Naive Bayes Event Models

CS114 Lab 4
February 8, 2018
Kenneth Lai

Event Models

- Recall that Naive Bayes models are generative
  - Assume the data are generated according to an underlying distribution
- Event models are models of the underlying distribution

Multinomial Naive Bayes

- Data generated by multinomial distribution
  - “rolling a k-sided die n times”
  - Feature (word) counts do matter
  - Features (words) that don’t appear in a test document don’t matter

Multinomial Naive Bayes

- Positions ← all word positions in test document
- V ← vocabulary

$$c_{NB} = \underset{c \in C}{argmax}\, P(c) \prod_{i \in positions} P(w_i | c)$$

$$P(w_i | c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)}$$

Text Classification and Naïve Bayes

Multinomial Naïve Bayes: A Worked Example (from Jurafsky and Martin)

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|c) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|c) = (0+1) / (8+6) = 1/14

P(Japan|c) = (0+1) / (8+6) = 1/14

P(Chinese|j) = (1+1) / (3+6) = 2/9

P(Tokyo|j) = (1+1) / (3+6) = 2/9

P(Japan|j) = (1+1) / (3+6) = 2/9

Choosing a class:

P(c|d) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14 $\approx$ 0.0003

P(j|d) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9 $\approx$ 0.0001

Bernoulli Naive Bayes

- Data generated by Bernoulli distribution
  - Flipping a (possibly unfair) coin once
  - Feature (word) counts don’t matter
  - Features (words) that don’t appear in a test document do matter

Bernoulli Naive Bayes

- V ← vocabulary
- B ← 1 if word i appears in test document, else 0
- N ← number of documents

$$c_{NB} = \underset{c \in C}{argmax}\, P(c) \prod_{i \in V} \left(B_i P(w_i|c) + (1 - B_i) P(-w_i|c)\right)$$

$$P(w_i|c) = \frac{N_{w_i,c}}{N_{doc}}$$

$$P(-w_i|c) = \frac{N_{-w_i,c}}{N_{doc}}$$

Text Classification and Naïve Bayes

Bernoulli Naïve Bayes: A Worked Example

$$\hat{P}(c)=\frac{N_c}{N}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c)=\frac{3}{4}$

$P(j)=\frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (3+1) / (3+2) = 4/5
P(Tokyo|*c*) = (0+1) / (3+2) = 1/5
P(Japan|*c*) = (0+1) / (3+2) = 1/5
P(-Beijing|*c*) = (2+1) / (3+2) = 3/5
P(-Shanghai|*c*) = (2+1) / (3+2) = 3/5
P(-Macao|*c*) = (2+1) / (3+2) = 3/5
P(Chinese|*j*) = (1+1) / (1+2) = 2/3
P(Tokyo|*j*) = (1+1) / (1+2) = 2/3
P(Japan|*j*) = (1+1) / (1+2) = 2/3
P(-Beijing|*j*) = (1+1) / (1+2) = 2/3
P(-Shanghai|*j*) = (1+1) / (1+2) = 2/3
P(-Macao|*j*) = (1+1) / (1+2) = 2/3

Choosing a class:

P(c|d) $\propto$ 3/4 * 4/5 * 1/5 * 1/5 * 3/5 * 3/5 * 3/5
$\approx$ 0.0052

P(j|d) $\propto$ 1/4 * 2/3 * 2/3 * 2/3 * 2/3 * 2/3 * 2/3
$\approx$ 0.0219

Common to both models

- P(class) is the same
- Features (words) that don’t appear in the training data don’t matter

Which event model to use?

- Both are widely used, both in CL and elsewhere
- The book and HW 3 both use multinomial NB
  - But if you’re ever unsure, ask!