Lab2

bbagliotto

December 2018

1 Introduction

In this lab we will do examples from a book of machine learning which contains exercices. This book is available at this adress: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html

2 Part 1

Lets use a Naive Bayes classifier for the following inputs:

	Words in document	China?
training set	Taipei Taiwan	yes
	Macao Taiwan Shanghai	yes
	Japan Sapporo	no
	Sapporo Osaka Taiwan	no
test set	Taiwan Taiwan Sapporo	?

Let's first calculate the probability of our X vector apparition:
P(X) = P(Taiwan)*P(Taiwan)*P(Sapporo)
P(X) = 3/4 * 3/4 * 2/4
P(X) = 9/32

Then,

P(X $\mid China = yes) = P(Taiwan \mid China = yes) * P(Taiwan \mid China = yes) * P(Sapporo \mid China = yes) * P(China = yes)$

P(X $\mid China = yes) = 1 * 1 * 0 * 1/2$

$\boxed{P(X \mid China = yes) = 0}$

$P(X \mid China = no) = P(Taiwan \mid China = no) * P(Taiwan \mid China = no) * P(Sapporo \mid China = no) * P(China = no)$

$P(X \mid China = no) = 1/2 * 1/2 * 1 * 1/2$

$\boxed{P(X \mid China = no) = 1/8}$

Now,

$P(China = yes \mid X) = \frac{P(X|China=yes)*P(China=yes)}{P(X)}$

$\boxed{P(China = yes \mid X) = \frac{0}{9/32} = 0}$

$P(China = no \mid X) = \frac{P(X|China=no)*P(China=no)}{P(X)}$

$\boxed{P(China = no \mid X) = \frac{1/8}{9/32} = 8/18}$

After that we can note that $P(China = no \mid X) > P(China = yes \mid X)$, we can say that, for our Bayes classifier, the sentence "Taiwan Taiwan Sapporo" has more chance not to be from China.

3 Part2

The algorithm for applying the Multinomial Nave-Bayes is described in algorithm 1. The complexity of this algorithm depends on the number of classes C and the size of the number of tokens vocabulary W, which leads to a complexity O($\mid C \mid La$).

One way to optimize this algorithm is to modify the function $ExtractTermsFromDocuments(v, d)$ in such a way that group the tokens that share the same conditional probability. After that, each unique probability will be multiplied by the number of tokens that share each probability. That leads to reduce the complexity to $O(La + \mid C \mid Ma) = O(|C|Ma)$, in algorithm 2.

4 Next parts

Let's see the notebooks in the github repository.