General Mathematics Vol. 14, No. 4 (2006), 135–138

An approach of the Naive Bayes classifier for the document classification [1]

Ioan Pop

Abstract

To perform the ranking document or the Web Mining tasks we have considered an approach based on the Naive Bayesian algorithm. The implementation of Naive Bayes algorithm was made in the PMML language.

2000 Mathematics Subject Classification: 60-04, 65C60
Keywords: Naive Bayesian classifier, document classification, Web Mining

1 Introduction

This article approaches the implementation of the Naive Bayes (shortly NB) classifier. It shows that the algorithm NB improves the tasks of the Web Mining by the accuracy documents classification. Its applications are important in the following areas: e-mail spamming; filtering spam results out of search queries; mining log files for computing system management; machine learning for Semantic Web; document ranking by text classification; hierarchical text categorization; managing content with automatic classification and other areas from Web Mining.

[1] *Received 19 August, 2007*
Accepted for publication (in revised form) 17 October, 2007

2 The NB probabilistic model

Abstractly, the probability model for a classifier is a conditional model $p(C|F_1, F_2, ..., F_n)$ over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. The Bayes′ theorem relates the conditional and marginal probabilities of stochastic events C, and F:

$$Pr(C|F) = \frac{Pr(F|C)Pr(C)}{Pr(F)} \tag{1}$$

where: $P(C)$ is the prior probability of hypothesis C; $P(F)$ is the prior probability of training data F; $P(C|F)$ is the probability of given F and; $P(F|C)$ is the probability of F given C. Using Bayes′ theorem for several feature variables F_n, we can rewrite this as:

$$p(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)}. \tag{2}$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features F_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model (1) which can be rewritten using repeated applications of the definition of conditional probability as:

$$p(C, F_1, ..., F_n) = \tag{3}$$

$$= p(C)p(F_1|C)p(F_2|C, F_1), p(F_3|C, F_1, F_2)...p(F_n|C, F_1, F_2, ..., F_{n-1})$$

This means: assuming that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$ and $p(F_i|C, F_j) = p(F_i|C)$ the model (1) can be expressed as:

$$p(C, F_1, ..., F_n) = p(C)p(F_1|C)p(F_2|C), ... = p(C)\prod_i p(F_i|C). \tag{4}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C can be expressed like this:

$$p(C, F_1, ..., F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C), \tag{5}$$

where Z is a scaling factor dependent only on $F_1, ..., F_n$, i.e., a constant if the values of the feature variables are known. The corresponding classifier for this model is the classify function defined as follows:

$$classify(f_1, ..., f_n) = argmax_c P(C = c) \prod_{i=1}^{n} p(F_i = f_i|C = c). \tag{6}$$

3 The NB model for the document classification

Consider the problem of classifying documents by their content, for example spam and non-spam E-mails. Imagine that documents are drawn from a number of classes of documents which can be modelled as sets of words where the (independent) probability that the $i-th$ word of a given document occurs in a document from class C can be written as $p(w_i|C)$. (Simply, we assume that the probability of a word in a document is independent of the length of a document, or that all documents are of the same length.) Then the probability of a given document D, given a class C, is

$$p(D|C) = \prod_i p(w_1|C). \tag{7}$$

Using the Bayesian result above, and the assuming that there are only two classes, S and $\neg S$ (e.g. spam and not spam) we can write:

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S)}{p\neg S} \prod_i \frac{p(w_i|S)}{p(w_i|\neg S)}. \tag{8}$$

Thus, the probability ratio $p(S|D)/p(\neg S|D)$ can be expressed in terms of a series of likelihood ratios. The actual probability $p(S|D)$ can be easily computed from $\log(p(S|D)/p(\neg S|D))$ based on the observation that

$p(S|D) + p(\neg S|D) = 1$. Taking the logarithm of all these ratios, we have:

$$\ln \frac{p(S|D)}{p(\neg S|D)} = \ln \frac{p(S)}{p\neg S} + \sum_i \ln \frac{p(w_i|S)}{p(w_i|\neg S)}. \tag{9}$$

Finally, the document can be classified as follows. It is spam if $\ln \frac{p(S|D)}{p(\neg S|D)} > 0$, otherwise it is not spam.

References

[1] I. Rish, *An empirical study of the naive Bayes classifier*, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. (available online: PDF, PostScript).

[2] M. Mozina, J. Demsar, M. Kattan, and B. Zupan, *Nomograms for Visualization of Naive Bayesian Classifier*, In Proc. of PKDD-2004, pages 337-348. (available online: PDF), 2004.

[3] O. R. Duda, P. E. Hart and D. G. Stork, *Pattern classification (2nd edition)*, Section 9.6.5, p. 487-489, Wiley, ISBN 0471056693,2000.

[4] PMML 3.0, http://www.dmg.org/pmml-v3-0.html

[5] PMML 3.1, http://www.dmg.org/pmml-v3-1.html

[6] PMML, http://www.sourceforge.net/projects/pmml

Department of Informatics, Faculty of Sciences
"Lucian Blaga" University of Sibiu
Str. Dr. I. Raţiu nr. 5-7, 550012 - Sibiu, România,
E-mail: *ioan.pop@ulbsibiu.ro*