University of York

MLAP

Machine Learning and Applications

Open Examination

Examination number:

Y3606797

Contents

1 Task 1

1.1 What sort of model?

The naive Bayes model makes the assumption that all features are conditionally independent given the class. The given Bayesian network explicitly models the conditional independence in it's structure.

1.2 EM in general

The EM algorithm is an iterative approach to maximising the likelihood when we have hidden variables. We have a model $p(\upsilon, h|\theta)$ and we wish to find θ by maximising the marginal likelihood $p(\upsilon|\theta)$. EM achieves this by replacing the marginal likelihood with a lower bound, the bound depends on θ and the set of variational distributions $\{q\}$, we optimise this bound. This is done by first fixing θ and then optimising $w.r.t$ $\{q\}$, then fixing $\{q\}$ and optimising $w.r.t$ θ. These are known as the 'E' and 'M' steps and are repeated until convergence. The EM algorithm may not always return the correct MLEs because it can get stuck on local optima.

1.3 EM for the current model

Firstly we assume some initial parameters for the distribution, so we have θ^0_{X1}, θ^0_{X2}, θ^0_{X3}, θ^0_H and θ^0_C. Then we compute the distribution for the hidden data H:

$$q^{n=1}_{t=1}(H) = p(H|C=0, X1=0, X2=0, X3=1, \theta^0),$$

$$q^{n=2}_{t=1}(H) = p(H|C=1, X1=1, X2=0, X3=0, \theta^0)...$$

and so on for each data point in the training data. This is first E-step in the algorithm.

The first M-step involves maximising the energy term by choosing new values for θ. The energy term is given by the following equation:

$$E(\theta) = \sum_{n=1}^{N} \{\langle \log p(C^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X1^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X2^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X3^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(H^n)\rangle_{q^n_t(H)}\} \quad (1)$$

To calculate the M-step update for each table we have to maximise each term of the above equation individually. The contribution of the term $p(C = i|H = j)$ is given by:

$$\sum_n \mathbb{I}[C^n = i] q^n(H = j) \log p(C = i|H = j))$$

where the indicator function $\mathbb{I}[C^n = i]$ equals 1 if C is in state i and is zero otherwise. We normalise the table by adding a Lagrange term:

$$\sum_n \mathbb{I}[C^n = i] q^n(H = j) \log p(C = i|H = j)) + \lambda\{1 - \sum_k p(C = k|H = j)\}$$

We differentiate with respect to $p(C = i|H = j)$ and equate to zero to find the maximum, this gives us:

$$\sum_n \mathbb{I}[C^n = i] \frac{q^n(H = j)}{p(C = i|H = j)} = \lambda$$

Hence

$$p(C = i|H = j) = \frac{\sum_n \mathbb{I}[C^n = i] q^n(H = j)}{\sum_{n,k} \mathbb{I}[C^n = k] q^n(H = j)}$$

This result highlights a relationship between the table updates and the standard MLE. That is, when we have missing data we replace the functions such as $\mathbb{I}[C^n = i]$ which represent the real counts of the data, by the assumed distributions q. We use this relationship to calculate the table updates for each other term in the energy equation.

We can then perform the E-step again using our updated parameters such that we have new assumed distributions for the hidden variable:

$$q_t^{n=1}(H) = p(H|C = 0, X1 = 0, X2 = 0, X3 = 1, \theta^{t-1}),$$

$$q_t^{n=2}(H) = p(H|C = 1, X1 = 1, X2 = 0, X3 = 0, \theta^{t-1})...$$

We perform the steps E and M iteratively until the algorithm converges to a local optima.

1.4 Gibbs sampling for the current model

We start by initially setting the 'state' of the variables to some values, usually according to their prior distributions. Then we pick a variable, C for example, and sample from it's conditional distribution.

From Bayes theorem:

$$P(X1, X2, X3, C, H) = P(C|X1, X2, X3, H)P(X1, X2, X3, H)$$

We could sample from the following distribution:

$$P(C|X1, X2, X3, H)$$

But we only need to take into account the values of the variables within the currently sampled variables Markov blanket, as defined by the nodes in the BN which are the parent, child or co-parents of the current node.

Hence we can sample from: $P(C|H)/Z$ where Z is a normalisation factor. We update the variable value to the sample in a new instance of the 'state' and we repeat for the remaining variables.

The theory of MCMC guarantees that the stationary distribution of the samples generated by Gibbs sampling will be an approximation of the joint posterior distribution for a large enough number of samples.

1.5 Running Gibbs sampling for the current model

I ran the sampler for 1000 iterations as a burn in, this was done to essentially discard initial samples that may not be representative of the stationary distribution. I then ran around 500000 iterations to converge on the distribution.

The kernel density distribution in fig. 1 represents the approximate distribution for the hidden variable H. The shape resembles the uniform distribution which means that the two possible values for H are approximately equally probable.

1.5.1 Compare contrast EM/Gibbs sampling

A downside to Gibbs sampling is that the samples are very dependent given the previous samples, due to the nature of MCMC. A downside of EM is that is can get stuck on local optima and not return the true maximum value for the parameters. Similarly for Gibbs sampling it is unknown how many iterations are required for convergence.

In the Bayesian approach we also have the option to incorporate informative priors into our learning, which increases the accuracy of our results, but can also complicate them. Both approaches are fairly easy computationally which is why they are popular methods for inference problems.

parameter	mean	sd
h.theta[1]	0.4991	0.2907
h.theta[2]	0.5009	0.2907
c.theta[1,1]	0.5008	0.2674
c.theta[1,2]	0.4992	0.2674
c.theta[2,1]	0.4984	0.2671
c.theta[2,2]	0.5016	0.2671
x1.theta[1,1]	0.5015	0.2671
x1.theta[1,2]	0.4985	0.2671
x1.theta[2,1]	0.4985	0.2668
x1.theta[2,2]	0.5015	0.2668
x2.theta[1,1]	0.708	0.2445
x2.theta[1,2]	0.292	0.2445
x2.theta[2,1]	0.7088	0.244
x2.theta[2,2]	0.2912	0.244
x3.theta[1,1]	0.708	0.2445
x3.theta[1,2]	0.292	0.2445
x3.theta[2,1]	0.709	0.2445
x3.theta[2,2]	0.291	0.2445

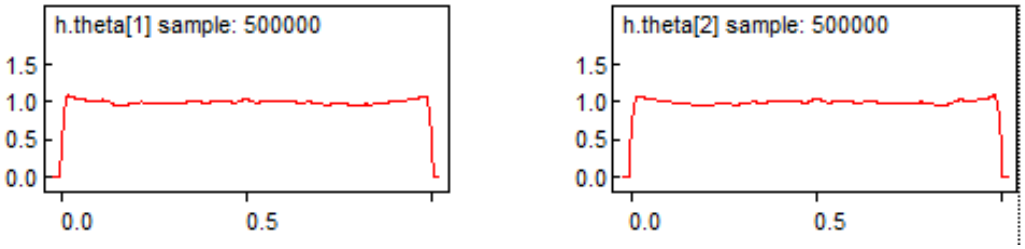

Figure 1: Approximate posterior distribution for H

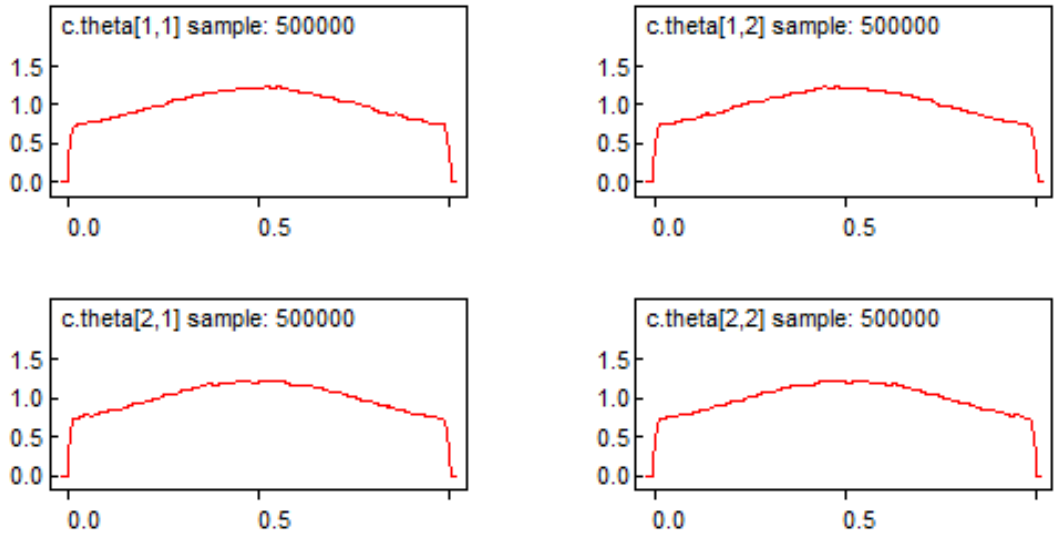

Figure 2: Approximate posterior distribution for C

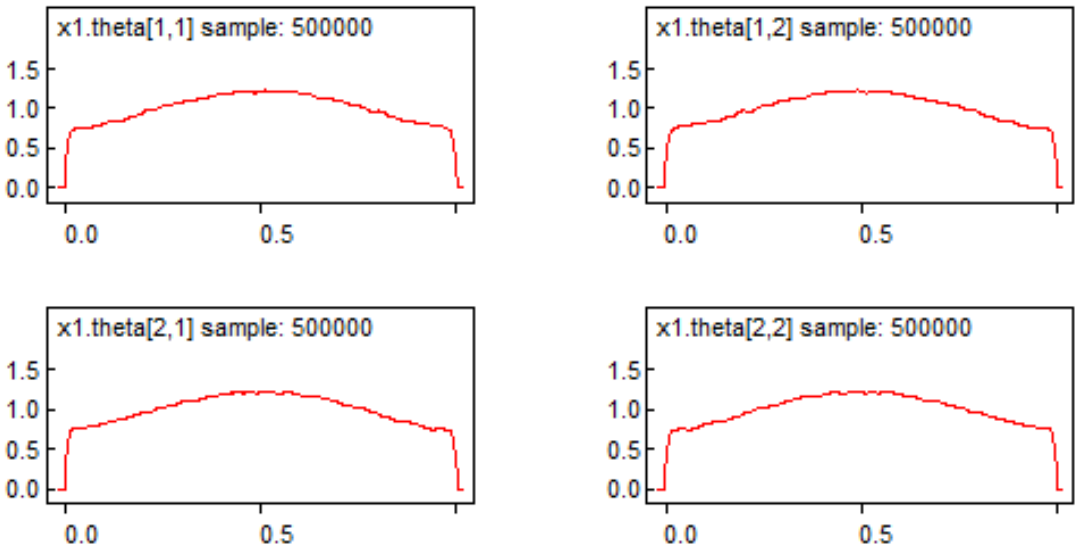

Figure 3: Approximate posterior distribution for X1

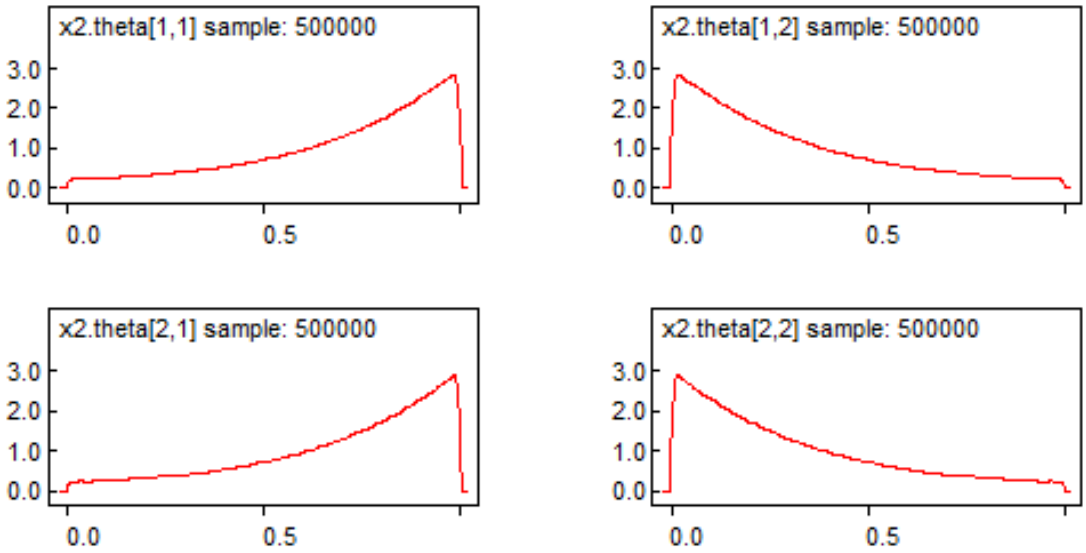

Figure 4: Approximate posterior distribution for X2

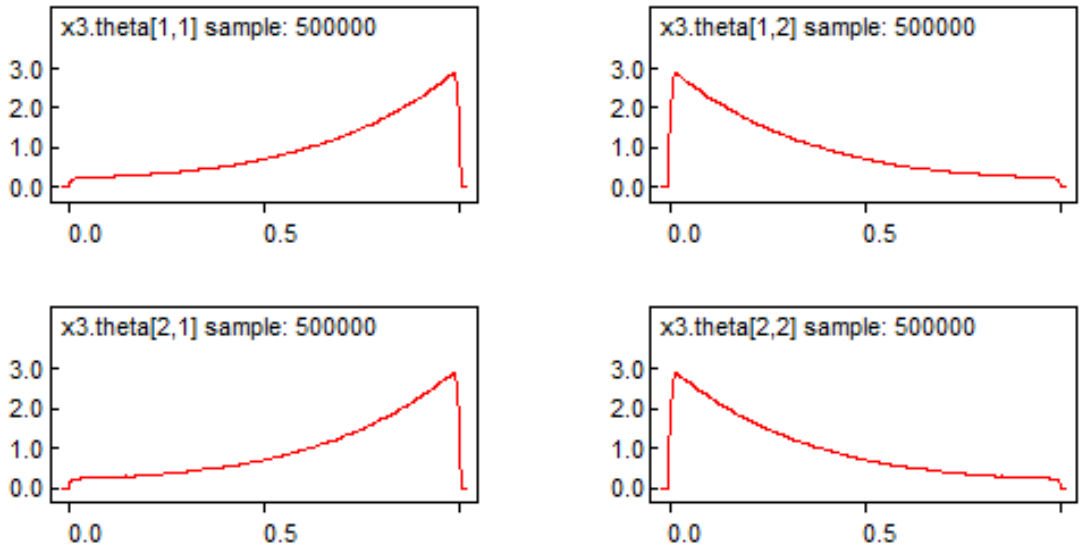

Figure 5: Approximate posterior distribution for X3

1.6 Using informative priors

I edited the prior for variables H and C to express that they were both highly likely to be the same value, the result of this can be seen in fig. 6 and fig. 7.

This change resulted in completely different shapes for the estimate distributions of the posteriors for variables X1, X2 and X3 as seen in fig. 8, fig. 9 and fig. 10. H and C now obviously tend towards the more probable values for the variables as defined by my chosen prior.

The values of the variables X1, X2, and X3 now seem to be more evenly distributed for the improbable value of H, they resemble uniform distributions (slightly skewed for X2 and X3). The values of X1, X2 and X3 seem to remain biased against the probable value of H under these prior assumptions.

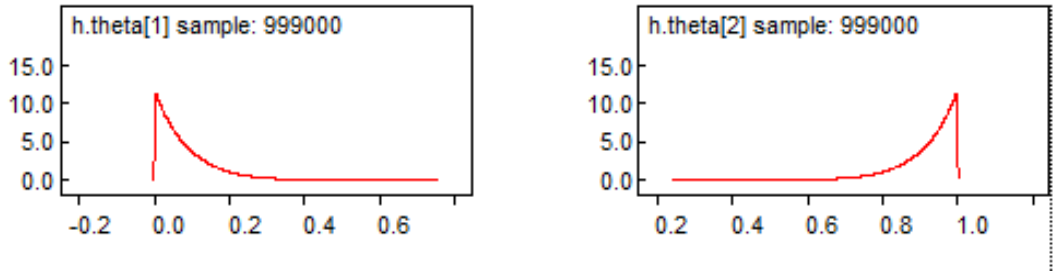

Figure 6: Approximate posterior distribution for H with informative prior

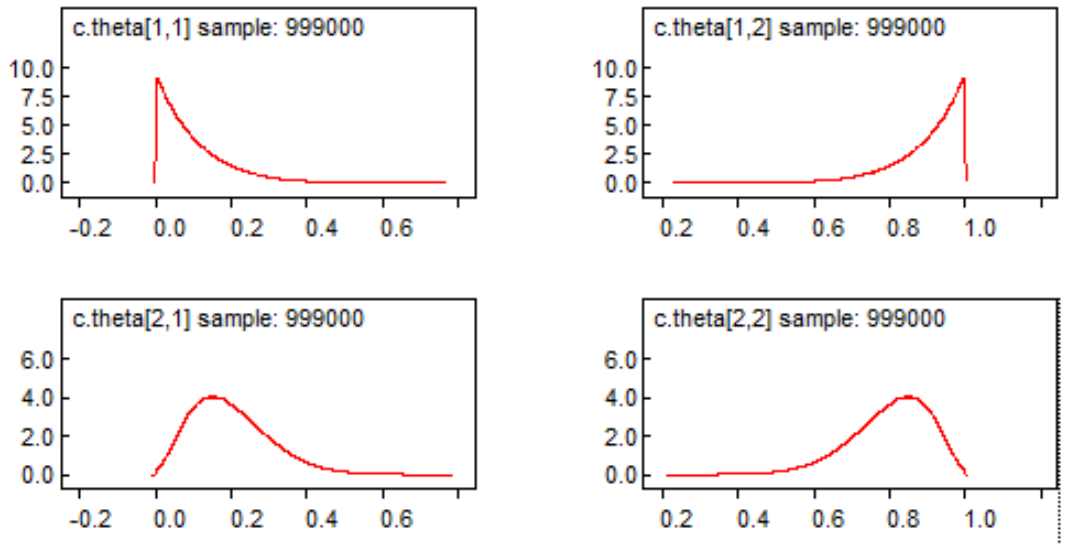

Figure 7: Approximate posterior distribution for C with informative prior

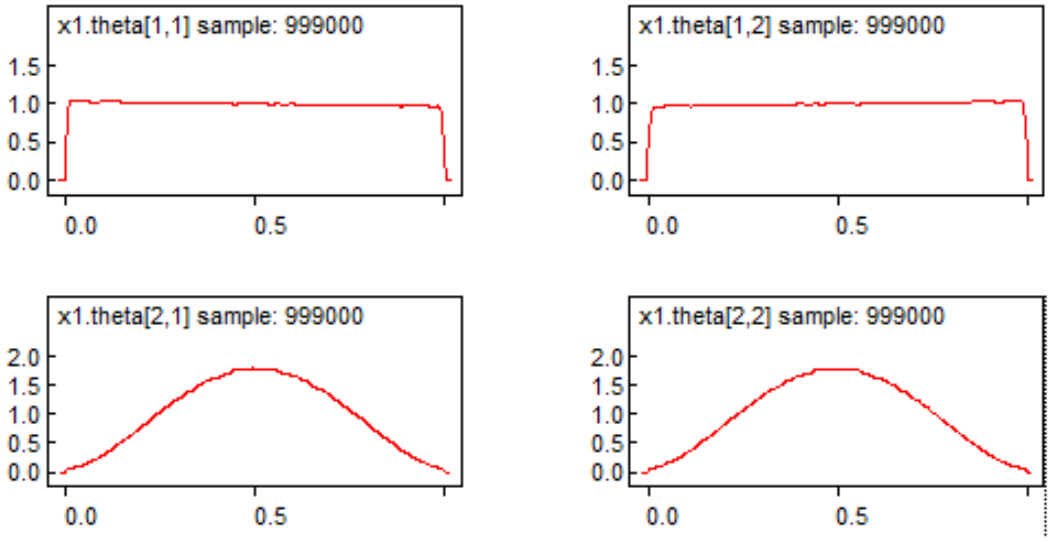

Figure 8: Approximate posterior distribution for X1

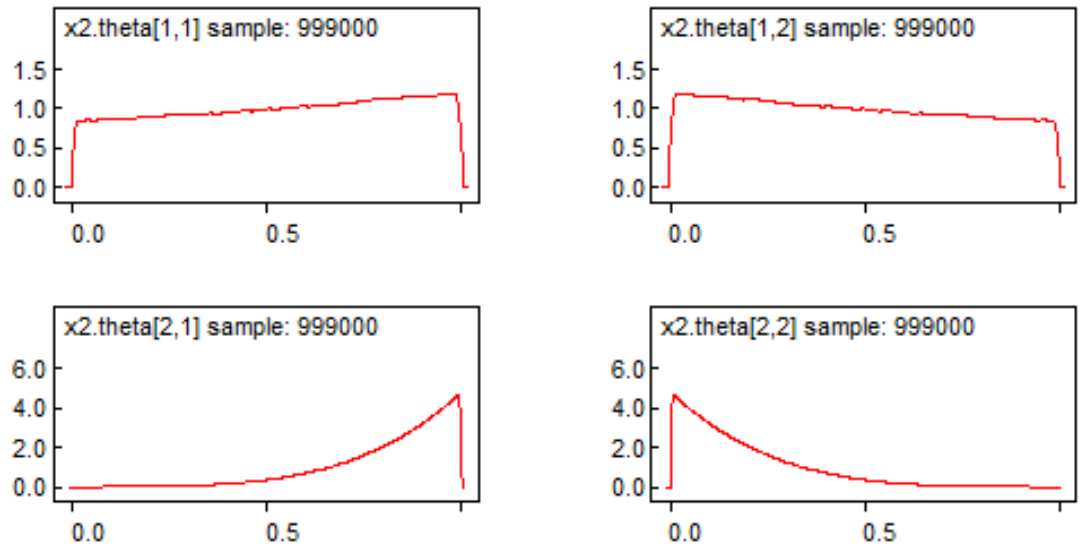

Figure 9: Approximate posterior distribution for X2

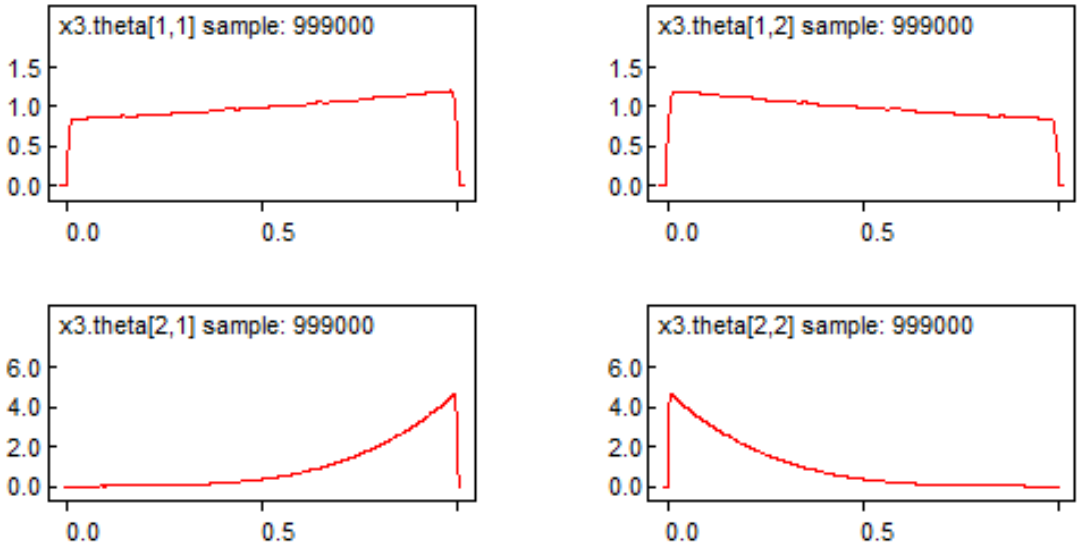

Figure 10: Approximate posterior distribution for X3

2 Task 2

2.1 Relationship to random walks on graph

2.2 Commute time method advantages

2.3 Commute time criticisms