Fundamental Mathematic Report

NAME: XIONG, ZHIPENG SID: 44231536
zhipeng@akane.waseda.jp
Graduate School of Information, Production and Systems, Waseda University

The Mathematic methods used in Artifical Intellgence

The text cllassification problem

In text classification, we are given a description d ∈ **X**. of a document, where **X** is the document sapce; and a fixed set of classes $\mathbf{C} = \{c_1, c_2 \ldots, c_j\}$. Classes are also called *categories* or *labels*. Typiaclly, the document space **X** is some type of high-dimensional space, and the classes are human defined for the needs of an application.

Bayes theorem

As we all know, Bayes' theorem, is a fundamental concept in probability theory and statistics.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- P(A|B) is the conditional probability of event A given event B
- P(B|A) is the conditional probability of event B given event A
- P(A) and P(B) are the probability of events A and B, respectively.

In a word, Bayes' theorem states that the probability of event A occurring given that event B has occurred is equal to the probability event B occurring given that event A has occurred, multiiplied by the probability of event A occurring(prior pprobability of A), divided by the probability of event B occurring.

Naive Bayes text classification

The probability of a document *d* being in class *c* is computed as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

where $P(t_k|c)$ is conditional probability of term t_k occurring in a document of class *c*. We interpret $P(t_k|c)$ as a measure of how much evidence t_k contributes that *c*. is the correct class. P(c) is the prior pobability of document occcurring in class *c*. If a document's term do not provide clear evidence for one

class versus another, we choose the one that has a higher prior probability. $< t_1, t_2, \ldots, t_{n_d} >$ are the tokens in *d* that are part of the vocabulary we use for classification and n_d is the number of such tokens in d. For example, $< t_1, t_2, \ldots, t_{n_d} >$ for the one-sentence document *Beijing and Taipei join the WTO* might be <Beijing, Taipei, join, WTO>, with $n_d = 4$, if we the terms *and* and *the* as stop words.

```
In natural language processing(NLP), stop words are common wods that are often considered
insignificant or irrelevant for the analysis of text data
```

Justification

In the case of Naive Bayes text classification, we assume that the words are independent of each other, meaning that the presence or absence of one word does not affect the probability of another word being present or absent. This allows us to simplify the formula for P(d|c) as follows:

$$P(d|c) = \prod_{i=1}^{n} P(t_i|c)$$

where $P(t_i|c)$ is the probability of the i-th word x_i, given the class c.

Substituting this formula into Bayes' theroem, we get:

$$P(c|d) = P(c) \cdot \frac{\prod_{i=1}^{n} P(t_i|c)}{P(d)} \rightarrow P(c|d) \propto P(c) \cdot \prod_{i=1}^{n} P(t_i|c)$$

Solution

In text classification, our goal is to find the best class for the document. The best class in NB classification is the most likely or *maximum a posteriori* (MAP) class c_{map}:

$$c_{map} = \arg\max_{c \in C} \hat{P}(c|d) = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

We write $\hat{P}$ for P because we do not know the true values of the parameters P(c) and $P(t_k|c)$, but estimate them from the training set as we will see in a moment.

We first try the maximum likelihood estimate, which is simply the relative frequency and corresponds to the most likely value of each parameter given the training data. For the priors this estimate is:

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

where N_c is the number of documents in class c and N is the total number of documents.

We estimate the conditional probability $\hat{P}(t|c)$ as the relative frequency of term t in documents belonging to class c:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum\limits_{t' \in V} T_{ct'}}$$

where T_{ct} is the number of occurrences of t in training documents from class c, including multiple occurrences of a term in a document.

To eliminate zeros, we use add-one or Laplace smoothing, which simply adds one to each count:
$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)} = \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+B')}$

where B = |V| is the number of terms in the vocabulary. Add-one smoothing can be interpreted as a uniform prior (each term occurs once for each class) that is then updated as evidence from the training data comes in.

Example

	docID	words in document	in c = China?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

the multinomial parameters we need to classify the test document are the priors $\hat{P}(c) = \frac{3}{4}$ and $\hat{P}(c) = \frac{1}{4}$ and the following conditional probabilities:

$$\hat{P}(Chinese|c) = (5+1)/(8+6) = 6/14 = 3/7$$
$$\hat{P}(Tokyo|c) = \hat{P}(Japan|c) = (0+1)/(8+6) = 1/14$$
$$\hat{P}(Chinese|\overline{c}) = (1+1)/(3+6) = 2/9$$
$$\hat{P}(Tokyo|\overline{c}) = \hat{P}(Japan|\overline{c}) = (1+1)/(3+6) = 2/9$$

The denominators are (8 + 6) and (3 + 6) because the lengths of $text_c$ and $text_{\overline{c}}$ are 8 and 3, respectively, and because the constant B is 6 as the vocabulary consists of six terms.

We then get:

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$
$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Thus, the classifier assigns the test document to c = China. The reason for this classification decision is that the three occurrences of the positive indicator Chinese in d5 outweigh the occurrences of the two negative indicators Japan and Tokyo.