Review: Naïve Bayes

CS114B Lab 1

Kenneth Lai

February 5, 2021

Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?

Naïve Bayes

- Training data:

document	class
just plain boring	negative
entirely predictable and lacks energy	negative
no surprises and very few laughs	negative
very powerful	positive
the most fun film of the summer	positive

Naïve Bayes

- Naïve Bayes models are generative

Naïve Bayes

- Naïve Bayes models are generative
  - Assume the data are generated according to an underlying distribution

Multinomial Naïve Bayes

- Documents are bags of words

Multinomial Naïve Bayes

- Documents are bags of words
- Data generated by multinomial distribution

Multinomial Naïve Bayes

- Documents are bags of words
- Data generated by multinomial distribution
  - "rolling a $|V|$-sided die n times"

Multinomial Naïve Bayes

- Documents are bags of words
- Data generated by multinomial distribution
  - "rolling a $|V|$-sided die n times"
  - $V =$ vocabulary, $n =$ length of document

Multinomial Naïve Bayes

- $c =$ negative
- $d =$"predictable with no fun"

Multinomial Naïve Bayes

- $c =$ negative
- $d =$ "predictable with no fun"
  - $w_1 =$ predictable
  - $w_2 =$ with
  - $w_3 =$ no
  - $w_4 =$ fun

Multinomial Naïve Bayes

- Bayes' Rule: $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$

Multinomial Naïve Bayes

- Bayes' Rule: $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- $\hat{c} = \operatorname*{argmax}_{c \in C} P(d|c)P(c)$

Multinomial Naïve Bayes

- Bayes' Rule: $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c)$
- What about $P(d)$?

Multinomial Naïve Bayes

- Bayes' Rule: $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- $\hat{c} = \operatorname*{argmax}_{c \in C} P(d|c)P(c)$
- What about $P(d)$?
  - $P(d)$ is the same for each class

Multinomial Naïve Bayes

- $\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c)P(d|c)$

Multinomial Naïve Bayes

- $$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c)P(d|c)$$
  $$= \underset{c \in C}{\operatorname{argmax}} P(c)P(w_1, ..., w_n|c)$$

Multinomial Naïve Bayes

- $$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, P(c)P(d|c)$$
  $$= \underset{c \in C}{\operatorname{argmax}}\, P(c)P(w_1, ..., w_n|c)$$
- Chain Rule: $\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, P(c) \prod_{i=1}^{n} P\left(w_i \middle| \bigcap_{j=1}^{i-1} w_j, c\right)$

Independence Assumptions

- Bag of Words Assumption: position doesn't matter

Independence Assumptions

- Bag of Words Assumption: position doesn't matter
- Naïve Bayes Assumption: features (words) are independent given the class

Independence Assumptions

- Bag of Words Assumption: position doesn't matter
- Naïve Bayes Assumption: features (words) are independent given the class
  - $$\prod_{i=1}^{n} P\left(w_i \middle| \bigcap_{j=1}^{i-1} w_j, c\right) = \prod_{i=1}^{n} P(w_i|c)$$

Training Naïve Bayes

- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$

Training Naïve Bayes

- $c_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i=1}^{n} P(w_i|c)$
- Everything is counting!

Training Naïve Bayes

- $c_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i=1}^{n} P(w_i|c)$
- Everything is counting!
- $\hat{P}(c) = \dfrac{\text{doccount}(c)}{\sum_{c' \in C} \text{doccount}(c')}$

Training Naïve Bayes

- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
- Everything is counting!
- $\hat{P}(c) = \dfrac{\text{doccount}(c)}{\sum_{c' \in C} \text{doccount}(c')}$
- $\hat{P}(w_i|c) = \dfrac{\text{wordcount}(w_i, c)}{\sum_{w \in V} \text{wordcount}(w, c)}$

Training Naïve Bayes

- What if w_i does not appear in any documents of class c?

Training Naïve Bayes

- What if w_i does not appear in any documents of class c?
  - $\hat{P}(w_i|c) = 0$

Training Naïve Bayes

- What if w_i does not appear in any documents of class c?
  - $\hat{P}(w_i|c) = 0$
- Suppose we observe a movie review $D' =$ "just fun". Is the review positive or negative?

Training Naïve Bayes

- What if w_i does not appear in any documents of class c?
  - $\hat{P}(w_i|c) = 0$
- Suppose we observe a movie review $D' =$"just fun". Is the review positive or negative?
  - $\text{argmax}(0, 0) =?$

Smoothing

- Laplace (add-1) smoothing: add 1 to all word counts

Smoothing

- Laplace (add-1) smoothing: add 1 to all word counts
- $\hat{P}(w_i|c) = \dfrac{\text{wordcount}(w_i, c) + 1}{\sum\limits_{w \in V} (\text{wordcount}(w, c) + 1)}$

$$= \frac{\text{wordcount}(w_i, c) + 1}{\left(\sum\limits_{w \in V} \text{wordcount}(w, c) \right) + |V|}$$

Training Naïve Bayes

- $\hat{P}(c) = \dfrac{\text{doccount}(c)}{\sum_{c' \in C} \text{doccount}(c')}$
- $\hat{P}(w_i|c) = \dfrac{\text{wordcount}(w_i, c) + 1}{\left(\sum_{w \in V} \text{wordcount}(w, c)\right) + |V|}$

Training Naïve Bayes

- $\hat{P}(c) = \dfrac{\text{doccount}(c)}{\displaystyle\sum_{c' \in C} \text{doccount}(c')}$
- $\hat{P}(w_i|c) = \dfrac{\text{wordcount}(w_i, c) + 1}{\left(\displaystyle\sum_{w \in V} \text{wordcount}(w, c)\right) + |V|}$

document	class
just plain boring	negative
entirely predictable and lacks energy	negative
no surprises and very few laughs	negative
very powerful	positive
the most fun film of the summer	positive

Training Naïve Bayes

- $\hat{P}(c) = \dfrac{\text{doccount}(c)}{\sum_{c' \in C} \text{doccount}(c')}$
- $\hat{P}(w_i|c) = \dfrac{\text{wordcount}(w_i, c) + 1}{\left(\sum_{w \in V} \text{wordcount}(w, c)\right) + |V|}$

document	class
just plain boring	negative
entirely predictable and lacks energy	negative
no surprises and very few laughs	negative
very powerful	positive
the most fun film of the summer	positive

- $\hat{P}(\text{negative}) = 3/5$
- $\hat{P}(\text{positive}) = 2/5$

Training Naïve Bayes

$\text{wordcount}(w, c)$		w			
		predictable	no	fun	...
c	negative	1	1	0	...
	positive	0	0	1	...

Training Naïve Bayes

-

wordcount(w, c) + 1		w			
		predictable	no	fun	...
c	negative	$1+1$	$1+1$	$0+1$	...
	positive	$0+1$	$0+1$	$1+1$	...

Training Naïve Bayes

$\text{wordcount}(w, c) + 1$		w			
		predictable	no	fun	...
c	negative	2	2	1	...
	positive	1	1	2	...

Training Naïve Bayes

$\text{wordcount}(w, c) + 1$		w			
		predictable	no	fun	...
c	negative	2	2	1	...
	positive	1	1	2	...

- $\sum_{w \in V} \text{wordcount}(w, \text{negative}) = 14$
- $\sum_{w \in V} \text{wordcount}(w, \text{positive}) = 9$

Training Naïve Bayes

wordcount$(w, c) + 1$		w			
		predictable	no	fun	...
c	negative	2	2	1	...
	positive	1	1	2	...

- $\sum_{w \in V} \text{wordcount}(w, \text{negative}) = 14$
- $\sum_{w \in V} \text{wordcount}(w, \text{positive}) = 9$
- $|V| = 20$

Training Naïve Bayes

$\hat{P}(w\|c)$		w			
		predictable	no	fun	...
c	negative	$2/(14+20)$	$2/(14+20)$	$1/(14+20)$	...
	positive	$1/(9+20)$	$1/(9+20)$	$2/(9+20)$	...

- $\sum_{w \in V} \text{wordcount}(w, \text{negative}) = 14$
- $\sum_{w \in V} \text{wordcount}(w, \text{positive}) = 9$
- $|V| = 20$

Training Naïve Bayes

$\hat{P}(w\|c)$		w			
		predictable	no	fun	...
c	negative	1/17	1/17	1/34	...
	positive	1/29	1/29	2/29	...

- $\sum_{w \in V} \text{wordcount}(w, \text{negative}) = 14$
- $\sum_{w \in V} \text{wordcount}(w, \text{positive}) = 9$
- $|V| = 20$

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i | c)$

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
  - Ignore unknown word "with"

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
  - Ignore unknown word "with"
- $P(\text{negative}|d) \propto 3/5 \times (1/17)^2 \times 1/34 \approx 6.1 \times 10^{-5}$

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?
- $c_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i=1}^{n} P(w_i|c)$
  - Ignore unknown word "with"
- $P(\text{negative}|d) \propto 3/5 \times (1/17)^2 \times 1/34 \approx 6.1 \times 10^{-5}$
- $P(\text{positive}|d) \propto 2/5 \times (1/29)^2 \times 2/29 \approx 3.2 \times 10^{-5}$

Testing Naïve Bayes

- Suppose we observe a movie review $d =$ "predictable with no fun". Is the review positive or negative?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
  - Ignore unknown word "with"
- $P(\text{negative}|d) \propto 3/5 \times (1/17)^2 \times 1/34 \approx 6.1 \times 10^{-5}$
- $P(\text{positive}|d) \propto 2/5 \times (1/29)^2 \times 2/29 \approx 3.2 \times 10^{-5}$
- negative

Words and Features

- Not all features are (necessarily) words

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features
  - Ignore unknown words

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features
  - Ignore unknown words
  - Ignore non-feature words

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features
  - Ignore unknown words
  - Ignore non-feature words
    - $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features
  - Ignore unknown words
  - Ignore non-feature words
    - $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$
  - Importantly, V should still be the entire vocabulary

Words and Features

- Not all features are (necessarily) words
  - Character n-grams, specific phrases, non-linguistic features, etc.
- Not all words are (necessarily) features
  - Ignore unknown words
  - Ignore non-feature words
    - $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$
  - Importantly, V should still be the entire vocabulary
    - The other words are still there, even if we are not using them

Working with Logs

- If $x \times y = z$, then $\log(x) + \log(y) = \log(z)$

Working with Logs

- If $x \times y = z$, then $\log(x) + \log(y) = \log(z)$
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log(P(c)) + \sum_{i=1}^{n} \log(P(w_i|c))$

Working with Logs

- If $x \times y = z$, then $\log(x) + \log(y) = \log(z)$
- $c_{NB} = \operatorname*{argmax}_{c \in C} \log(P(c)) + \sum_{i=1}^{n} \log(P(w_i|c))$
- Avoid floating-point underflow

Working with Logs

- If $x \times y = z$, then $\log(x) + \log(y) = \log(z)$
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log(P(c)) + \sum_{i=1}^{n} \log(P(w_i|c))$
- Avoid floating-point underflow
  - (You will need to do this for PA, but not for HW)