Machine Learning 2003

Assignment 4: Text classification using the naive Bayesian classifier.

- Due by: Part A: 23h59 on 3/04/2003; Part B: 23h59 on 10/04/2003.
- Remember that discussion and conclusions are the most important part of an assignment!
- Note: this is **not** a team project. Your implementation and experiments should be your own work.

1 Part A: Getting acquainted with Bayesian classifiers

Implement the naive Bayesian classifier and test it using the PlayTennis data set in Table 3.2 (p. 59 in Mitchell) as training data. Make up a few unseen test examples of your own to test the algorithm's ability to generalise. Write a short report, giving the learned parameters for this data set and the classification of each training and test example. Does data sparsity seem to have an effect on the performance of the algorithm? Discuss the relative suitability of decision trees and Bayesian classifiers for this problem.

Warning: Part B is more work and counts more than part A, so you'll need to finish the warm-up excercise (part A) early.

2 Part B: Application to text classification

Classification of newsgroup postings: Choose 100, 200, 300, 400, and 500 articles from each of the following 5 newsgroups: comp.ai, comp.ai.neural-nets, comp.theory, comp.graphics, and comp.windows.x to train a Bayesian classifier using the algorithm on page 183. Choose another set articles from the same 5 newsgroups and predict which newsgroup they were posted in using the trained Bayesian classifier.

2.1 Deliverables

1. Investigate whether the prediction improves as the size of the training set increases and show the learning curve.

2. A program demonstration where your best classifier is given a set of newsgroup postings (without header information!) which it has to classify as belonging to one of the 5 newsgroups.