Multinomial NB for Text Classification

Dr. Muhammad Wasim

Text Classification

- Textual documents are different from relational or tabular data
- So, they need to be transformed and represented in a format suitable for classifiers
- Moreover, textual documents may have variations and features not helpful for classification

An example of a text corpus

Example	Document	Class
Doc-1	I love playing cricket.	Sports
Doc-2	I often play badminton.	Sports
Doc-3	I am taking a vaccine	Medical
Doc-4	headache	Medical

Doc	I	love	playing	cricket	often	play	badminton	am	taking	a	vaccine	headache	Class
1	1	1	1	1	0	0	0	0	0	0	0	0	Sports
2	1	0	0	0	1	1	1	0	0	0	0	0	Sports
3	1	0	0	0	0	0	0	1	1	1	1	0	Medical
4	0	0	0	0	0	0	0	0	0	0	0	1	Medical

Preprocessing Text for Feature usefulness, variation, and weighting

Doc	I	love	playing	cricket	often	play	badminton	am	taking	a	vaccine	headache	Class
1	1	1	1	1	0	0	0	0	0	0	0	0	Sports
2	1	0	0	0	1	1	1	0	0	0	0	0	Sports
3	1	0	0	0	0	0	0	1	1	1	1	0	Medical
4	0	0	0	0	0	0	0	0	0	0	0	1	Medical

- Feature usefulness: repeating features may not help in classification. Such features are often call stopwords and removed in preprocessing.
- Variation: words with different forms. Normalization is used to handle variations.
- Normalization includes lowercasing words, removing special characters, and stemming /lemmatization

Preprocessing Text for Feature usefulness, variation, and weighting

Doc	I	love	playing	cricket	often	play	badminton	am	taking	a	vaccine	headache	Class
1	1	1	1	1	0	0	0	0	0	0	0	0	Sports
2	1	0	0	0	1	1	1	0	0	0	0	0	Sports
3	1	0	0	0	0	0	0	1	1	1	1	0	Medical
4	0	0	0	0	0	0	0	0	0	0	0	1	Medical

Weighting:

- Term document incidence matrix (Binary Features)
- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- TF-IDF which is achieved by multiplying TF with its IDF value.

Bayes' rule applied to documents and classes

For a document d and class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes' Classifier (I)

$$C_{MAP} = argmax_{c \in C}\, P(c|d)$$

$$C_{MAP} = argmax_{c \in C} \frac{P(d|c)\, P(c)}{P(d)}$$

$$C_{MAP} = argmax_{c \in C}\, P(d|c)P(c)$$

Naïve Bayes' Classifier (II)

$$C_{MAP} = argmax_{c \in C}\, P(d|c)P(c)$$

$$C_{MAP} = argmax_{c \in C}\, P(x_1, x_2, \ldots, x_n\, |c)P(c)$$

Multinomial Naïve Bayes Independence Assumptions

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

Bag of words assumption: Assume position doesn't matter.

Conditional Independence: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c.

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

$$\mathrm{C_{NB}} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i} P(x_i|c_j)$$

Multinomial NB for Text Classification - II

Dr. Muhammad Wasim

Multinomial Naïve Bayes

$$\mathrm{C_{NB}} = \mathrm{argmax}_{c_j \in C}\, P(c_j) \prod_i P(x_i|c_j)$$

Learning the Multinomial Naïve Bayes Model

Maximum likelihood estimates

Simply use the frequencies in data

Example	Document	Class
Doc-1	I love cricket.	Sports
Doc-2	I often play badminton	Sports
Doc-3	I am taking a vaccine	Medical
Doc-4	Having headache	Medical

$$P(C_j) = \frac{docCount(C = c_J)}{N_{doc}}$$

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

Problem with Maximum Likelihood

What if we have seen no training documents with the word fantastic and has the topic positive

$$P(\text{"}fantastic\text{"}|positive) = \frac{count(\text{"}fantastic\text{"}, positive)}{\sum_{w \in V} count(w, positive)} = 0$$

Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\mathrm{C_{MAP}} = \mathrm{argmax}_c\, P(c) \prod_i P(x_i|c)$$

Laplace (add-1) smoothing for Naïve Bayes

$$P(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} count(w, c_j) + 1}$$

$$P(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\left[\sum_{w \in V} count(w, c_j)\right] + |V|}$$

A Walkthrough Example

	Doc	Document	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	

$$P(C_j) = \frac{docCount(C = c_J)}{N_{doc}}$$

$$P(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\left[\sum_{w \in V} count(w, c_j)\right] + |V|}$$

Priors: $P(c) = \frac{3}{4}$ $P(j) = \frac{1}{4}$

Choosing a class:

$$P(c|d5) = \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} = 0.0003$$

$$P(j|d5) = \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \frac{2}{9} \times \frac{2}{9} = 0.0001$$

Conditional Probabilities

$$P(Chinese|c) = \frac{5+1}{8+6} = \frac{3}{7} \quad P(Chinese|j) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(Tokyo|c) = \frac{0+1}{8+6} = \frac{1}{14} \quad P(Tokyo|j) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(Japan|c) = \frac{0+1}{8+6} = \frac{1}{14} \quad P(Japan|j) = \frac{1+1}{3+6} = \frac{2}{9}$$