Multinomial Naive Bayes

Document classification is a classical machine learning problem. If there is a set of documents that are already categorized/labeled in existing categories, the task is to automatically categorize a new document into one of the existing categories.

**Step 1**

Calculate prior probabilities. These are the probability of a document being in a specific category from the given set of documents.

Equation is :

P(Category) = (No. of documents classified into the category) divided by (Total number of documents)

**Step 2**

Calculate Likelihood. Likelihood is the conditional probability of a word occurring in a document given that the document belongs to a particular category.

Equation is :

P(Word/Category) = (Number of occurrence of the word in all the documents from a category+1) divided by (All the words in every document from a category + Total number of unique words in all the documents)

**Step 3**

Calculate P(Category/Document) = P(Category) \* P(Word1/Category) \* P(Word2/Category) \* P(Word3/ Category) …….. \* P[ (Word)n / Category ]

|  |  |  |  |
| --- | --- | --- | --- |
| Serial | Action | Word | Class |
| 1 | Training | Chinese Beijing Chinese | C |
| 2 | Chinese Chinese Shanghai | C |
| 3 | Chinese Macao | C |
| 4 | Tokyo Japan Chinese | J |
| 5 | Test | Chinese Chinese Chinese Tokyo Japan | ? |

**Data Table**

All unique words and their frequencies are:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Words | Class | Beijing | Chinese | Japan | Macao | Shanghai | Tokyo |
| Chinese Beijing Chinese | C | 1 | 2 | 0 | 0 | 0 | 0 |
| Chinese Chinese Shanghai | C | 0 | 2 | 0 | 0 | 1 | 0 |
| Chinese Macao | C | 0 | 1 | 0 | 1 | 0 | 0 |
| Tokyo Japan Chinese | J | 0 | 1 | 1 | 0 | 0 | 1 |

**Priors:**

P( C ) = 3 / 4

P( J ) = 1 / 4

**Conditional Probabilities:**

P( Beijing | C ) = (1+1) / (8+6) = 1/7

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

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

P( Macao | C ) = (1+1) / (8+6) = 1/7

P( Shanghai | C ) = (1+1) / (8+6) = 1/7

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

P( Beijing | J ) = (0+1) / (3+6) = 1/9

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

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

P( Macao | J ) = (0+1) / (3+6) = 1/9

P( Shanghai | J ) = (0+1) / (3+6) = 1/9

P( Tokyo | J ) = (0+1) / (3+6) = 2/9

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Word | C | J | P( Word | C ) | P( Word | J ) |
| Beijing | 1 | 0 | 1/7 | 1/9 |
| Chinese | 5 | 1 | 3/7 | 2/9 |
| Japan | 0 | 1 | 1/14 | 2/9 |
| Macao | 1 | 0 | 1/7 | 1/9 |
| Shanghai | 1 | 0 | 1/7 | 1/9 |
| Tokyo | 0 | 1 | 1/14 | 2/9 |

**For a test text, Choosing a class:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Frequency ( f ) | | | | | |
| Word | Beijing | Chinese | Japan | Macao | Shanghai | Tokyo |
| Chinese Chinese Chinese Tokyo Japan | 0 | 3 | 1 | 0 | 0 | 1 |

**Step -1**

P( C ) = 3 / 4

P( J ) = 1 / 4

**Step -2**

For class J ;

P( Chinese | J ) = 2/9

P( Tokyo | J ) = 2/9

P( Japan | J ) = 2/9

For class C ;

P( Chinese | C ) = 3/7

P( Tokyo | C ) = 1/14

P( Japan | C ) = 1/14

**Step -3**

P( C | Test ) = P( C ) \* { P( Chinese | C ) } f \* { P( Tokyo | C ) } f \* { P( Japan | C ) } f

= (3/4) \* ( 3/7 )3 \* (1/14) \* (1/14) = 0.0003

P( J | Test ) = P( J ) \* { P( Chinese | J ) } f \* { P( Tokyo | J ) } f \* { P( Japan | J ) } f

= (1/4) \* ( 2/9 )3 \* (2/9) \* (2/9) = 0.0001

As , P( C | Test ) > P( J | Test ) or ( 0.0003 ) > ( 0.0001 ). So, we can choose **Class C.**