10-602: Statistical Approaches to Learning and Discovery

Problem Set 3
Date: February 19, 2001 **Due: start of class February 27, 2001**

This problem set examines the models for text documents described in the paper "A Comparison of Event Models for Naive Bayes Text Classification," by Andrew McCallum and Kamal Nigam, AAAI-98 Workshop on "Learning for Text Categorization", `http://www.cs.cmu.edu/~mccallum/papers/multinomial-aaai98w.ps`.

The problem set involves answering questions about the models as well as implementing the models using a software package of your choice (`matlab` is recommended). The dataset you will use, available on the course web page, is a subset of the "20 newsgroups" dataset at the CMU data archive: `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-3/www/`. Only the articles from "talk.politics.misc" and "talk.religion.misc" are used, giving a two class problem. Note that the categorization was made without inspecting the actual content of the articles, so some outliers can be expected.

From the course Web page you can get the file `politics`, where each line has three numbers (d, w, c): c is the count of word w in document d. The file `religion` has the same thing for the religion class. The file `tpm_trm.mat` has the same information as matrices in `matlab` format. Words are numbered according to the `words` file. If you are using C for this problem set, the file `gammaln.c` might be helpful.

1. What are the two models compared in the paper? What estimation technique are the authors using in their 'Bayes-optimal' estimators? What is their prior?

2. The paper's results show that feature selection, i.e. selectively ignoring some of the data, can lower test error, especially with the multivariate Bernoulli model. This problem explores some reasons why this might be so.

   (a) Give a set of parameters for the multivariate Bernoulli model which have the effect of ignoring a particular word during classification. Do the same for the multinomial model.

   (b) Consider the following scenario: class 2 has exactly the same documents as class 1, except that the word "very" has been added in various places, e.g. "big" becomes "very big." For this classification task, what are the true multivariate Bernoulli parameters? What are the "ignored words" for these parameters? What are the true multinomial parameters? What words will the multinomial model ignore?

   (c) Suppose a word w always occurs in every document in both classes. The training set has a different number of documents in each class. If we use the "Bayes-optimal" estimator on the training set, will the multivariate Bernoulli model ignore this word? Explain.

(d) Generalizing from the last two questions, are parameter sets which ignore words likely to be the true parameters in practice? Are such parameter sets likely to arise from the training procedure? When should we expect feature selection to help?

3. Is there a way to statistically justify the paper's way of evaluating the quality of a feature? What about in the multinomial case? Use formulas from class to improve their evaluation measure. Hint: Are asymptotic expansions valid for text counts on a large vocabulary?

4. Based on your improved measure in the last problem, implement a routine which evaluates an individual word as a feature for classification. There should be one routine for the multivariate Bernoulli model and another for the multinomial model. According to each model, what are the top ten words for the politics/religion dataset?

5. Give a third model which you would like to see compared with the two models in the paper. It should be a probabilistic model relating the document content to the class index, as in the paper. How would it be used for classification purposes? Explain how your model uses the word counts differently from the models in the paper and why/how you might expect your model to do a better job of classifying documents.