ELN 2016
Homework 2

Naïve Bayes Classifier

Download the Movie Review dataset from http://www.cs.cornell.edu/People/pabo/movie-review-data/review_polarity.tar.gz

It contains 1000 positive and 1000 negative reviews, preprocessed and tokenized. The directory neg contains the negative reviews, the directory pos contains the positive reviews.

The collection is split into 10 folds: fold 1 consists of documents in files cv000* to cv099*, fold 2 those in files cv100* to cv199*, etc.

Create a Naïve Bayes classifier. It should consist of two parts:

a) Training: the program takes two lists of files: one containing the positive review files, the other containing the negative ones, and outputs a model file.
b) Testing and evaluation: the program takes a model file and two lists of files: positive and negative review lists, and outputs the classification accuracy for the test set, in terms of accuracy, precision, recall and F1 measure.

Test the classifier on the Movie Review dataset, using fold 10 for testing and the rest for training.

Explore the technique of cross validation, i.e. perform 10 runs of the classifier, using a different fold for testing and compute the average accuracy, precision, recall and F1 macro-average.

Optional

Experiment using the Maximum Entropy classifier from NLTK (`nltk.classify.MaxentClassifier`) and compare the results.

The train data for the classifier are pairs of (features, class). In this case the features are binary features corresponding to the words present in a document (bag of words). For example:

```
train = [
     (dict(I=1,like=1,the=1,movie=1), 'pos'),
     (dict(A=1,terrible=1,movie=1), 'neg'),
      ...]
```