CSI4106: Introduction to Artificial Intelligence
Fall 2017

Assignment 3

Handed in on: November 22nd, 2017
Due on: December 6th, 2017 (11h55 pm)

Learning objective: Experiment supervised learning using WEKA on textual data

Requirements: The assignment MUST be done in a group of two.

Preparation:

1. Download WEKA and its documentation from http://www.cs.waikato.ac.nz/~ml/weka/
2. Watch the video: https://www.youtube.com/watch?v=IY29uC4uem8

This video is a tutorial that you must follow stop by step to perform your experiments. This assignment requires that you learn by yourself how to manipulate WEKA. Be proactive. Read WEKA manual. Google. This is a self-learning and learning by doing experience!

Sentiment analysis based on text

Sentiment Analysis (and/or Opinion Mining) is one of the hottest topics in Natural Language Processing nowadays. The task consists of determining the polarity (the sentiment) of a text as positive or negative. In this assignment, you will be comparing the performance of 5 types of classifiers on a collections of movie-review documents labeled with respect to their overall sentiment polarity (positive or negative):

1. ZeroR (your baseline, you must definitely beat it)
2. A decision tree (J48)
3. K-Nearest neighbour (IBk)
4. Naïve Bayes (NaiveBayes)
5. Support Vector Machines (SVM)

(The words in parenthesis correspond to the names of these classifiers in WEKA).

You will use the `movie_review_polarity.arff` dataset.

Tasks

1. Load the arff file in WEKA
2. Following the YouTube tutorial video, you have to create **a bag of word model** that uses the WEKA StringToWordVector Filter to convert text into vectors.

3. Run the classifiers indicated above with the following configurations:
   a. Build a **Boolean unigram bag of words** model with a WordTokenizer;
   b. Build an **N-grams bag of words model** (use the NGrams Tokenizer - unigrams, bigrams and trigrams must be considered) associated to the following parameters: TF-IDF transformations + output word counts + stemming + min Frequency 5 + word count 1000 ;
   c. Build an N-grams model with your own configuration, which must improve the result of the previous experiments. Explain your configuration and the techniques used. (HINT: watch the YouTube video!).
   Note that simply changing the machine learning algorithm will not be considered sufficient. You must either perform a **feature selection** technique, and/or try **additional/other features** to improve the performance of the classifiers. Your result should be at least >=0.85. The best model obtained by the class will obtain the full mark for this part of the assignment.

   For all your models, use **10-Folds cross validation**.
   Save your models using the convention: classifierName_ExpNumber. E.g. J48_Exp_a. You should end up with 3 (configurations)*5 (classifiers) = 15 models.
4. Report
   a. Explain briefly and in your own words the principle behind each of the classifiers used in this experiment
   b. Explain in detail the configuration and the techniques used to obtain your best result (experiment c) such as the algorithm, its parameters, the feature selection techniques, and the additional features if applicable, etc.
   c. Report the results of all your classifiers in a table:

Machine learning classifier	**Experiment a**	**Experiment b**	**Experiment c**
Zero R - Accuracy			
J48 - Accuracy			
Etc.			

   d. Discuss your results. For example, you might want to reflect on the following questions: Does any of the configurations consistently improve the results across classifiers? Is one classifier generally better than another for this task? How about the errors? Could you find another configuration that improves the result? Any other aspect you can think of.

You must submit a zip file *yourname1_yourname2_A3.zip* containing

a) Your arff files after your transformations
b) Your machine learning models for each configuration – we must be able to load your models and run them again on your arff files. Carefully check that point before submitting your work
c) A **report** entitled *yourname1_yourname2_A3.pdf* that provides answers to the questions and requirements of the assignment.

Appendix

Using WEKA is simple. Everything can be done with the graphical interface included in the package (you can also call it from Java but this is not required in the assignment). When you run WEKA, you are given the choice of four user modes. I suggest that you use the mode "explorer" to start the graphical interface.

You will see several tabs, including "preprocess" and "classify". The "preprocess" tab is used to open your data file (already in .arff format) and choose filters and clean your data. The "classify" tab (see picture below) will allow you to choose the algorithm, its parameters and the evaluation method. Click on "Choose" to choose a classification algorithm.

To modify the parameters of your algorithm, select the algorithm first and click "more" in order to obtain information about the algorithm and its parameters. Try to optimize the performance of your classifiers by modifying the parameters appropriately.

Choose 10-fold cross-validation as your evaluation method in the « test options » section on the left, and click on the "Start" button. The results will appear on the right. You can save your models by right clicking on it in the box below start. You can also visualize more details about the output of your model (for example, you can see the obtained decision tree).