Artificial Intelligence and Knowledge Engineering Laboratory

Task 4. Document classification

Authors: Maciej Piasecki, Jan Kocoń

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.
We will focus on the two machine learning algorithms discussed during the lectures, namely

- Naive Bayesian Classifier (NB),
- Decision Trees (DT).

Subtasks

The obligatory part:

1. Read chapter 13 from "Introduction to Information Retrieval" – described in the bibliography.
2. Download Weka environment for Data Mining and Machine Learning: http://www.cs.waikato.ac.nz/ml/weka/ or any other environment for Machine Learning. Learn about functions of the selected environment and implementations of the machine learning algorithms pre-selected for this assignment.
3. Download the collection of news group documents (an newsgroups archive) called 20 Newsgroups: http://qwone.com/~jason/20Newsgroups/ (see **Additional Information** below).
4. Select for the experiments 5 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 newsgroups documents into the vectors of the word frequencies:
   doc_id, category, word1, frequency_of _w1, …, word1, frequency_of _w1
6. Select k=10 000 words that are most frequent and occur in not too small number of documents.

7. Convert document vectors to a format appropriate for the selected environment, e.g. ARFF in the case of Weka (all vectors in one file), in which the attributes are words selected in the step 6 and the decision class is the document category.
8. Upload the ARFF file into Weka system (or another selected environment).
9. Divide the created data into appropriate subsets according to the rules presented during lectures or in literature. In every case k-fold cross-validation is mandatory.
10. Plan experiments (on the basis of careful consideration of the problem and good understanding, this should not be done in a mechanical way) aimed at careful investigation of both algorithms in their different variants in application to the formulated task. A plan of experiments should be presented and explained in the report. At minimum, the following aspects should be taken into account:
    - influence of different parameters characteristic for the algorithms,
    - feature selection methods and their parameters,
    - discovering bad and strong aspects of the examined algorithms,
    - finding out sets of features that are characteristic for different classes of documents and what they tell us about these classes,
    - behaviour of classifiers for different subsets of the decision classes (it would be good to formulate earlier what do we expect to discover).
11. Configure the best possible text classification system on the basis of the pre-selected algorithms and what have been learned about the during the experiments. Present the system properties and performance in a convincing way in the report.
12. Decide about a method for feature selection.
13. Prepare a report: describe the classifier and the feature selection method in your own words, analyse the results of the evaluation, draw conclusions.

Additional Part:

14. Test one more classifier that is based on a different Machine Learning algorithm than the two pre-selected ones.
15. Analyse learning curve for the pre-selected Machine Learning algorithms on the basis of different amounts of training data.
16. Try to inflict overfitting on both pre-selected Machine Learning algorithms and analyse effects of this phenomenon.

Report

The report should include description of the undertaken decision together with their motivation, explanation and justification. Special attention should be given to planning the experiments, achieved results and formulated conclusions. The report does not need to be too long, but it should be written in a consice and informative way.

Additional information

The 20NewsGroups collection can be easily obtained, e.g., by using the Python *sklearn* package (it would not be also difficult to write a simple program to directly read from the downloaded collection):

```
>>> from sklearn.datasets import fetch_20newsgroups
>>> newsgroups_train = fetch_20newsgroups(subset='train')

>>> from pprint import pprint

>>> pprint(list(newsgroups_train.target_names))
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']
```

The real data lies in the filenames and target attributes. The target attribute is the integer index of the category:

```
>>> newsgroups_train.filenames.shape
(11314,)
>>> newsgroups_train.target.shape
(11314,)
>>> newsgroups_train.target[:10]
array([12,  6,  9,  8,  6,  7,  9,  2, 13, 19])
```

It is important to remove the header/footer from the texts in the dataset so that the classifier learns to recognize categories solely from the content of the message (red parts below):

```
>>> newsgroups_train.data[10]
From: irwin@cmptrc.lonestar.org (Irwin Arnstein)
Subject: Re: Recommendation on Duc
Summary: What's it worth?
Distribution: usa
Expires: Sat, 1 May 1993 05:00:00 GMT
Organization: CompuTrac Inc., Richardson TX
Keywords: Ducati, GTS, How much?
Lines: 13
```

```
I have a line on a Ducati 900GTS 1978 model with 17k on the clock.  Runs
very well, paint is the bronze/brown/orange faded out, leaks a bit of oil
and pops out of 1st with hard accel.  The shop will fix trans and oil
leak.  They sold the bike to the 1 and only owner.  They want $3495, and
I am thinking more like $3K.  Any opinions out there?  Please email me.
Thanks.  It would be a nice stable mate to the Beemer.  Then I'll get
a jap bike and call myself Axis Motors!
--
-----------------------------------------------------------------------
"Tuba" (Irwin)      "I honk therefore I am"     CompuTrac-Richardson,Tx
irwin@cmptrc.lonestar.org    DoD #0826          (R75/6)
-----------------------------------------------------------------------
```

Task rating

2 points – designing representation of data, loading data and generation of the required format
1 point – designing and setting up a system for Machine Learning.
1 point – dividing data (and their conversion whenever needed) for the needs of experimental research.
4 points – planning and conducting experiments followed by preparing an analysis of their results in a report.
2 points – setting up the best possible solution for classifying the specified dataset of documents with the help of the classification algorithms defined in the assignment.

Extra Task up to 5 points:

- performing works described in Additional Part.

Bibliography

1. Christopher D. Manning. Prabhakar Raghavan. Hinrich Schütze. *Introduction* to. *Information*. *Retrieval*. Cambridge University Press, 2008.:
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).