Text classification using Naive Bayes

Naive Bayes

- Naive bayes is a probabilistic algorithm

Naive Bayes

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

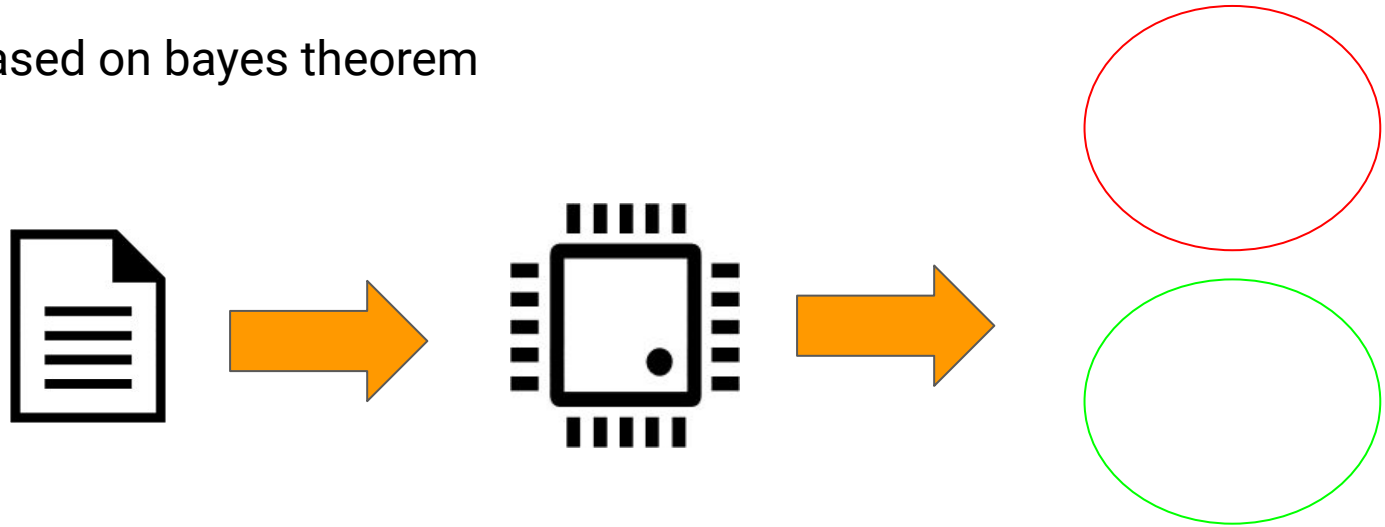
Naive Bayes

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

$$P(E_1 \mid E_2) = \frac{P(E_2 \mid E_1) * P(E_1)}{P(E_2)}$$

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem



Document (d)

Features (x_1, x_2, …, x_n)

Class (c)

Analytics Vidhya

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm

- Based on bayes theorem

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

$$C = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

Analytics Vidhya

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

$$C = \underset{c \in C}{\operatorname{argmax}}\, P(c \mid d) = \frac{P(d \mid c) * P(c)}{\boxed{P(d)}}$$

How likely the document will occur

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

$$C = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = P(d \mid c) * P(c)$$

Naive Bayes for text classification

- Naive bayes is a probabilistic algorithm
- Based on bayes theorem

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)}$$

$$C = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = P(d \mid c) * P(c)$$

$$= P(x_1, x_2, \ldots, x_n \mid c) * P(c)$$

Analytics Vidhya

Naive Bayes for text classification

$$C = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = P(d \mid c) * P(c)$$

$$= P(x_1, x_2, \ldots, x_n \mid c) * P(c)$$

$$P(x_1, x_2, \ldots, x_n \mid c) =$$

Analytics Vidhya

Naive Bayes for text classification

$$C = \underset{c \in C}{\operatorname{argmax}} P(c \mid d) = P(d \mid c) * P(c)$$

$$= P(x_1, x_2, \ldots, x_n \mid c) * P(c)$$

$$P(x_1, x_2, \ldots, x_n \mid c) = P(x_1|c) * P(x_2|c) * \ldots\ldots * P(x_n|c)$$

Analytics Vidhya

Naive Bayes for text classification

$$C = \underset{c \in C}{\text{argmax}}\, P(c \mid d) = P(d \mid c) * P(c)$$

$$= P(x_1, x_2, \ldots, x_n \mid c) * P(c)$$

$$P(x_1, x_2, \ldots, x_n \mid c) = P(x_1|c) * P(x_2|c) * \ldots\ldots * P(x_n|c)$$

$$P(c) = \frac{\text{Number of documents belonging to class c}}{\text{Total number of documents}}$$

Analytics Vidhya

Naive Bayes for text classification

- Calculate the probability of each class P(c)

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)
  - P(w|c) = Count(w,c) / Count(c)

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)
  - P(w|c) = Count(w,c) / Count(c)
  - Laplace add 1 smoothing to deal with new words

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)
  - P(w|c) = Count(w,c) / Count(c)
  - Laplace add 1 smoothing to deal with new words
  - P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)
  - P(w|c) = Count(w,c) / Count(c)
  - Laplace add 1 smoothing to deal with new words
  - P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]
  - Count(w,c) = Number of times w occurs in documents of class c
  - Count(c) = Number of words in documents of class c

Naive Bayes for text classification

- Calculate the probability of each class P(c)
  - For each class c in C:
    - P(c) = | docs | / | total number of documents |
    - docs is the number of documents belonging to class c

- Calculate the conditional probabilities P(w|c)
  - P(w|c) = Count(w,c) / Count(c)
  - Laplace add 1 smoothing to deal with new words
  - P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]
  - Count(w,c) = Number of times w occurs in documents of class c
  - Count(c) = Number of words in documents of class c
  - |V| = Vocabulary size

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

$P(i) = 3/4$

$P(c) = 1/4$

$$P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]$$

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]

P(Delhi | i) =

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

$P(i) = 3/4$

$P(c) = 1/4$

$P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]$

$P(Delhi \mid i) = [4 + 1] /$

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

$P(i) = 3/4$

$P(c) = 1/4$

$P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]$

$P(Delhi \mid i) = [4 + 1] / [8 +$

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

$P(i) = 3/4$

$P(c) = 1/4$

$P(w|c) = [Count(w,c) + 1] / [Count(c) + |V|]$

$P(Delhi \mid i) = [4 + 1] / [8 + 7]$

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

$P(i) = 3/4$

$P(c) = 1/4$

$P(\text{Delhi} \mid i) = 5/15$

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

Analytics Vidhya

Naive Bayes for text classification

	Doc_ID	Document	Class
Train	1	Chennai Delhi Mumbai	i
	2	Delhi Delhi Kolkata	i
	3	Delhi Gurgaon	i
	4	Beijing Shanghai Delhi	c
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) =

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = P(i)

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = P(i) * P(Delhi | i) * P(Delhi | i) * P(Kolkata | i) * P(Kolkata | i) * P(Beijing | i)

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = P(i) * P(Delhi | i) * P(Delhi | i) * P(Kolkata | i) * P(Kolkata | i) * P(Beijing | i)

P(i | D5) = 3/4 * 5/15 * 5/15 * 2/15 * 2/15 * 1/15

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = P(i) * P(Delhi | i) * P(Delhi | i) * P(Kolkata | i) * P(Kolkata | i) * P(Beijing | i)

P(i | D5) = 3/4 * 5/15 * 5/15 * 2/15 * 2/15 * 1/15 = 0.000098

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = 0.000098

P(c | D5) = P(c) * P(Delhi | c) * P(Delhi | c) * P(Kolkata | c) * P(Kolkata | c) * P(Beijing | c)

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = 0.000098

P(c | D5) = P(c) * P(Delhi | c) * P(Delhi | c) * P(Kolkata | c) * P(Kolkata | c) * P(Beijing | c)

P(c | D5) = 1/4 * 2/10 * 2/10 * 1/10 * 1/10 * 2/10

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = 0.000098

P(c | D5) = P(c) * P(Delhi | c) * P(Delhi | c) * P(Kolkata | c) * P(Kolkata | c) * P(Beijing | c)

P(c | D5) = 1/4 * 2/10 * 2/10 * 1/10 * 1/10 * 2/10 = 0.00002

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = 0.000098

P(c | D5) = 0.00002

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	

P(i | D5) = 0.000098

P(c | D5) = 0.00002

P(i | D5) > P(c | D5)

Analytics Vidhya

Naive Bayes for text classification

P(i) = 3/4

P(c) = 1/4

P(Delhi | i) = 5/15

P(Kolkata | i) = 2/15

P(Beijing | i) = 1/15

P(Delhi | c) = 2/10

P(Kolkata | c) = 1/10

P(Beijing | c) = 2/10

	Doc_ID	Document	Class
Test	5	Delhi Delhi Kolkata Kolkata Beijing	i

P(i | D5) = 0.000098

P(c | D5) = 0.00002

P(i | D5) > P(c | D5)

Analytics Vidhya

Thank You

Analytics Vidhya