Miłosz Staniszewski 220629

Artificial Intelligence and Knowledge Engineering Laboratory

1. Document Classification
2. **Introduction**

The main task of this assignment is to train classifier for assigning document to a particular category. The documents should be converted into a vector of word frequencies and submitted into a program specialized in classification. Then the texts would be classified by using one of the available classification methods.

1. **Text Preparation**

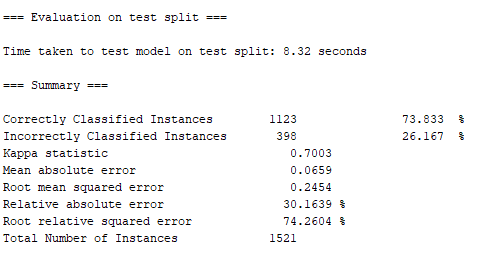
Before the classification, from available set of various text documents, taken from the 20 Newsgroups database, 8 different categories were chosen – Graphics, Hardware, For Sale, Motorcycles, Hockey, Space, Guns, Religion. Each category contains hundreds of documents related to the category.

All of those documents were adjusted by a Java program that deleted all headers and punctuation from the text in order to obtain pure words. Documents prepared in that way were converted into a vector of word frequencies by the classification tool, WEKA.

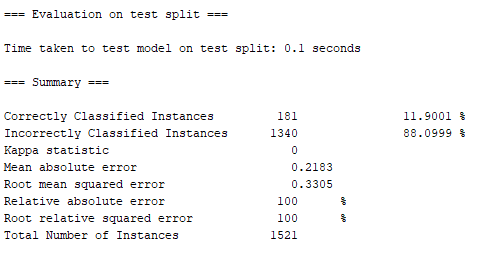
1. **Classification**

Documents were classified using two different methods of classification: Naïve Bayes and ZeroR. To keep the results unbiased, set of documents was divided – 66% of the documents were used as the training set, the remaining 34% were used as test set. The results are as follows:

Naïve Bayes



ZeroR



1. **Applied methods**

Naïve Bayes

Naïve Bayes classifier is based on the Bayes theorem and assumes independency between values of the various features. Even if the features depend on each other or upon the existence of the other features, all of the properties of an object independently contribute to the probability of an event.



ZeroR

ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. In order to use the algorithm it is required to construct a frequency table for the target and select its most frequent value.

1. **Conclusion**

Before classifying, it is important to remove all unimportant characters and headers from the document, as it can provide noise for the classifier that will modify the final result of the classification. ZeroR classifier correct classification percentage is so low due to the lack of predictions and the values being just the words assigned to the classes. Naïve Bayes classifier works much better, because of the probabilities of predictions that may be the correct ones.

1. **Sources**
2. C. Manning, P. Raghavan, H. Schütze, *An Introduction to Information Retrieval*, Cambridge, 2009
3. http://www.analyticsvidhya.com
4. <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>
5. <https://www.cs.waikato.ac.nz/ml/weka/>