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

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

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

Probability of each word in test set depending of class
Vocabulary has 7 words. Each class have 5 terms.

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

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

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

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

1.2 Applying the classifier to the test document

Using formula

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

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

Document in $\overline{c}$

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.

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

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

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

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

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

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

1.4 Applying the classifier to the test document

Using formula

$$\hat{P}(c|doc5) = \hat{P}(c)*\hat{P}(Taiwan|c)*\hat{P}(Sapporo|c)*(1-\hat{P}(Osaka|c))*(1-\hat{P}(Japan|c))*(1-\hat{P}(Shanghai|c))*(1-\hat{P}($$

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

Document in $\overline{c}$

2 Algorithm