N-grams

CS114 Lab 6

Kenneth Lai

February 28, 2020

Bayes' Rule

- $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$

Bayes' Rule

- $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- Logistic regression/neural networks: model $P(c|d)$

Bayes' Rule

- $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- Logistic regression/neural networks: model $P(c|d)$
- Naïve Bayes: model $P(c, d) = P(d|c)P(c)$

Bayes' Rule

- $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- Logistic regression/neural networks: model $P(c|d)$
- Naïve Bayes: model $P(c, d) = P(d|c)P(c)$
- Language models: model $P(d)$

Language Modeling

- Suppose we observe a document (sentence) $d =$ "Chinese Chinese Chinese Tokyo Japan". What is the probability of the sentence?

Language Modeling

- Training data:

sentence
Chinese Beijing Chinese
Chinese Chinese Shanghai
Chinese Macao
Tokyo Japan Chinese

Language Modeling

- Sentences are sequences of words

Language Modeling

- Sentences are sequences of words
  - No bag of words assumption this time: position matters

Language Modeling

- Sentences are sequences of words
  - No bag of words assumption this time: position matters
- $d =$ "Chinese Chinese Chinese Tokyo Japan"

Language Modeling

- Sentences are sequences of words
  - No bag of words assumption this time: position matters
- $d =$ "Chinese Chinese Chinese Tokyo Japan"
  - $w_1 =$ Chinese
  - $w_2 =$ Chinese
  - $w_3 =$ Chinese
  - $w_4 =$ Tokyo
  - $w_5 =$ Japan

Language Modeling

- Chain Rule: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P\left(w_i \middle| \bigcap_{j=1}^{i-1} w_j\right)$

Independence Assumption

- Markov Assumption: the probability of a word depends only on the previous word(s)

Independence Assumption

- Markov Assumption: the probability of a word depends only on the previous word(s)
  - Zeroth-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i)$

Independence Assumption

- Markov Assumption: the probability of a word depends only on the previous word(s)
  - Zeroth-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i)$
  - First-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i|w_{i-1})$

Independence Assumption

- Markov Assumption: the probability of a word depends only on the previous word(s)
  - Zeroth-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i)$
  - First-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i|w_{i-1})$
  - Second-order: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1})$

Independence Assumption

- Markov Assumption: the probability of a word depends only on the previous word(s)
  - Unigram: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i)$
  - Bigram: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i|w_{i-1})$
  - Trigram: $P\left(\bigcap_{i=1}^{n} w_i\right) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1})$

Training Language Models

- Everything is counting (again)!

Training Language Models

- Everything is counting (again)!
- $\hat{P}(w_i) = \dfrac{\text{count}(w_i)}{\displaystyle\sum_{w \in V} \text{count}(w)}$

Training Language Models

- Everything is counting (again)!
- $\hat{P}(w_i) = \dfrac{\text{count}(w_i)}{\sum\limits_{w \in V} \text{count}(w)}$
- $\hat{P}(w_i | w_{i-1}) = \dfrac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$
- ...

Training Language Models

sentence
Chinese Beijing Chinese
Chinese Chinese Shanghai
Chinese Macao
Tokyo Japan Chinese

Training Language Models

sentence
Chinese Beijing Chinese </S>
Chinese Chinese Shanghai </S>
Chinese Macao </S>
Tokyo Japan Chinese </S>

Training Language Models

sentence
Chinese Beijing Chinese </S>
Chinese Chinese Shanghai </S>
Chinese Macao </S>
Tokyo Japan Chinese </S>

- Stop symbol </S>

Training Language Models

sentence
Chinese Beijing Chinese </S>
Chinese Chinese Shanghai </S>
Chinese Macao </S>
Tokyo Japan Chinese </S>

- Stop symbol </S>
  - Why?

Training Language Models

sentence
Chinese Beijing Chinese </S>
Chinese Chinese Shanghai </S>
Chinese Macao </S>
Tokyo Japan Chinese </S>

- Stop symbol </S>
  - Why?
    - (see Jurafsky and Martin exercise 3.5)

Unknown Words

- Words that do not appear in the training data

Unknown Words

- Words that do not appear in the training data
- Suppose we observe a sentence $D' =$ "Japanese Kyoto". What is the probability of the sentence?

Unknown Words

- Words that do not appear in the training data
- Suppose we observe a sentence $D' =$ "Japanese Kyoto". What is the probability of the sentence?
  - $\text{count}(\text{Japanese}) = \text{count}(\text{Kyoto}) = 0$

Unknown Words

- Words that do not appear in the training data
- Suppose we observe a sentence $D' =$ "Japanese Kyoto". What is the probability of the sentence?
  - $\text{count}(\text{Japanese}) = \text{count}(\text{Kyoto}) = 0$
  - $P(\text{Kyoto}|\text{Japanese}) = 0/0$

Unknown Words

- Words that do not appear in the training data
- Suppose we observe a sentence $D' =$ "Japanese Kyoto". What is the probability of the sentence?
  - $\text{count}(\text{Japanese}) = \text{count}(\text{Kyoto}) = 0$
  - $P(\text{Kyoto}|\text{Japanese}) = 0/0$
- Solution: unknown word `<UNK>`

Unknown Words

- Words that do not appear in the training data
- Suppose we observe a sentence $D' =$"Japanese Kyoto". What is the probability of the sentence?
  - $\text{count}(\text{Japanese}) = \text{count}(\text{Kyoto}) = 0$
  - $P(\text{Kyoto}|\text{Japanese}) = 0/0$
- Solution: unknown word `<UNK>`
  - For HW/PA, you can simply set all the `<UNK>`-related counts equal to 1

Unigram Language Models

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
count(w_i)	6	1	1	1	1	1	4	1

Unigram Language Models

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	6/16	1/16	1/16	1/16	1/16	1/16	4/16	1/16

- $\sum_{w \in V} \text{count}(w) = 16$

Unigram Language Models

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	3/8	1/16	1/16	1/16	1/16	1/16	1/4	1/16

- $\sum_{w \in V} \text{count}(w) = 16$
- $P(\text{Chinese Chinese Chinese Tokyo Japan}) = (3/8)^3 \times 1/16 \times 1/16 \times 1/4 = 27/524288 \approx 0.00005$

Unigram Language Models

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	3/8	1/16	1/16	1/16	1/16	1/16	1/4	1/16

- $\sum_{w \in V} \text{count}(w) = 16$
- $P(\text{Chinese Chinese Chinese Tokyo Japan}) = (3/8)^3 \times 1/16 \times 1/16 \times 1/4 = 27/524288 \approx 0.00005$
- $P(\text{Japanese Kyoto}) = 1/16 \times 1/16 \times 1/4 = 1/1024 \approx 0.001$

Bigram Language Models

sentence
Chinese Beijing Chinese </S>
Chinese Chinese Shanghai </S>
Chinese Macao </S>
Tokyo Japan Chinese </S>

Bigram Language Models

sentence
<S> Chinese Beijing Chinese </S>
<S> Chinese Chinese Shanghai </S>
<S> Chinese Macao </S>
<S> Tokyo Japan Chinese </S>

Bigram Language Models

sentence
<S> Chinese Beijing Chinese </S>
<S> Chinese Chinese Shanghai </S>
<S> Chinese Macao </S>
<S> Tokyo Japan Chinese </S>

- Start symbol <S>

Bigram Language Models

sentence
<S> Chinese Beijing Chinese </S>
<S> Chinese Chinese Shanghai </S>
<S> Chinese Macao </S>
<S> Tokyo Japan Chinese </S>

- Start symbol <S>
  - Why?

Bigram Language Models

sentence
<S> Chinese Beijing Chinese </S>
<S> Chinese Chinese Shanghai </S>
<S> Chinese Macao </S>
<S> Tokyo Japan Chinese </S>

- Start symbol <S>
  - Why?
    - Beginning of the sentence is a context, too!

Bigram Language Models

count(w_{i-1}, w_i)		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3	0	0	0	1	0	0	1
	Chinese	1	1	1	1	0	0	2	1
	Beijing	1	0	0	0	0	0	0	1
	Shanghai	0	0	0	0	0	0	1	1
	Macao	0	0	0	0	0	0	1	1
	Tokyo	0	0	0	0	0	1	0	1
	Japan	1	0	0	0	0	0	0	1
	<UNK>	1	1	1	1	1	1	1	1

Bigram Language Models

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Bigram Language Models

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- $P(\text{Chinese Chinese Chinese Tokyo Japan}) =$
  $3/5 \times (1/7)^2 \times 0 \times 1/2 \times 0 =$
  0

Bigram Language Models

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- $P(\text{Japanese Kyoto}) =$
  $1/5 \times 1/8 \times 1/8 =$
  $1/320 \approx 0.003$

Smoothing

- Words that appear in the training data, but not in a given context

Smoothing

- Words that appear in the training data, but not in a given context
- Laplace (add-1) smoothing: add 1 to all word counts

Smoothing

- Words that appear in the training data, but not in a given context
- Laplace (add-1) smoothing: add 1 to all word counts
  - $\hat{P}(w_i|w_{i-1}) = \dfrac{\text{count}(w_{i-1}, w_i) + 1}{\text{count}(w_i) + |V|}$

Smoothing

- Words that appear in the training data, but not in a given context
- Laplace (add-1) smoothing: add 1 to all word counts
  - $\hat{P}(w_i|w_{i-1}) = \dfrac{\text{count}(w_{i-1}, w_i) + 1}{\text{count}(w_i) + |V|}$
- Add-k smoothing: add k to all word counts

Smoothing

- Words that appear in the training data, but not in a given context
- Laplace (add-1) smoothing: add 1 to all word counts
  - $\hat{P}(w_i|w_{i-1}) = \dfrac{\text{count}(w_{i-1}, w_i) + 1}{\text{count}(w_i) + |V|}$
- Add-k smoothing: add k to all word counts
  - $\hat{P}(w_i|w_{i-1}) = \dfrac{\text{count}(w_{i-1}, w_i) + k}{\text{count}(w_i) + k|V|}$

Smoothing

$\text{count}(w_{i-1}, w_i)$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3	0	0	0	1	0	0	1
	Chinese	1	1	1	1	0	0	2	1
	Beijing	1	0	0	0	0	0	0	1
	Shanghai	0	0	0	0	0	0	1	1
	Macao	0	0	0	0	0	0	1	1
	Tokyo	0	0	0	0	0	1	0	1
	Japan	1	0	0	0	0	0	0	1
	<UNK>	1	1	1	1	1	1	1	1

Smoothing

$\mathrm{count}(w_{i-1}, w_i) + 1$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	4	1	1	1	2	1	1	2
	Chinese	2	2	2	2	1	1	3	2
	Beijing	2	1	1	1	1	1	1	2
	Shanghai	1	1	1	1	1	1	2	2
	Macao	1	1	1	1	1	1	2	2
	Tokyo	1	1	1	1	1	2	1	2
	Japan	2	1	1	1	1	1	1	2
	<UNK>	2	2	2	2	2	2	2	2

Smoothing

$\hat{P}(w_i \| w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	4/13	1/13	1/13	1/13	2/13	1/13	1/13	2/13
	Chinese	2/15	2/15	2/15	2/15	1/15	1/15	3/15	2/15
	Beijing	2/10	1/10	1/10	1/10	1/10	1/10	1/10	2/10
	Shanghai	1/10	1/10	1/10	1/10	1/10	1/10	2/10	2/10
	Macao	1/10	1/10	1/10	1/10	1/10	1/10	2/10	2/10
	Tokyo	1/10	1/10	1/10	1/10	1/10	2/10	1/10	2/10
	Japan	2/10	1/10	1/10	1/10	1/10	1/10	1/10	2/10
	<UNK>	2/16	2/16	2/16	2/16	2/16	2/16	2/16	2/16

Smoothing

$\hat{P}(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	4/13	1/13	1/13	1/13	2/13	1/13	1/13	2/13
	Chinese	2/15	2/15	2/15	2/15	1/15	1/15	1/5	2/15
	Beijing	1/5	1/10	1/10	1/10	1/10	1/10	1/10	1/5
	Shanghai	1/10	1/10	1/10	1/10	1/10	1/10	1/5	1/5
	Macao	1/10	1/10	1/10	1/10	1/10	1/10	1/5	1/5
	Tokyo	1/10	1/10	1/10	1/10	1/10	1/5	1/10	1/5
	Japan	1/5	1/10	1/10	1/10	1/10	1/10	1/10	1/5
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- P(Chinese Chinese Chinese Tokyo Japan) $=$
  $4/13 \times (2/15)^2 \times 1/15 \times 1/5 \times 1/10 =$
  $8/1096875 \approx 0.000007$

Smoothing

- Another solution to zeros: use less context

Smoothing

- Another solution to zeros: use less context
- Backoff: if the n-gram is not available, back off to the (n-1)-gram

Smoothing

- Another solution to zeros: use less context
- Backoff: if the n-gram is not available, back off to the (n-1)-gram
- Interpolation: weighted average of n-gram, (n-1)-gram, etc.

Smoothing

- Another solution to zeros: use less context
- Backoff: if the n-gram is not available, back off to the (n-1)-gram
- Interpolation: weighted average of n-gram, (n-1)-gram, etc.
  - $\hat{P}(w_i|w_{i-1}) = \lambda_1 P(w_i|w_{i-1}) + \lambda_2 P(w_i)$
    - $\sum_i \lambda_i = 1$

Interpolation

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	3/8	1/16	1/16	1/16	1/16	1/16	1/4	1/16

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Interpolation

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	3/8	1/16	1/16	1/16	1/16	1/16	1/4	1/16

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- Let $\lambda_1 = \lambda_2 = 1/2$

Interpolation

w_i	Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
$P(w_i)$	3/8	1/16	1/16	1/16	1/16	1/16	1/4	1/16

$P(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	3/5	0	0	0	1/5	0	0	1/5
	Chinese	1/7	1/7	1/7	1/7	0	0	2/7	1/7
	Beijing	1/2	0	0	0	0	0	0	1/2
	Shanghai	0	0	0	0	0	0	1/2	1/2
	Macao	0	0	0	0	0	0	1/2	1/2
	Tokyo	0	0	0	0	0	1/2	0	1/2
	Japan	1/2	0	0	0	0	0	0	1/2
	<UNK>	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- Let $\lambda_1 = \lambda_2 = 1/2$
  - $\hat{P}(\text{Chinese}|\text{<S>}) = \frac{1}{2}P(\text{Chinese}|\text{<S>}) + \frac{1}{2}P(\text{Chinese})$

$$= \frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{3}{8}$$

$$= \frac{39}{80} = 0.4875$$

Interpolation

$\hat{P}(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	.4875	.03125	.03125	.03125	.03125	.13125	.125	.13125
	Chinese	.25893	.10268	.10268	.10268	.03125	.03125	.26786	.10268
	Beijing	.4375	.03125	.03125	.03125	.03125	.03125	.125	.28125
	Shanghai	.1875	.03125	.03125	.03125	.03125	.03125	.375	.28125
	Macao	.1875	.03125	.03125	.03125	.03125	.03125	.375	.28125
	Tokyo	.1875	.03125	.03125	.03125	.03125	.28125	.125	.28125
	Japan	.4375	.03125	.03125	.03125	.03125	.03125	.125	.28125
	<UNK>	.25	.09375	.09375	.09375	.09375	.09375	.1875	.09375

Interpolation

$\hat{P}(w_i \mid w_{i-1})$		w_i							
		Chinese	Beijing	Shanghai	Macao	Tokyo	Japan	</S>	<UNK>
w_{i-1}	<S>	.4875	.03125	.03125	.03125	.03125	.13125	.125	.13125
	Chinese	.25893	.10268	.10268	.10268	.03125	.03125	.26786	.10268
	Beijing	.4375	.03125	.03125	.03125	.03125	.03125	.125	.28125
	Shanghai	.1875	.03125	.03125	.03125	.03125	.03125	.375	.28125
	Macao	.1875	.03125	.03125	.03125	.03125	.03125	.375	.28125
	Tokyo	.1875	.03125	.03125	.03125	.03125	.28125	.125	.28125
	Japan	.4375	.03125	.03125	.03125	.03125	.03125	.125	.28125
	<UNK>	.25	.09375	.09375	.09375	.09375	.09375	.1875	.09375

- $P(\text{Chinese Chinese Chinese Tokyo Japan}) =$
  $0.4875 \times (0.25893)^2 \times 0.03125 \times 0.28125 \times 0.125 =$
  ≈ 0.000036