CSCI 1950-F Homework 1: Naive Bayes

Brown University, Spring 2011

Homework due at 12:00pm on February 10, 2011

The Nursery database records a series of admission decisions to a nursery in Ljubljana, Slovenia. We downloaded this data from `http://archive.ics.uci.edu/ml/datasets/Nursery`, which you can see for more details if you're interested.

The database contains one tuple for each admission decision. The features or attributes include financial status of the parents, the number of other children in the house, etc. The first three tuples in the dataset are as follows:

```
usual,proper,complete,1,convenient,convenient,nonprob,recommended,recommend
usual,proper,complete,1,convenient,convenient,nonprob,priority,priority
usual,proper,complete,1,convenient,convenient,nonprob,not_recom,not_recom
```

where the first 8 values are features or attributes $\mathbf{x}_i$, and the 9th value is the class assigned y_i (i.e., the admission decision recommendation).

Your job is to build a Naive Bayes classifier that will make admission recommendations. The file `/course/cs195f/asgn/naive_bayes/handout/nursery/nursery.mat` contains this data pre-processed in a matrix format that Matlab can directly read. All of the symbols have been replaced with identifying integers. The first three rows of this matrix are:

```
>> load('/course/cs195f/asgn/naive_bayes/handout/nursery/nursery.mat');
>> data(1:3,:)
ans =
     1     1     1     1     1     1     1     1     2
     1     1     1     1     1     1     1     2     4
     1     1     1     1     1     1     1     3     1
```

You should divide this data into equal-sized training and testing data sets as follows (the `reset` ensures that we'll all use the same training/test split).

```
load('/course/cs195f/asgn/naive_bayes/handout/nursery/nursery.mat');
reset(RandStream.getDefaultStream)
data = data(randperm(size(data,1)),:);
train = data(1:size(data,1)/2,:);
test = data(size(data,1)/2+1:end,:);
```

Question 1:

Give an equation for the joint log-likelihood of a Naive Bayes model for this data, defining parameters as appropriate. Specify the equations for maximum likelihood (ML) estimation of the model parameters from `train`*, implement a parameter estimation algorithm based on these equations, and use it to predict the most likely labels (admission recommendations) for the* `test` *data. Report the accuracy of your classifier, and submit your code.*

Question 2:

Modify the Nursery data by duplicating the last attribute 20 more times. You can do this by executing

```
data = [data(:,1:end-1),repmat(data(:,end-1),1,20),data(:,end)];
```

before splitting into `train` *and* `test`*. Then run your ML Naive Bayes estimator on the new training data, and evaluate it on the new testing data. Explain in words why you see the change in accuracy that you observe.*

Question 3:

Using the original data and the model you estimated in Question 1, calculate the (joint) log-likelihood $\sum_i \log p(\mathbf{x}_i, y_i \mid \theta)$ *of the training data. Using this model, is it possible to calculate the (joint) log-likelihood of the test data? Explain your answer.*

Question 4:

Now use a Bayesian estimator with a uniform prior (i.e., a flat Dirichlet prior with parameters $\alpha_k = 1$ *for all dimensions* k*) to estimate all of the multinomial distributions in your Naive Bayes model from the original training data. Specify the equations for estimation of the model parameters via their posterior mean (see MLaPP Sec. 4.5.2-4.5.3), and implement a corresponding parameter estimation algorithm. What accuracy do you obtain on the test data using this model?*

Question 5:

Calculate the (joint) log-likelihood of the training data under the model estimated via the Bayesian approach of Question 4. Is it higher or lower than the log-likelihood of the training data under the ML model? Explain your answer. Now calculate the log-likelihood of the test data given these Bayesian parameter estimates. Explain why you don't run into the same problems you encountered in Question 3.