Text Classification

Homework 1 Week 7
Machine Learning and Business Intelligence CS 550
Presented by Imran Noor Saleh ID 19648

Introduction

- Text classification is the process of assigning predefined categories to text data based on their content.
- The objective of our project is to develop a machine learning model that can accurately classify text data into different categories.
- We used various machine learning algorithms and techniques to achieve our objective.
- To prepare the dataset for machine learning, we performed feature extraction and selection techniques. We used bag-of-words, TF-IDF, and word embeddings to represent the text data as numerical features.
- We also performed feature selection techniques to identify the most relevant features that contribute to the classification task.
- We used various machine learning algorithms such as Naive Bayes, Support Vector Machines, and Neural Networks for text classification.

Dataset

The objective is to test the Text Classifier to predict who the real author of Hamlet is, using this dataset.

	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	?

Manual Calculation

To predict who the real author of Hamlet is, we have the training data and the probability of each to be calculated

P(C) : The probability of class C = 3/7

P(W) : The probability of class W = 2/7

P(F) : The probability of class F = 2/7

P(W1|C): The probability that the word "W1" appears on the 3 class c documents

= (count(W1, C) + 1) / (count(C)+|V|) = (4+1) / (12+6) = 5/18

4: how many times the word "W1" appear on the 3 class C documents., 12: how many words in the 3 class C documents, 6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W1|W) : The probability that the word "W1" appears on the 3 class W documents

= (count(W1, W) + 1) / (count(W)+|V|)

= (1+1) / (8+6) = 2/14 = 1/7

1: how many times the word "W1" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W1|F) : The probability that the word "W1" appears on the 2 class F documents

= (count(W1, F) + 1) / (count(F)+|V|)

= (0+1) / (9+6) = 1/15

0: how many times the word "W1" appear on the 2 class F documents.

9: how many words in the 3 class W documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W3|C) : The probability that the word "W3" appears on the 3 class C documents

= (count(W3, C) + 1) / (count(C)+|V|)

= (2+1) / (12+6) = 3/18 = ⅙

2: how many times the word "W3" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W3|W) : The probability that the word "W3" appears on the 3 class W documents

= (count(W3, W) + 1) / (count(W)+|V|)

= (1+1) / (8+6) = 2/14 = 1/7

1: how many times the word "W3" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W3|F) : The probability that the word "W3" appears on the 2 class F documents

= (count(W3, F) + 1) / (count(F)+|V|)

= (2+1) / (9+6) = 3/15 = 1/5

2: how many times the word "W3" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W4|C) : The probability that the word "W4" appears on the 3 class C documents

= (count(W4, C) + 1) / (count(C)+|V|)

= (2+1) / (12+6) = 3/18 = 1/

2: how many times the word "W4" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W4|W) : The probability that the word "W4" appears on the 3 class W documents

= (count(W4, W) + 1) / (count(W)+|V|)

= (1+1) / (8+6) = 2/14 = 1/7

1: how many times the word "W4" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W4|F) : The probability that the word "W4" appears on the 2 class F documents

= (count(W4, F) + 1) / (count(F)+|V|)

= (2+1) / (9+6) = 3/15

2: how many times the word "W4" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W5|C): The probability that the word "W5" appears on the 3 class C documents

= (count(W5, C) + 1) / (count(C)+|V|)

= (2+1) / (12+6) = 3/18 = 1/6

2: how many times the word "W5" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W5|W): The probability that the word "W5" appears on the 3 class W documents

= (count(W5, W) + 1) / (count(W)+|V|)

= (2+1) / (8+6) = 3/14

2: how many times the word "W5" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W5|F): The probability that the word "W5" appears on the 2 class F documents

= (count(W5, F) + 1) / (count(F)+|V|)

= (2+1) / (9+6) = 3/15

2: how many times the word "W5" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6).

Manual Calculation Contd.

P(W6|C): The probability that the word "W6" appears on the 3 class C documents

= (count(W6, C) + 1) / (count(C)+|V|)

= (0+1) / (12+6) = 1/18

0: how many times the word "W6" appear on the 3 class C documents.

12 : how many words in the 3 class C documents.

6 : number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W6|W): The probability that the word "W6" appears on the 2 class W documents

= (count(W6, W) + 1) / (count(W)+|V|)

= (2+1) / (8+6) = 3/14

2: how many times the word "W6" appear on the 2 class W documents.

8 : how many words in the 3 class W documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(W6|F) : The probability that the word "W6" appears on the 2 class F documents

= (count(W6, F) + 1) / (count(F)+|V|)

= (1+1) / (9+6) = 2/15

1: how many times the word "W6" appear on the 2 class F documents.

9: how many words in the 3 class F documents.

6: number of vocabulary: (W1 W2 W3 W4 W5 W6)

Manual Calculation Contd.

P(C|d8) : P(C) * P(W1|C) * P(W4|C)* P(W6|C) * P(W5|C) * P(W3|C)

= ((3/7) * (5/18)* (1/6)* (1/18) *(1/6) *(1/6))

= 0.00003061924 , approx 0.00003

= 3/7: prior : P(C)

There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has P(W1|C) = 5/18, The word "W4" has P(W4|C) =3/18 = ⅙, The word "W6" has P(W6|C) = 1/18, The word "W5" has P(W5|C) = 3/18 = ⅙, The word "W3" has P(W3|C) = 3/18 = 1/6

Manual Calculation Contd.

P(W|d8) = P(W) * P(W1|W) * P(W4|W)* P(W6|W) * P(W5|W) * P(W3|W)

= (2/7* 2/14 * 2/14 * 3/14 * 3/14 * 2/14)

= 0.00004 = 2/7: prior : P(W)

There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has P(W1|W) = 2/14, The word "W4" has P(W4|W)= 2/14, The word "W6" has P(W6|W)= 3/14, The word "W5" has P(W5|W) = 3/14, The word "W3" has P(W3|W) = 2/14

Manual Calculation Contd.

P(F|d8) = P(F) * P(W1|F) * P(W4|F)* P(W6|F) * P(W5|F) * P(W3|F)

=((2/7) * (1/15)*(3/15) * (2/15) * (3/15) * (3/15))

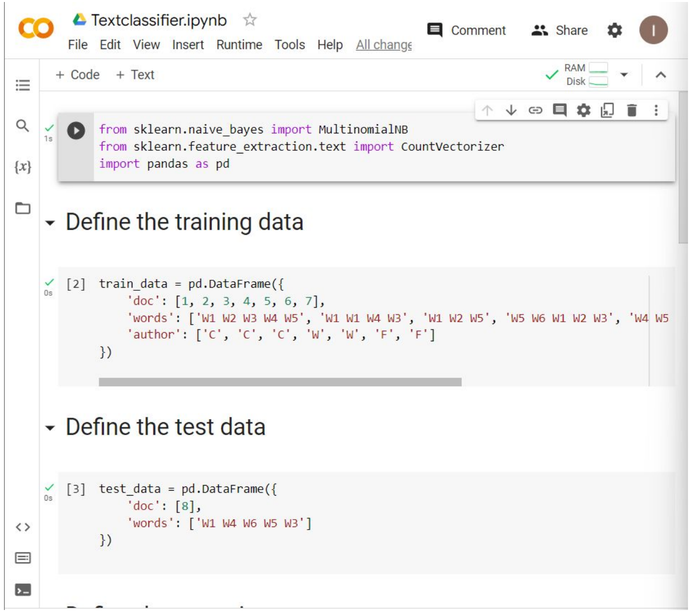
= 0.00002 = 2/7:prior : P(F)

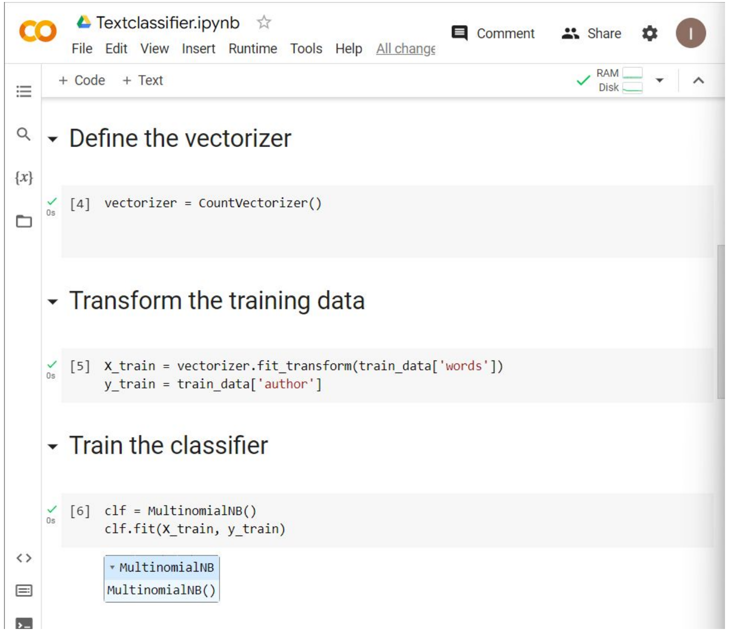
There are 5 words in d8 : W1 W4 W6 W5 W3

Each word "W1" has P(W1|F) = 1/15, The word "W4" has P(W4|F)= 3/15, The word "W6" has P(W6|F)= 2/15 , The word "W5" has P(W5|F)= 3/15, The word "W3" has P(W3|F) = 3/15

Document 8 should belong to class W because it has the highest probability calculation.

Programming Solution





Programming Solution Contd.

So, the manually calculated and programmed solution is the same for this problem.