Artificial Intelligence and Knowledge Engineering Laboratory

Task 4. Document classification

Authors: Maciej Piasecki

Task Objectives

Getting familiar with the representation of text documents as vectors of word frequencies. Learning basic methods for feature selection and Machine Learning algorithms for classification. Obtaining basic skills in using Weka environment for Machine Learning.
The main task is to train a classifier based on Machine Learning for recognising documents that belong to one of the selected categories on the basis of the content of the documents.

Subtasks

The minimal part:

1. Read chapter 13 from "Introduction to Information Retrieval" from (Manning et al., 2008), see the bibliography.
2. Select a toolkit for Machine Learning, e.g.
   a. Weka environment for Data Mining and Machine Learning: http://www.cs.waikato.ac.nz/ml/weka/
   b. or scikit-learn Machine Learning in Python, http://scikit-learn.org/stable/
3. Download the collection of news group documents (an newsgroups archive) called 20 Newsgroups: http://qwone.com/~jason/20Newsgroups/
4. Select for the experiments 8 different categories (if they have very different topics, it will be easier to obtain better results during the experiments).
5. Write a program for converting the main body of the newsgroups documents into the vectors of the word frequencies:
   doc_id, category, word1, frequency_of _w1, …, word1, frequency_of _w1
   The statistics should be computed only from the main body of the documents. Documents' headers and all kind of meta-data should be excluded. The headers can include the name of a group or a name of the author who can be very characteristic for the given group. So it would be fun to have such obvious clues included into the training-testing data.
6. Select k=10 000 words that are most frequent and occur in not too small number of documents.

7. Convert document vectors to the training-testing data format required (or suitable) for the selected ML toolkit, in which the attributes are words selected in the step 6 and the decision class is the document category.
8. Upload the training-testing file into the toolkit.
9. Select and test (in combination with the next step) at least two different methods for feature selection. It is good to create a separate data subset for the feature selection wich is different from both the training and testing subsets.
10. Choose at least two different ML algorithms for the construction of a classifier, e.g. you can start with a version of Naïve Bayes.
11. Evaluate the performance of different classifiers in 10-fold cross validation scheme for different parameter setting and different methods for feature selection.
12. Prepare a report: describe shortly in your own words, but enough comprehensively, the applied feature selection methods and ML algorithms, compare and analyse the results of the evaluation, analyse the error distribution and draw conclusions.

The extended part (for extra score points):

13. Test one more classifier that is based on a different Machine Learning algorithm than the first two.
14. Add to the report information about this classifier and the obtained results
15. Implementation of a feature weighting algorithm, e.g. tf.idf and testing the classifier on the weighted values of features instead of the raw frequencies – this step must be described in the report too.

Task rating

2 points – building vector representation
2 points – preparing the data for the ML toolkit and uploading the data to the ML toolkit
3 points – running feature selection and classification for different settings
3 points – writing report
Extra: up to 3 points

Bibliography

1. Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. *Introduction* to. *Information*. *Retrieval*. Cambridge University Press, 2008. (there will be also a copy in the Board):
http://www-nlp.stanford.edu/IR-book
or
https://archive.org/details/AnIntroductionToInformationRetrieval
or
http://www-connex.lip6.fr/~gallinar/livres%20-%20fichiers/2007-%20Manning-irbookonlinereading.pdf
2. Weka documentation: http://www.cs.waikato.ac.nz/ml/weka/documentation.html
3. Papers suggested in Weka for the selected classifier(s).