MLTA Tutorial 2: Bayesian Inference

David Barber
Department of Computer Science, University College London

1 Setting up

To get the MATLAB code for this tutorial, please go to

`http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.Software`

and download

`http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/290313oo.zip`

You'll then need to unzip the file in a local directory. Open
`matlab` and type

```
>> cd localdir
```

where `localdir` is the name of the directory where you downloaded the software. Now type

```
>> setup
```

This will initialise the files and paths and enable you to run the demos.

2 Naive Bayes

Whizzco decide to make a text classifier. To begin with they attempt to classify documents as either sport or politics. They decide to represent each document as a (row) vector of attributes describing the presence or absence of words.

$$\mathbf{x} = (\text{goal}, \text{football}, \text{golf}, \text{defence}, \text{offence}, \text{wicket}, \text{office}, \text{strategy})$$

Training data from sport documents and from politics documents is represented below in MATLAB using a matrix in which each row represents the 8 attributes.

```
xP=[1 0 1 1 1 0 1 1;  % Politics
    0 0 0 1 0 0 1 1;
    1 0 0 1 1 0 1 0;
    0 1 0 0 1 1 0 1;
    0 0 0 1 1 0 1 1;
    0 0 0 1 1 0 0 1]
```

```
xS=[1 1 0 0 0 0 0 0; % Sport
    0 0 1 0 0 0 0 0;
    1 1 0 1 0 0 0 0;
    1 1 0 1 0 0 0 1;
    1 1 0 1 1 0 0 0;
    0 0 0 1 0 1 0 0;
    1 1 1 1 1 0 1 0]
```

Using a maximum likelihood naive Bayes classifier, what is the probability that the document $\mathbf{x} = (1, 0, 0, 1, 1, 1, 1, 0)$ is about politics?

3 EM for Chest Clinic

From MATLAB, type

```
>> demoEMchestclinic
```

This shows how to learn the tables of the Chest Clinic Belief network (from tutorial 1) using the Expectation Maximisation (EM) algorithm. There are some parameters of the tables, from which we can draw samples from the belief network. We now pretend we don't know the table parameters and try to learn them based only on the observed data (which has some entries missing). We can then compare how close the learned tables are to the true tables.

Type

```
>> edit demoEMchestclinic.m
```

and adjust line 8 to draw now 150 samples and rerun the demo. What do you think will happen to the accuracy of learning the tables as the number of samples increases?

4 Fitting a Mixture Model to Digits

From MATLAB, type

```
>> demoMixBernoulliDigits
```

This runs a demo in which we use the EM algorithm to fit a mixture model to a set of handwritten digits.

Adjust the code (line 23) to use a larger number of mixture components, by setting $H = 30$; similarly run the code with $H = 5$. Can you explain your results?

5 Gaussian Mixture Model

From MATLAB, type

```
>> demoGMMem
```

Which runs a demo of fitting a Gaussian Mixture model for two-dimensional data.

5.1 Fitting the digits again

Now we will try to cluster the digit data again, but this time using the GMM.

First get the digit data. From MATLAB, type:

```
import brml.*; v=[];
for d=0:9
    load(['digit',num2str(d),'.mat'])
    xx=x';
    v=[v xx(:,1:500)];
end
```

In this case the digits are greyscale, with values from 0 to 255. We will treat these as 'continuous' values and use EM to train the GMM:

```
>> opts.plotlik=1; opts.plotsolution=1; opts.maxit=10; opts.minDeterminant=0.0001;
>> figure; [P,m,S,loglik,phgn]=GMMem(v,10,opts);
```

Now plot the solution:

```
>> figure; for i=1:10; subplot(1,10,i); imagesc(reshape(m(:,i),28,28)'); end
```