Links

Google Slides:
https://docs.google.com/presentation/d/1bozof8hrTUrcuXt2obVfjLMgg1zzrHzGEScG-GuOXog/edit?usp=sharing

Github:

https://github.com/ademiltonnunes/Machine-Learning/tree/main/Text%20Classification

Text Classification

Week 7 - Homework 1
CS550 - Machine Learning and Business Intelligence

Ademilton Marcelo da Cruz Nunes (19679)

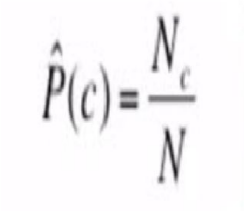
Table of Content

Introduction

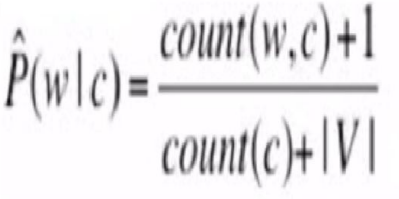
This project aims to classify who is the real author of Hamlet.

Introduction - Text Classification

Text classification is a type of machine learning algorithm that categorizes text into categories or classes. The text classification formulas are:



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



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

Where:

- P(c) = probability of a classe
  - C = a class
  - Nc = Number documents in that class
  - N = Total number in all classes

- P(w|c) = probability of a word belongs to a class
  - W = the word
  - count(w|c) = how many times a word appears in a class
  - count(c) = how many word are in a class
  - |V|= Total vocabulary (words) in all classes

Text Classifier

Test the Text Classifier to predict who the real author of Hamlet is.

	Doc	Words	Author
Training	**1**	**W1 W2 W3 W4 W5**	**C (Christopher Marlowe)**
	2	**W1 W1 W4 W3**	**C (Christopher Marlowe)**
	3	**W1 W2 W5**	**C (Christopher Marlowe)**
	4	**W5 W6 W1 W2 W3**	**W (William Stanley)**
	5	**W4 W5 W6**	**W (William Stanley)**
	6	**W4 W6 W3**	**F (Francis Bacon)**
	7	**W2 W2 W4 W3 W5 W5**	**F (Francis Bacon)**
Test	**8 (Hamlet)**	**W1 W4 W6 W5 W3**	**???**

Step 1: Training - Verify each author probability

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

Let's verify what is the probability of a document belongs to each author.

Applying the probability formula:
$P(c) = Nc+1/N$

Author C
$P(c) = 3/7$

Author W
$P(w) = 2/7$

Author F
$P(f) = 2/7$

Step 2: Training - Verify each word probability

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

Let's verify what is the probability of each word of Hamlet belongs to each author.

Applying the probability formula:
$P(w|c) = count(w,c)+1/count(c)+|V|$

Author C

$P(W1|C) = 4+1/12+6 = 5/18$
$P(W4|C) = 2+1/12+6 = 3/18$
$P(W6|C) = 0+1/12+6 = 1/18$
$P(W5|C) = 2+1/12+6 = 3/18$
$P(W3|C) = 2+1/12+6 = 3/18$

Step 2: Training - Verify each word probability

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

Author W

$P(W1|W) = 1+1/8+6 = 2/14$

$P(W4|W) = 1+1/8+6 = 2/14$

$P(W6|W) = 2+1/8+6 = 3/14$

$P(W5|W) = 2+1/8+6 = 3/14$

$P(W3|W) = 1+1/8+6 = 2/14$

Author F

$P(W1|F) = 0+1/9+6 = 1/15$

$P(W4|F) = 2+1/9+6 = 3/15$

$P(W6|F) = 1+1/9+6 = 2/15$

$P(W5|F) = 2+1/9+6 = 3/15$

$P(W3|F) = 2+1/9+6 = 3/15$

Step 3: Test- Verify what author Hamlet belongs to

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

The Bayes' theorem is:

$$P(h|D) = P(h) \cdot \frac{P(D|h)}{P(D)}$$

Where:

- P(h|D) = Probability of a document belongs to a class
- P(h) = probability of a class
- P(D|h) = Probability of a class has a document
- P(D) = probability of a document

The Naive Bayes Classifier assumes that a conditional is independence. Therefore, Naive Bayes compared model with Bayes theorem has the formula:

P(h|D) = P(h)* P(D|h)

Step 3: Test- Verify what author Hamlet belongs to

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

The probability of a Hamlet belong to author C through Bayes' theorem is:

$$P(c|D8) = P(c)*P(D8|c) / P(D8)$$
$$P(c|D8) = P(c)* P(W1|c)* P(W4|c)* P(W6|c)* P(W5|c)* P(W3|c)/P(D8)$$

The compare model of Bayes theorem and Naive Bayes is:

$$P(c|D8) = P(c)* P(W1|c)* P(W4|c)* P(W6|c)* P(W5|c)* P(W3|c)$$
$$P(c|D8) = 3/7* 5/18* 3/18*1/18*3/18*3/18$$
$$P(c|D8) = 0.43*0.28*0.17*0.06*0.17*0.17$$
$$P(c|D8) =0.000031$$

Step 3: Test- Verify what author Hamlet belongs to

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

The probability of a Hamlet belong to author W through compare model of Bayes' and Naive Bayes is:

$$P(w|D8) = P(w)* P(W1|w)* P(W4|w)* P(W6|w)* P(W5|w)* P(W3|w)$$
$$P(w|D8) = 2/7* 2/14*2/14*3/14*3/14*2/14$$
$$P(w|D8) = 0.29*0.14*0.14*0.21*0.21*0.14$$
$$P(w|D8) = 0.000038$$

Step 3: Test- Verify what author Hamlet belongs to

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	???

The probability of a Hamlet belong to author F through compare model of Bayes' and Naive Bayes is:

$$P(F|D8) = P(F)* P(W1|F)* P(W4|F)* P(W6|F)* P(W5|F)* P(W3|F)$$
$$P(F|D8) = 2/7* 1/15*3/15*2/15*3/15*3/15$$
$$P(F|D8)= 0.29*0.07*0.20*0.13*0.20*0.20$$
$$P(F|D8) = 0.000020$$

Conclusion

Compare the probability result those three authors:

- P(c|D8) =0.000031
- P(w|D8) = 0.000038
- P(F|D8) = 0.000020

There highest probability is P(w|D8), therefore the author of Hamlet is W (William Stanley).

References

labnet. (n.d.). *Text Classifier*. labnet. Retrieved March 2, 2023, from

https://hc.labnet.sfbu.edu/~henry/sfbu/course/mllib/naive_bayes/slide/text_classifier.html