinic is where nearest the

x_4 ,	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
	shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where

A “bag of words”

clinic
nearest
where
the
is

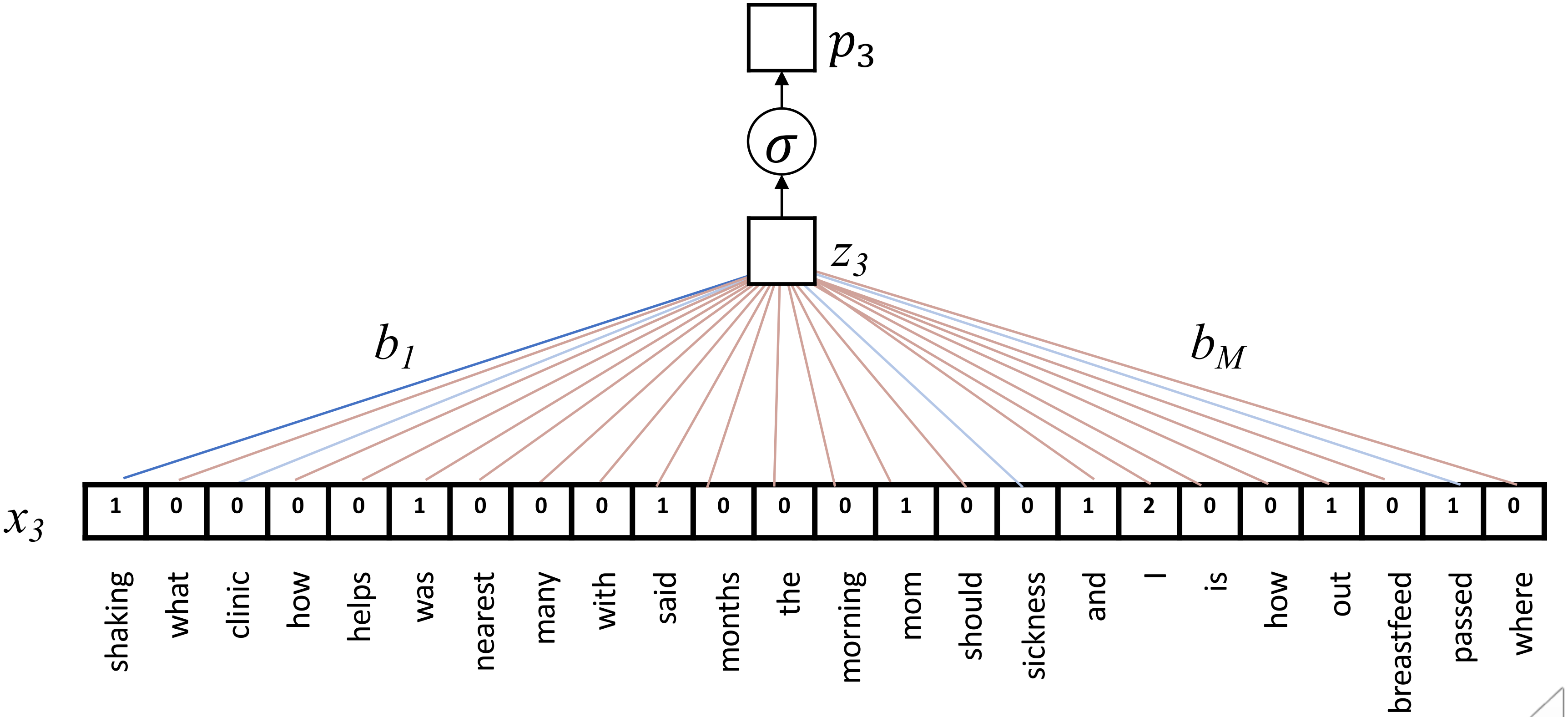
Logistic Regression for Text Classification



shaking	what	clinic	how	helps	was	nearest	many	with	said	months	the	morning	mom	should	sickness	and	I	is	how	out	breastfeed	passed	where
1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	2	0	0	1	0	1	0

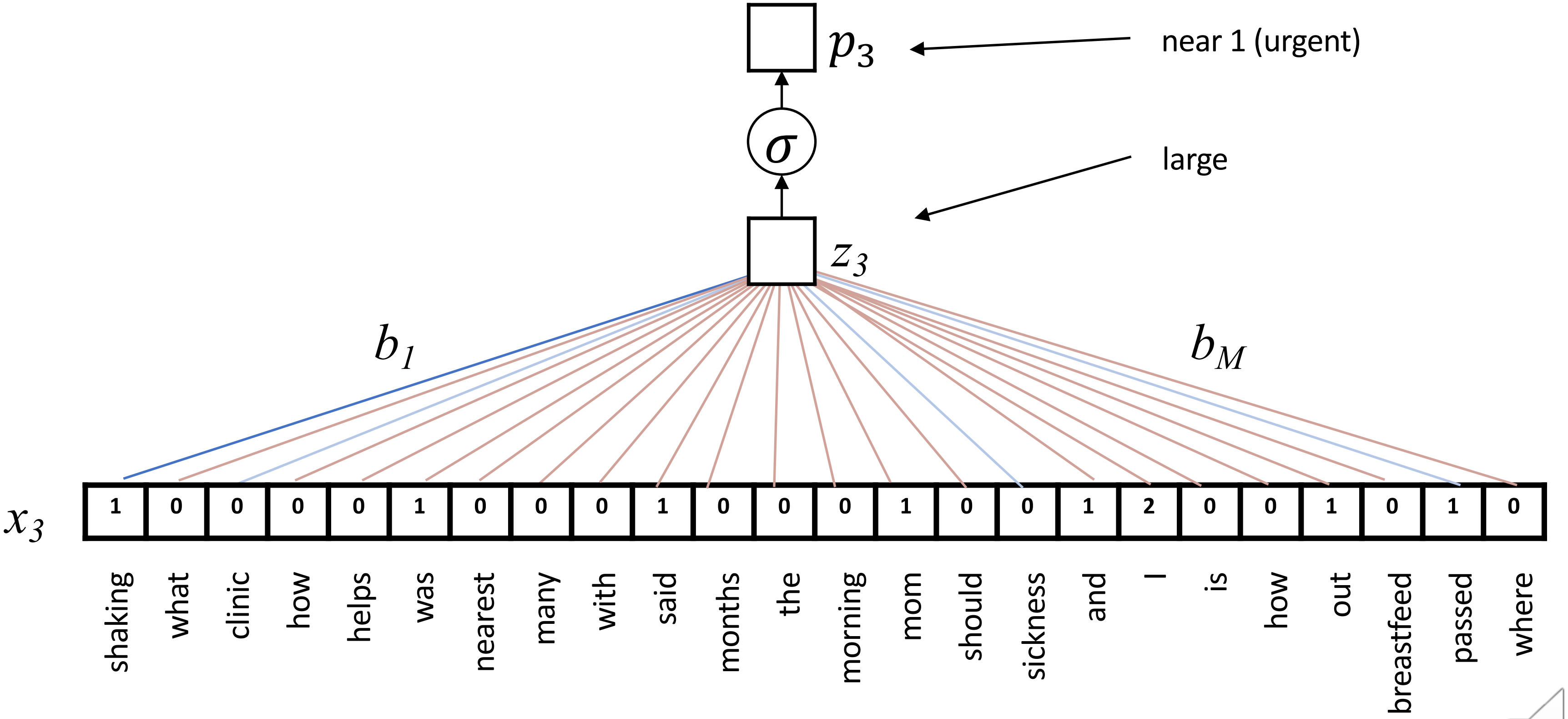
I passed out and Mom said I was shaking

Logistic Regression for Text Classification



I passed out and Mom said I was shaking

Logistic Regression for Text Classification

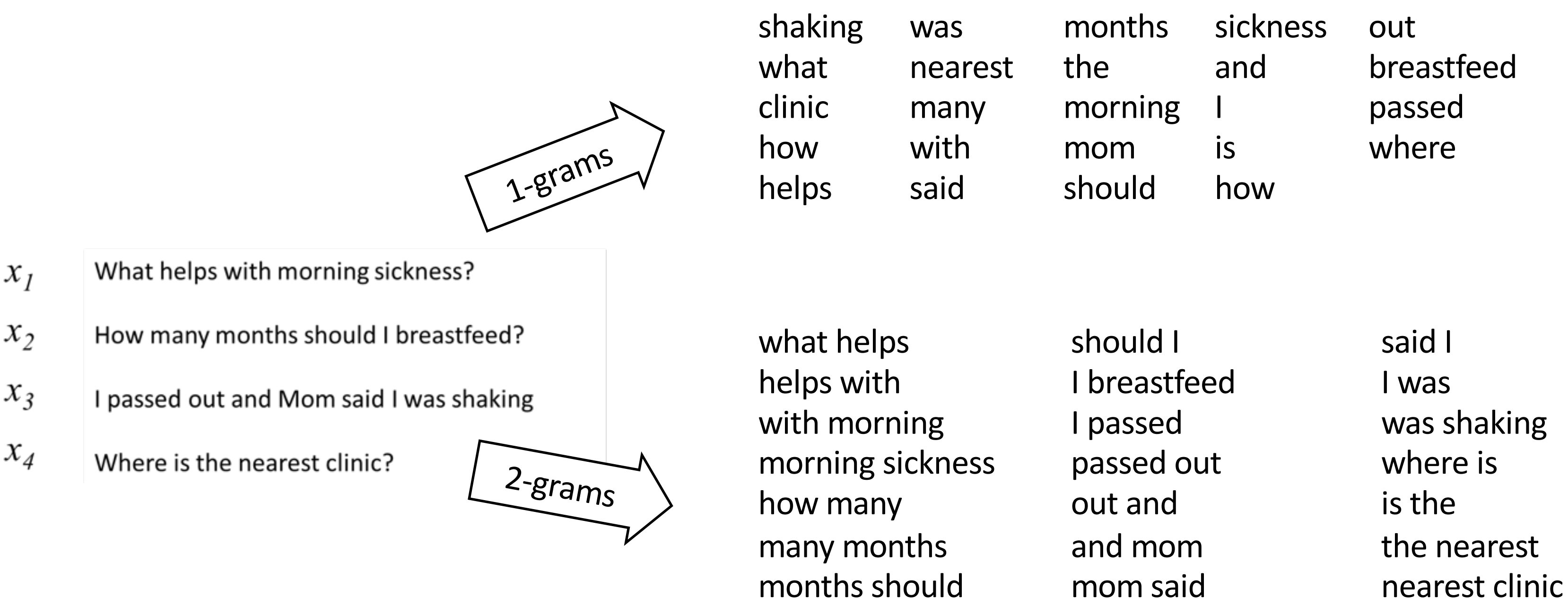


I passed out and Mom said I was shaking

Strengths and Weaknesses

- (+) This approach is simple and works surprisingly well in practice
- (+) Often the best approach with small datasets
- (-) Does not capture word order
- (-) Does not group synonyms together or understand semantic relationships between words

2nd try: count 1- and 2-grams in each SMS (i.e. extend vocabulary to include 2-word phrases)



x_1	What helps with morning sickness?
x_2	How many months should I breastfeed?
x_3	I passed out and Mom said I was shaking
x_4	Where is the nearest clinic?

shaking
what
clinic
how
helps
was
nearest
many
with
said
months
the
morning
mom
should
sickness
and
I
is
how
out
breastfeed
passed
where

what helps
helps with
with morning
morning sickness
how many
many months
months should
should I
I breastfeed
I passed
passed out
out and
and mom
mom said
said I
I was
was shaking
where is
is the
the nearest
nearest clinic

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick

Bag of 1-grams: no difference between these sentences

n-grams can be very helpful!

I am not sick and feel great

I am not great and feel sick

Bag of 1- and 2-grams:

not sick, **feel great**

versus

not great, **feel sick**

3rd try: more powerful methods to work with…

- (a) word meaning: assign words to vectors that encode their meaning numerically

- (b) words in context: neural network architectures that act on *sequences* of words (rather than a bag of words)

More Text Processing Details

(for bag of words models)

Variations on counting: term frequency

term count: ‘times’

2

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

1

“And the first one now
Will later be last
For the times they are a-changin'.”

Variations on counting: term frequency

term frequency: ‘times’

2/119

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

1/16

“And the first one now
Will later be last
For the times they are a-changin'.”

-> better measure of the importance of the term within a given text sample

Variations on counting: inverse document frequency

2/2 **document frequency: ‘times’**

✓ “It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

✓ “And the first one now
Will later be last
For the times they are a-changin'.”

Variations on counting: inverse document frequency

1/2

document frequency: ‘evil’

✓

“It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.”

✗

“And the first one now
Will later be last
For the times they are a-changin'.”

term frequency-inverse document frequency (tf-idf)

- What helps with morning sickness?
- How many months should I breastfeed?
- I passed out and Mom said I was shaking
- Where is the nearest clinic?
- I am having heavy bleeding, what should I do?
- What foods should I eat while pregnant?
- My heart is racing and I can’t catch my breath

$$\frac{\text{term frequency}}{\text{document frequency}} \quad \text{for ‘shaking’}$$

$$\frac{1/9}{1/7} = .78$$

$$\frac{\text{term frequency}}{\text{document frequency}} \quad \text{for ‘I’}$$

$$\frac{2/9}{5/7} = .31$$

Preprocessing

```
I passed out, and Mom said I was shaking.
```

- remove punctuation

```
I passed out and Mom said I was shaking
```

- to lowercase

```
i passed out and mom said i was shaking
```

- “tokenization”

```
[i, passed, out, and, mom, said, i, was, shaking]
```

- “stemming”

```
[i, pass, out, and, mom, said, i, wa, shak
```

Summary

- A central challenge of NLP lies in converting text documents into feature vectors that can be used in a predictive model

- Bag of words models solve this challenge by constructing a feature vector based on counts of each word of interest

- Even though they ignore word order and semantic relationships, these models are very powerful