LANGUAGE MODELING

David Kauchak
CS159 – Fall 2024

some slides adapted from Jason Eisner

1

Admin

How did assignment 1 finish up?

Assignment 2two part assignment)

- Two part assignment
- 2a out now: due next Thursday (work through calculations by hand)
- 2b out soon: start looking at this one too
- Can start now, but will finish discussion on Tuesday

2

Assign 0: 343rd word

this	8
except	4
let	2
very	2
and	
collaboration	
honesty	
however	
in	
know	
me	
not	
under	

3

Independence

Two variables are independent if they do not affect each other

For two independent variables, knowing the value of one does not change the probability distribution of the other variable

- the result of the toss of a coin is independent of a roll of a dice
- price of tea in England is independent of the whether or not you get an A in NLP

4

Independent or Dependent?

You catching a cold and a butterfly flapping its wings in Africa

Miles per gallon and driving habits

Height and longevity of life

5

Independent variables

How does independence affect our probability equations/properties?

If A and B are independent, written $A \perp B$

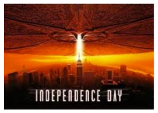
- P(A|B) = P(A)
- P(B|A) = P(B)

What does that mean about P(A,B)?

6

Independent variables

How does independence affect our probability equations/properties?



If A and B are independent, written $A \perp B$

- P(A|B) = P(A)
- P(B|A) = P(B)
- P(A,B) = P(A|B) P(B) = P(A) P(B)
- P(A,B) = P(B|A) P(A) = P(A) P(B)

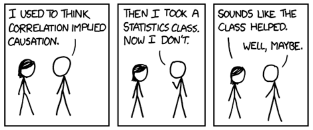
7

Conditional Independence

Dependent events can become independent given certain other events

Examples,

- height and length of life
- "correlation" studies
  - size of your lawn and length of life



http://xkcd.com/552/

8

Conditional Independence

Dependent events can become independent given certain other events

Examples,
- height and length of life
- "correlation" studies
  - size of your lawn and length of life

If A, B are conditionally independent given C $A \perp B \mid C$
- P(A,B | C) = P(A | C) P(B | C)
- P(A | B,C) = P(A | C)
- P(B | A,C) = P(B | C)
- but P(A,B) ≠ P(A)P(B)

9

Assume independence

Sometimes we will assume two variables are independent (or conditionally independent) even though they're not

Why?
- Creates a simpler model
  - p(X,Y) many more variables than just P(X) and P(Y)
- May not be able to estimate the more complicated model

10

Language modeling

What does natural language look like?

More specifically in NLP, probabilistic model

p(sentence)
- p("I like to eat pizza")
- p("pizza like I eat")

Often is posed as: p(word | previous words) – or some other notion of context
- p("pizza" | "I like to eat")
- p("garbage" | "I like to eat")
- p("run" | "I like to eat")

11

Language modeling

How might these models be useful?
- Language generation tasks
  - machine translation
  - summarization
  - simplification
  - speech recognition
  - …
- Text correction
  - spelling correction
  - grammar correction

12

Ideas?

p("I like to eat pizza")

p("pizza like I eat")

p("pizza" | "I like to eat")

p("garbage" | "I like to eat")

p("run" | "I like to eat")

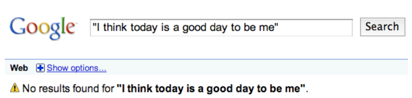
13

Look at a corpus

14

Language modeling

I think today is a good day to be me



Language modeling is about dealing with data sparsity!

15

Probabilistic Language modeling

A probabilistic explanation of how the sentence was generated

Key idea:

- break the generation process into smaller steps
- estimate the probabilities of these smaller steps
- the overall probability is the product of the steps

16

Language modeling

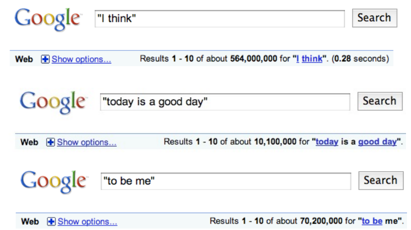
Many approaches:

- n-gram language modeling
  - Start at the beginning of the sentence
  - Generate one word at a time based on the previous words
- syntax-based language modeling
  - Construct the syntactic tree from the top down
  - e.g. context free grammar
  - eventually at the leaves, generate the words
- Neural language models
  - Predict the likelihood of the word based on the context
  - Often allows for generalization beyond the lexical strings

17

n-gram language modeling

I think today is a good day to be me



18

Our friend the chain rule

Step 1: decompose the probability

P(I think today is a good day to be me) =

P(I | <start>) x

P(think | I) x

P(today | I think) x

P(is | I think today) x

P(a | I think today is) x

P(good | I think today is a) x

…

How can we simplify these?

19

The n-gram approximation

Assume each word depends only on the previous n-1 words
(e.g. trigram: three words total)

P(is | I think today) ≈ P(is | think today)

P(a | I think today is) ≈ P(a | today is)

P(good | I think today is a) ≈ P(good | is a)

20

Estimating probabilities

How do we find probabilities?

P(is | think today)

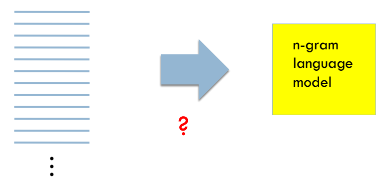
Get real text, and start counting (MLE)!

$$P(\text{is} \mid \text{think today}) = \frac{\text{count(think today is)}}{\text{count(think today)}}$$

21

Estimating from a corpus

Corpus of sentences
(e.g. gigaword corpus)



22

Estimating from a corpus

I am a happy Pomona College student .

count all of the trigrams

<start> <start> I
<start> I am
I am a
am a happy
a happy Pomona
happy Pomona College
Pomona College student
College student .
student . <end>
. <end> <end>

why do we need <start> and <end>?

23

Estimating from a corpus

I am a happy Pomona College student .

count all of the trigrams

<start> <start> I
<start> I am
I am a
am a happy
a happy Pomona
happy Pomona College
Pomona College student
College student .
student . <end>
. <end> <end>

Do we need to count anything else?

24

Estimating from a corpus

I am a happy Pomona College student .

count all of the bigrams

<start> <start>
<start> I
I am
am a
a happy
happy Pomona
Pomona College
College student
student .
. <end>

$$p(c \mid a\ b) = \frac{count(a\ b\ c)}{count(a\ b)}$$

25

Estimating from a corpus

1. Go through all sentences and count trigrams and bigrams
   - usually you store these in some kind of data structure
2. Now, go through all of the trigrams and use the count and the bigram count to calculate MLE probabilities
   - do we need to worry about divide by zero?

$$p(c \mid a\ b) = \frac{count(a\ b\ c)}{count(a\ b)}$$

26

Applying a model

Given a new sentence, we can apply the model

p(Pomona College students are the best .) = ?

p(Pomona | <start> <start>) *
p(College | <start> Pomona) *
p(students | Pomona College) *
⋮
p(<end> | . <end>) *

27

Generating examples

We can also use a trained model to generate a random sentence

Ideas?

<start> <start> ______

We have a distribution over all possible starting words

Draw one from this distribution

p(A | <start> <start>)
p(Apples | <start> <start>)
p(I | <start> <start>)
p(The | <start> <start>)
⋮
p(Zebras | <start> <start>)

28

Generating examples

<start> <start> Zebras _______

repeat!

p(are | <start> Zebras)

p(eat | <start> Zebras)

p(think | <start> Zebras)

p(and| <start> Zebras)

⋮

p(mostly| <start> Zebras)

29

Generation examples

Unigram

are were that ères mammal naturally built describes jazz territory heteromyids film tenor prime live founding must on was feet negro legal gate in on beside . provincial san ; stephenson simply spaces stretched performance double-entry grove replacing station across to burma . repairing ères capital about double reached omnibus el time believed what hotels parameter jurisprudence words syndrome to ères profanity is administrators ères offices hilarius institutionalized remains writer royalty dennis , ères tyson , and objective , instructions seem timekeeper has ères valley ères " magnitudes for love on ères from allakaket , , ana central enlightened . to , ères is belongs fame they the corrected , . on in pressure %NUMBER% her flavored ères derogatory is won metcard indirectly of crop duty learn northbound ères ères dancing similarity ères named ères berkeley . . off-scale overtime . each mansfield stripes dānu traffic ossetic and at alpha popularity town

30

Generation examples

Bigrams

the wikipedia county , mexico .

maurice ravel . it is require that is sparta , where functions . most widely admired .

halogens chamiali cast jason against test site .

31

Generation examples

Trigrams

is widespread in north africa in june %NUMBER% %NUMBER% units were built by with .

jewish video spiritual are considered ircd , this season was an extratropical cyclone .

the british railways ' s strong and a spot .

32

Evaluation

We can train a language model on some data

How can we tell how well we're doing?

- for example
  - bigrams vs. trigrams
  - 100K sentence corpus vs. 100M
  - …

33

Evaluation

A very good option: extrinsic evaluation

If you're going to be using it for machine translation

- build a system with each language model
- compare the two based on their approach for machine translation

Sometimes we don't know the application

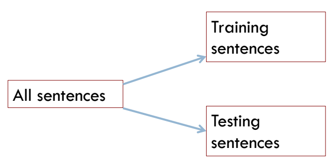
Can be time consuming

Granularity of results

34

Evaluation

Common NLP/machine learning/AI approach



35

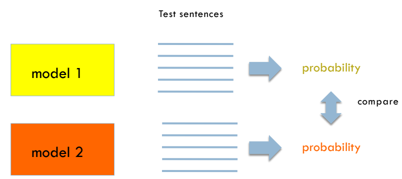
Evaluation

Ideas?

36

Evaluation

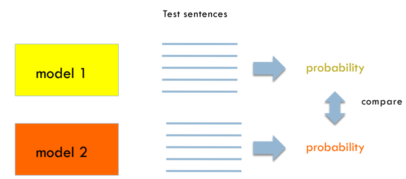
A good model should do a good job of predicting actual sentences



37

Evaluation

Pros: Fine for comparing two models
Cons: Doesn't give us a sense of how well any model is doing



38

The problem

Which of these sentences will have a higher probability based on a language model?

I like to eat banana peels .

I like to eat banana peels with peanut butter.

39

The problem

Which of these sentences will have a higher probability based on a language model?

I like to eat banana peels .

I like to eat banana peels with peanut butter.

Since probabilities are multiplicative (and between 0 and 1), they get smaller for longer sentences.

40

The solution: perplexity*

$$prob(w_{1..n}) = \prod_{i=1}^{n} p(w_i|w_{1..i-1})$$

average the probabilities

geometric mean

$$PP(w_{1..n}) = \sqrt[n]{\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}}$$

41

Calculating perplexity in practice

$$\log\left(\sqrt[n]{\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}}\right) = \log\left(\left(\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}\right)^{1/n}\right)$$

$$= \frac{\log\left(\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}\right)}{n}$$

$$= \frac{-\log\left(\prod_{i=1}^{n} p(w_i|w_{1..i-1})\right)}{n}$$

$$= -\frac{\sum_{i=1}^{n} \log p(w_i|w_{1..i-1})}{n}$$

What is this?

42

Calculating perplexity in practice

$$\log\left(\sqrt[n]{\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}}\right) = \log\left(\left(\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}\right)^{1/n}\right)$$

$$= \frac{\log\left(\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}\right)}{n}$$

$$= \frac{-\log\left(\prod_{i=1}^{n} p(w_i|w_{1..i-1})\right)}{n}$$

Average logprob per word!

$$= -\frac{\sum_{i=1}^{n} \log p(w_i|w_{1..i-1})}{n}$$

43

Calculating **perplexity**

$$PP(w_{1..n}) = \sqrt[n]{\frac{1}{\prod_{i=1}^{n} p(w_i|w_{1..i-1})}}$$

$$= 10^{-\frac{\sum_{i=1}^{n} \log_{10} p(w_i|w_{1..i-1})}{n}}$$

- This is often how it's calculated (and how we'll calculate it)
- Avoid underflow from multiplying too many small probabilities together

44

Another view of perplexity

Weighted average branching factor

- number of possible next words that can follow a word or phrase
- measure of the complexity/uncertainty of text (as viewed from the language models perspective)

45

Smoothing

What if our test set contains the following sentence, but one of the trigrams never occurred in our training data?

P(I think today is a good day to be me) =

P(I | <start> <start>) x

P(think | <start> I) x

P(today| I think) x

P(is| think today) x

P(a| today is) x

P(good| is a) x

...

If any of these has never been seen before, prob = 0!

46

A better approach

p(z | x y) = ?

Suppose our training data includes

... x y a ..
... x y d ...
... x y d ...

but never: xyz

We would conclude

p(a | x y) = 1/3?
p(d | x y) = 2/3?
p(z | x y) = 0/3?

Is this ok?

Intuitively, how should we fix these?

47

Smoothing the estimates

Basic idea:

p(a | x y) = 1/3? *reduce*
p(d | x y) = 2/3? *reduce*
p(z | x y) = 0/3? *increase*

Discount the positive counts somewhat

Reallocate that probability to the zeroes

Remember, it needs to stay a probability distribution

48

Other situations

p(z | x y) = ?

Suppose our training data includes
… x y a … (100 times)
… x y d … (100 times)
… x y d … (100 times)
but never: x y z

Suppose our training data includes
… x y a …
… x y d …
… x y d …
… x y … (300 times)
but never: x y z

Is this the same situation as before?

49

Smoothing the estimates

Should we conclude
p(a | xy) = 1/3? *reduce*
p(d | xy) = 2/3? *reduce*
p(z | xy) = 0/3? *increase*

$$p(c \mid a\ b) = \frac{count(a\ b\ c)}{count(a\ b)}$$

Readjusting the estimate is particularly important if:

- the denominator is small …
  - 1/3 probably too high, 100/300 probably about right
- numerator is small …
  - 1/300 is probably too high, 100/300 probably about right

50

Add-one (Laplacian) smoothing

Consider a LM with a vocabulary of just 26 (a-z)

xya	1	1/3
xyb	0	0/3
xyc	0	0/3
xyd	2	2/3
xye	0	0/3
…		
xyz	0	0/3
Total xy	3	3/3

51

Add-one (Laplacian) smoothing

Consider a LM with a vocabulary of just 26 (a-z)

xya	1	1/3	2	2/29
xyb	0	0/3	1	1/29
xyc	0	0/3	1	1/29
xyd	2	2/3	3	3/29
xye	0	0/3	1	1/29
…				
xyz	0	0/3	1	1/29
Total xy	3	3/3	29	29/29

52

Add-one (Laplacian) smoothing

300 observations instead of 3 – better data, less smoothing

xya	100	100/300	101	101/326
xyb	0	0/300	1	1/326
xyc	0	0/300	1	1/326
xyd	200	200/300	201	201/326
xye	0	0/300	1	1/326
…				
xyz	0	0/300	1	1/326
Total xy	300	300/300	326	326/326

53

Add-one (Laplacian) smoothing

What happens if we're now considering a vocabulary of 20,000 words?

xya	1	1/3	2	2/29
xyb	0	0/3	1	1/29
xyc	0	0/3	1	1/29
xyd	2	2/3	3	3/29
xye	0	0/3	1	1/29
…				
xyz	0	0/3	1	1/29
Total xy	3	3/3	29	29/29

54

Add-one (Laplacian) smoothing

20,000 words, not 26 letters

see the abacus	1	1/3	2	2/20003
see the abbot	0	0/3	1	1/20003
see the abduct	0	0/3	1	1/20003
see the above	2	2/3	3	3/20003
see the Abram	0	0/3	1	1/20003
…				
see the zygote	0	0/3	1	1/20003
Total	3	3/3	20003	20003/20003

Any problem with this?

55

Add-one (Laplacian) smoothing

An "unseen event" is a 0-count event

The probability of an unseen event is 19998/20003

- add one smoothing thinks it is very likely to see a novel event

The problem with add-one smoothing is it gives too much probability mass to unseen events

see the abacus	1	1/3	2	2/20003
see the abbot	0	0/3	1	1/20003
see the abduct	0	0/3	1	1/20003
see the above	2	2/3	3	3/20003
see the Abram	0	0/3	1	1/20003
…				
see the zygote	0	0/3	1	1/20003
Total	3	3/3	20003	20003/20003

56

The general smoothing problem

			modification	probability
see the abacus	1	1/3	?	?
see the abbot	0	0/3	?	?
see the abduct	0	0/3	?	?
see the above	2	2/3	?	?
see the Abram	0	0/3	?	?
…			?	?
see the zygote	0	0/3	?	?
Total	3	3/3	?	?

57

Add-lambda smoothing

A large dictionary makes novel events too probable.

Instead of adding 1 to all counts, add λ = 0.01?

- This gives much less probability to novel events

see the abacus	1	1/3	1.01	1.01/203
see the abbot	0	0/3	0.01	0.01/203
see the abduct	0	0/3	0.01	0.01/203
see the above	2	2/3	2.01	2.01/203
see the Abram	0	0/3	0.01	0.01/203
…			0.01	0.01/203
see the zygote	0	0/3	0.01	0.01/203
Total	3	3/3	203	

58