Probability and Naïve Bayes

CS114 Lab 3

Kenneth Lai

January 31, 2020

Probability Mass Function

- Wikipedia: a probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.

Probability Mass Function

- Wikipedia: a probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value.
  - In computational linguistics, most of what we are interested in can be represented by random variables.

Probability Mass Function

- $X =$ topic

x	$P(x)$
Chinese	0.75
Japanese	0.25

Probability Mass Function

- $X =$ part of speech

x	$P(x)$
verb	0.4
noun	0.3
preposition	0.1
adjective	0.1
adverb	0.1

Probability Mass Function

- Properties of probability distributions:

Probability Mass Function

- Properties of probability distributions:
  - All of the possible values must be mutually exclusive

Probability Mass Function

- Properties of probability distributions:
  - All of the possible values must be mutually exclusive
  - $\forall x \in X . 0 \leq P(x) \leq 1$ (The probability values must be between 0 and 1)

Probability Mass Function

- Properties of probability distributions:
  - All of the possible values must be mutually exclusive
  - $\forall x \in X . 0 \leq P(x) \leq 1$ (The probability values must be between 0 and 1)
  - $\sum_{x \in X} P(x) = 1$ (The probability values must sum to 1)

Joint Probability

- X = part of speech
  Y = capitalization

$P(x, y)$		y	
		Yes	no
x	verb	0.1	0.3
	noun	0.25	0.05
	preposition	0.04	0.06
	adjective	0.03	0.07
	adverb	0.04	0.06

Joint Probability

- Same properties, just with multiple random variables:

Joint Probability

- Same properties, just with multiple random variables:
  - All of the possible values must be mutually exclusive
  - $\forall x \in X . \forall y \in Y . 0 \leq P(x, y) \leq 1$ (The probability values must be between 0 and 1)
  - $\sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$ (The probability values must sum to 1)

Marginal Probability

- To extract $P(x)$ or $P(y)$ from $P(x, y)$, sum up in the margins:

$P(x, y)$		y	
		Yes	no
x	verb	0.1	0.3
	noun	0.25	0.05
	preposition	0.04	0.06
	adjective	0.03	0.07
	adverb	0.04	0.06

Marginal Probability

- To extract $P(x)$ or $P(y)$ from $P(x, y)$, sum up in the margins:

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1

Marginal Probability

- To extract $P(x)$ or $P(y)$ from $P(x, y)$, sum up in the margins:

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

Marginal Probability

- To extract $P(x)$ or $P(y)$ from $P(x, y)$, sum up in the margins:

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

- $P(x) = \sum_{y \in Y} P(x, y)$

$$P(y) = \sum_{x \in X} P(x, y)$$

Conditional Probability

- Suppose we observe that a word is capitalized. What is the probability that the word is a verb?

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

Conditional Probability

- Suppose we observe that a word is capitalized. What is the probability that the word is a verb?

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

- $P(x|y) = \frac{P(x, y)}{P(y)}$

Conditional Probability

- Suppose we observe that a word is capitalized. What is the probability that the word is a verb?

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

- $P(\text{verb}|\text{Yes}) = \dfrac{P(\text{verb, Yes})}{P(\text{Yes})} = \dfrac{0.1}{0.46} \approx 0.22$

Conditional Probability

$P(x\|y)$		y	
		Yes	no
x	verb	0.22	0.56
	noun	0.54	0.09
	preposition	0.09	0.11
	adjective	0.07	0.13
	adverb	0.09	0.11
- $P(x|y) = \frac{P(x,y)}{P(y)}$

Conditional Probability

$P(y\|x)$		y	
		Yes	no
x	verb	0.25	0.75
	noun	0.83	0.17
	preposition	0.4	0.6
	adjective	0.3	0.7
	adverb	0.4	0.6
- $P(y|x) = \frac{P(x,y)}{P(x)}$

Bayes' Rule

- $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

Bayes' Rule

- $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
- $P(X,Y) = P(X|Y)P(Y)$

Bayes' Rule

- $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
- $P(X,Y) = P(X|Y)P(Y)$
- $P(Y|X) = \frac{P(X,Y)}{P(X)}$

Bayes' Rule

- $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
- $P(X,Y) = P(X|Y)P(Y)$
- $P(Y|X) = \frac{P(X,Y)}{P(X)}$

- $\boxed{P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}}$

Bayes' Rule

- $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$
- $P(X,Y) = P(X|Y)P(Y)$
- $P(Y|X) = \dfrac{P(X,Y)}{P(X)}$

- $\boxed{P(Y|X) = \dfrac{P(X|Y)P(Y)}{P(X)}}$
- $\text{posterior} = \dfrac{\text{likelihood} \times \text{prior}}{\text{evidence}}$

Independence

- If two random variables are independent, it means that knowing one does not tell us anything about the other (they are not correlated).

Independence

- If two random variables are independent, it means that knowing one does not tell us anything about the other (they are not correlated).
  - $P(X|Y) = P(X)$

Independence

- If two random variables are independent, it means that knowing one does not tell us anything about the other (they are not correlated).
  - $P(X|Y) = P(X)$
  - It does not matter whether we know the value of Y or not, the probability distribution of X is the same.

Independence

- Are part of speech and capitalization independent?

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

- $P(\text{verb}|\text{Yes}) = \dfrac{P(\text{verb, Yes})}{P(\text{Yes})} = \dfrac{0.1}{0.46} \approx 0.22$

Independence

- Are part of speech and capitalization independent?

$P(x, y)$		y		$P(x)$
		Yes	no	
x	verb	0.1	0.3	0.4
	noun	0.25	0.05	0.3
	preposition	0.04	0.06	0.1
	adjective	0.03	0.07	0.1
	adverb	0.04	0.06	0.1
$P(y)$		0.46	0.54	

- $P(\text{verb}|\text{Yes}) = \dfrac{P(\text{verb, Yes})}{P(\text{Yes})} = \dfrac{0.1}{0.46} \approx 0.22$
- No

Chain Rule

- Repeated application of definition of conditional probability

Chain Rule

- Repeated application of definition of conditional probability
  - $P(X_2, X_1) = P(X_2|X_1)P(X_1)$

Chain Rule

- Repeated application of definition of conditional probability
  - $P(X_2, X_1) = P(X_2|X_1)P(X_1)$
  - $P(X_3, X_2, X_1) = P(X_3|X_2, X_1)P(X_2, X_1)$

Chain Rule

- Repeated application of definition of conditional probability
  - $P(X_2, X_1) = P(X_2|X_1)P(X_1)$
  - $P(X_3, X_2, X_1) = P(X_3|X_2, X_1)P(X_2, X_1)$
    $= P(X_3|X_2, X_1)P(X_2|X_1)P(X_1)$

Chain Rule

- Repeated application of definition of conditional probability
  - $P(X_2, X_1) = P(X_2|X_1)P(X_1)$
  - $P(X_3, X_2, X_1) = P(X_3|X_2, X_1)P(X_2, X_1)$
    $= P(X_3|X_2, X_1)P(X_2|X_1)P(X_1)$
  - $P\left(\bigcap_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} P\left(X_i \middle| \bigcap_{j=1}^{i-1} X_j\right)$

Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?

Naïve Bayes

- Training data:

document	class
Chinese Beijing Chinese	Chinese
Chinese Chinese Shanghai	Chinese
Chinese Macao	Chinese
Tokyo Japan Chinese	Japanese

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

- $c =$ Chinese
- $d =$ "Chinese Chinese Chinese Tokyo Japan"

Multinomial Naïve Bayes

- $c =$ Chinese
- $d =$ "Chinese Chinese Chinese Tokyo Japan"
  - $w_1 =$ Chinese
  - $w_2 =$ Chinese
  - $w_3 =$ Chinese
  - $w_4 =$ Tokyo
  - $w_5 =$ Japan

Multinomial Naïve Bayes

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

Multinomial Naïve Bayes

- Bayes' Rule: $P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$
- $\hat{c} = \underset{c \in C}{\text{argmax}}\, P(d|c)P(c)$

Multinomial Naïve Bayes

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

Multinomial Naïve Bayes

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

Multinomial Naïve Bayes

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

Multinomial Naïve Bayes

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

Multinomial Naïve Bayes

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

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

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

Training Naïve Bayes

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

Training Naïve Bayes

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

Training Naïve Bayes

- $c_{NB} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1}^{n} P(w_i|c)$
- 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 document D' = "Beijing Tokyo". Is the document Chinese or Japanese?

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 document D' = "Beijing Tokyo". Is the document Chinese or Japanese?
  - 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
- $$\hat{P}(w_i|c) = \frac{\text{wordcount}(w_i, c) + 1}{\sum_{w \in V} (\text{wordcount}(w, c) + 1)}$$
$$= \frac{\text{wordcount}(w_i, c) + 1}{\left(\sum_{w \in V} \text{wordcount}(w, c) \right) + |V|}$$

Training Naïve Bayes

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

Training Naïve Bayes

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

document	class
Chinese Beijing Chinese	Chinese
Chinese Chinese Shanghai	Chinese
Chinese Macao	Chinese
Tokyo Japan Chinese	Japanese

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
Chinese Beijing Chinese	Chinese
Chinese Chinese Shanghai	Chinese
Chinese Macao	Chinese
Tokyo Japan Chinese	Japanese

- $\hat{P}(\text{Chinese}) = 3/4$
- $\hat{P}(\text{Japanese}) = 1/4$

Training Naïve Bayes

wordcount(w, c)		c	
		Chinese	Japanese
w	Chinese	5	1
	Tokyo	0	1
	Japan	0	1
	...	...	...

Training Naïve Bayes

wordcount$(w, c) + 1$		c	
		Chinese	Japanese
w	Chinese	$5 + 1$	$1 + 1$
	Tokyo	$0 + 1$	$1 + 1$
	Japan	$0 + 1$	$1 + 1$
	...	...	...

Training Naïve Bayes

wordcount$(w, c) + 1$		c	
		Chinese	Japanese
w	Chinese	$5 + 1$	$1 + 1$
	Tokyo	$0 + 1$	$1 + 1$
	Japan	$0 + 1$	$1 + 1$
	...	...	...

- $\sum_{w \in V} \text{wordcount}(w, \text{Chinese}) = 8$
- $\sum_{w \in V} \text{wordcount}(w, \text{Japanese}) = 3$

Training Naïve Bayes

wordcount$(w, c) + 1$		c	
		Chinese	Japanese
w	Chinese	$5 + 1$	$1 + 1$
	Tokyo	$0 + 1$	$1 + 1$
	Japan	$0 + 1$	$1 + 1$
	...	...	...

- $\sum_{w \in V} \text{wordcount}(w, \text{Chinese}) = 8$
- $\sum_{w \in V} \text{wordcount}(w, \text{Japanese}) = 3$
- $|V| = 6$

Training Naïve Bayes

$\hat{P}(w\|c)$		c	
		Chinese	Japanese
w	Chinese	$(5+1)/(8+6)$	$(1+1)/(3+6)$
	Tokyo	$(0+1)/(8+6)$	$(1+1)/(3+6)$
	Japan	$(0+1)/(8+6)$	$(1+1)/(3+6)$
	...	...	...
- $\sum_{w \in V} \text{wordcount}(w, \text{Chinese}) = 8$
- $\sum_{w \in V} \text{wordcount}(w, \text{Japanese}) = 3$
- $|V| = 6$

Training Naïve Bayes

$\hat{P}(w\|c)$		c	
		Chinese	Japanese
w	Chinese	3/7	2/9
	Tokyo	1/14	2/9
	Japan	1/14	2/9
	...	...	...

- $\sum_{w \in V} \text{wordcount}(w, \text{Chinese}) = 8$
- $\sum_{w \in V} \text{wordcount}(w, \text{Japanese}) = 3$
- $|V| = 6$

Testing Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?

Testing Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$

Testing Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
- $P(\text{Chinese}|d) \propto 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$

Testing Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
- $P(\text{Chinese}|d) \propto 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$
- $P(\text{Japanese}|d) \propto 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001$

Testing Naïve Bayes

- Suppose we observe a document $d =$ "Chinese Chinese Chinese Tokyo Japan". Is the document Chinese or Japanese?
- $c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(w_i|c)$
- $P(\text{Chinese}|d) \propto 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$
- $P(\text{Japanese}|d) \propto 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001$
- Chinese

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 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 non-feature words
    - $c_{NB} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$

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 non-feature words
    - $c_{NB} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$
  - 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 non-feature words
    - $c_{NB} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{i=1, w_i \in \text{features}}^{n} P(w_i|c)$
  - 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 $\log(x) + \log(y) = \log(z)$
- $c_{NB} = \underset{c \in C}{\text{argmax}} \log(P(c)) + \sum_{i=1}^{n} \log(P(w_i|c))$

Working with Logs

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

Working with Logs

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