CS195f Homework 1: Naive Bayes

Mark Johnson and Erik Sudderth

Homework due at 2pm, 24th September 2009

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 and the 9th value is the class assigned (i.e., the admission decision recommendation).

Your job is to build a Naive Bayes classifier that will make admission recommendations. Luckily the really hard work of data preparation has been done for you by Deqing, our fearless TA. The file `/course/cs195f/asgn/naive_bayes/handout/nursery/nursery.mat` contains this 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:

Estimate a Naive Bayes model using Maximum Likelihood (ML) on `train` *and use it to predict the y labels of 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 $\sum_i \log \mathrm{P}(\mathbf{x}_i, y_i|\theta)$ *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 MAP estimator with a flat Dirichlet prior (i.e., $\alpha = 1$*) to estimate all of the multinomial distributions in your Naive Bayes model from the original training data. What accuracy do you obtain on the test data using this model? Calculate the (joint) likelihood of the training data under the MAP model. Is it higher or lower than the likelihood of the training data under the ML model? Explain your answer. Now calculate the likelihood of the test data under the MAP model. Explain why you don't run into the same problems you encountered in Question 3.*

Question 5:

In this question you'll produce an ROC curve for your Naive Bayes model. We'll do this for the class label $y^\star = 4$*. First, explain how to calculate* $\mathrm{P}(Y|\mathbf{X})$ *in a Naive Bayes model (this should take one line). Then calculate an ROC curve for class label* $y^\star = 4$ *using your Naive Bayes model estimated with the Bayesian MAP estimator from the last question. For each value of c in 0, 0.01, 0.02, . . . , 1, calculate the sensitivity and specificity of a classifier that predicts* $Y = 4$ *whenever* $\mathrm{P}(Y = 4|\mathbf{X}) > c$*, and plot the resulting ROC curve. Please submit this as a pdf file. Using your ROC curve, roughly estimate the specificity of your classifier when the sensitivity is 0.8?*

The last question concerns the Iris data set collected by R.A. Fisher, whose pioneering work in statistics about a century ago still influences the field today. This data is about classifying flowers into one of three possible kinds of Iris; we downloaded it from

`http://archive.ics.uci.edu/ml/datasets/Iris`. The first three lines of this data set are:

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
```

The first four components are various length measurements (in centimeters). Again, Deqing has converted this into Matlab matrix format for us. The corresponding lines of the data are:

```
>> data(1:3,:)

ans =

    5.1000    3.5000    1.4000    0.2000    1.0000
    4.9000    3.0000    1.4000    0.2000    1.0000
    4.7000    3.2000    1.3000    0.2000    1.0000
```

You should load the data and split it into train and test portions as follows:

```
load('/course/cs195f/asgn/naive_bayes/handout/iris/iris.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 6:

Estimate a Naive Bayes model with Gaussian features from the Iris `train` *data using Maximum Likelihood, and evaluate this model on* `test`*. Report the accuracy of your classifier on the* `test` *data, and submit your code. As in question 5, also plot and turn in an ROC curve for class label* $y^\star = 3$*. Hint: you may find the Matlab* `mean` *and* `var` *functions useful.*

Question 7: (200-level credit)

For this question you'll derive the Maximum Likelihood estimates of the mean μ *and the variance* σ^2 *of a Normal distribution given data* $D = (x_1, \ldots, x_n)$*.*

1. *Give the log likelihood of* D*. You can ignore terms that do not involve* μ *or* σ*.*
2. *Give the derivative of the log likelihood with respect to* μ*.*
3. *Solve for* μ *by setting this derivative to zero. Give your solution for* μ*.*
4. *Give the derivative of the log likelihood with respect to* σ*.*
5. *Solve for* σ *by setting this derivative to zero. Give your solution for* σ^2*.*

Hand-in procedure

You should do the following while logging in a CS department machine. Put all the files you want to submit in a folder and `cd` to that directory. Create a `README` file that describes the various parts of your submission (e.g., which files contain the various ROC curves). This `README` file is required even if it contains nothing. Then at command line, type:

```
/course/cs195f/bin/cs195f-handin hw1
```

Your CS account will receive a confirmation email after successful handin.