Journal Title
XX(X):1–2
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

SuperBayesCat - A pairwise extention to the Naive Bayes Classifier

Dr Barry D.O. Adams[1]

Abstract

We introduce an extention to the Naive Bayes Classifier, in which remove the assumption that all document features (words or ngrams) are independent, based on the computering the pairwise correlation of words (ngram) in the training documents, and find formulae for the training and test phases. A python implementation is available on github [?] and tests are run using the Reutes21578 accuracy increases from 79.1% to 81.1%, when the classifier also included features pair of word adjacent to each other. We discuss the speed and accuracy of the implementation and its advantages and disadvantages.

Keywords

bayes classifier document classification supervised

Introduction

The Naive Bayes Classifier is a subset of Bayesian decision theory and can classify documents suprisingly well [1].

In a naive Bayes a the probability of each feature, (typical words or ngrams) in a document for each document class is computered in the training phase, using the Bayes Theorem . When new documents are presented, the class of the document, is found to be

$$c = \arg\max_{c} Pr(c).\prod_{i=1}^{m} Pr(x_i|c) \tag{1}$$

This work was self funded, Dr Adams is employed in a Back End Developer Role not related to this work

Corresponding author:
77 Twyford Abbey Road, London NW10 7ET
Email: barry.david.adams@gmail.com

We product is typical turned into a sum by taking the logarithm of the probability. It can be seen that speed of the classifier depends on the product of number of features in document and the number of document classes in the training set.

in this work we relax the (Naive) assumption that each document feature, words or ngrams, are independent, instead computing the correlation of the features, (hereafter refered to a words), in the training phase, and produce a new fomula for class of the documents.

$$c = \arg\max_{c} Pr(c) . \sum_{i=1}^{m} 2^{-\max_j(C_{ij})} \log(Pr(x_i|c)) \tag{2}$$

Where the correlation error C is defined in the training phase as

References

[1] Pouria Kaviani, Mrs Sunita Dhotre, *Short Survey on Naivce Bayes Algorithm*, Internation Journal of Advance Engineering and Research Development, (2017)