

Machine learning - Lab2

Carlos Alfonso Gómez Hernández

October 8, 2018

1 Lab 2.1

Based on the table 1:

	docID	words in document	in c = China?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

Table 1: Data for parameter estimation exercise

1.1 Estimation - Multinomial Naive Bayes classifier

Probability of class (China, not China) in training set

$$\hat{P}(c) = \frac{1}{2} = 0.5$$

$$\hat{P}(\bar{c}) = \frac{1}{2} = 0.5$$

Probability of each word in test set depending of class

Vocabulary has 7 words. Each class have 5 terms.

$$\hat{P}(Taiwan|c) = \frac{1}{4} = 0.25$$

$$\hat{P}(Sapporo|c) = \frac{1}{12} = 0.0833333...$$

$$\hat{P}(Taiwan|\bar{c}) = \frac{1}{6} = 0.1666666...$$

$$\hat{P}(Sapporo|\bar{c}) = \frac{1}{4} = 0.25$$

1.2 Applying the classifier to the test document

Using formula

$$\hat{P}(c|doc5) = \frac{1}{2} * \left(\frac{1}{4}\right)^2 * \frac{1}{12} \approx 0.0026041666666666665$$

$$\hat{P}(\bar{c}|doc5) = \frac{1}{2} * \left(\frac{1}{6}\right)^2 * \frac{1}{4} \approx 0.0034722222222222222$$

Document in \bar{c}

1.3 Estimation - Bernoulli NB classifier

Probability of each word in test set depending of class

There is two classes (China or not China). Each class have 2 documents.

$$\hat{P}(Taiwan|c) = \frac{3}{4} = 0.75$$

$$\hat{P}(Sapporo|c) = \hat{P}(Osaka|c) = \hat{P}(Japan|c) = \frac{1}{4} = 0.25$$

$$\hat{P}(Taipei|c) = \hat{P}(Macao|c) = \hat{P}(Shanghai|c) = \frac{1}{2} = 0.5$$

$$\hat{P}(Sapporo|\bar{c}) = \frac{3}{4} = 0.75$$

$$\hat{P}(Taiwan|\bar{c}) = \hat{P}(Osaka|\bar{c}) = \hat{P}(Japan|\bar{c}) = \frac{1}{2} = 0.5$$

$$\hat{P}(Taipei|\bar{c}) = \hat{P}(Macao|\bar{c}) = \hat{P}(Shanghai|\bar{c}) = \frac{1}{4} = 0.25$$

1.4 Applying the classifier to the test document

Using formula and previous values calculated

$$\begin{aligned} \hat{P}(c|doc5) &= \hat{P}(c) * \hat{P}(Taiwan|c) * \hat{P}(Sapporo|c) * (1 - \hat{P}(Osaka|c)) * (1 - \hat{P}(Japan|c)) \\ &\quad * (1 - \hat{P}(Shanghai|c)) * (1 - \hat{P}(Macao|c)) * (1 - \hat{P}(Taipei|c)) \\ &\approx 0.006591796875 \end{aligned} \tag{1}$$

$$\begin{aligned} \hat{P}(\bar{c}|doc5) &= \hat{P}(\bar{c}) * \hat{P}(Taiwan|\bar{c}) * \hat{P}(Sapporo|\bar{c}) * (1 - \hat{P}(Osaka|\bar{c})) * (1 - \hat{P}(Japan|\bar{c})) \\ &\quad * (1 - \hat{P}(Shanghai|\bar{c})) * (1 - \hat{P}(Macao|\bar{c})) * (1 - \hat{P}(Taipei|\bar{c})) \\ &\approx 0.019775390625 \end{aligned} \tag{2}$$

Document in \bar{c}

2 Algorithm

To modify time complexity of ApplyMultinomialNB, I call a function named SelectDistinctTokens (line 3). This receives as parameter the list of tokens extracted from each document. Then, the conditional probability is only calculated one time for each token (line 6).

Algorithm 1 Apply Multinomial NB

```
1: procedure APPLYMULTINOMIALNB( $C, V, prior, condprob, d$ )
2:    $W \leftarrow ExtractTokensFromDoc(V, d)$ 
3:    $D \leftarrow SelectDistinctTokens(W)$ 
4:   foreach  $c \in C$  do
5:      $score[c] \leftarrow logprior[c]$ 
6:     foreach  $d \in D$  do
7:        $score[c] += logcondprob[d][c]$ 
8:     end foreach
9:   end foreach
10: return  $argmax_{e \in C} score[c]$  ▷ Naive Bayes algorithm
```
