Probablistic Graphical Models, Spring 2007
Homework 3

Due at the beginning of class on 11/05/07

Instructions

There are five questions in this homework. The last question involves some programming which should be done in MATLAB. Do *not* attach your code to the writeup. Make a tarball of your code and output, name it <userid>.tgz where your userid is your CS or Andrew id, and copy it to /afs/cs/academic/class/10708-f07/hw3/<userid> .

If you are not submitting this from an SCS machine, you might need to authenticate yourself first. See `http://www.cs.cmu.edu/~help/afs/cross_realm.html` for instructions. If you are not a CMU student and don't have a CS or Andrew id, email your submission to 10708-07-instr@cs.cmu.edu.

You are allowed to use any of (and only) the material distributed in class for the homeworks. This includes the slides and the handouts given in the class[1]. Refer to the web page for policies regarding collaboration, due dates, and extensions.

1 [10 pts] Score Equivalence

Recall the definition of score equivalence covered in class. A scoring function is called score-equivalent if the score of two I-equivalent BNs(ie with set of independencies) is the same. A simple prior, called the K2 prior, that can be used in the Bayesian score to score BNs is to take a fixed Dirichlet distribution $Dirichlet(\alpha, \alpha, \ldots, \alpha)$ (for this problem, assume α is 1) for every parameter.
Show that the Bayesian score with such a K2 prior is not score equivalent.
Hint: construct a data set for which the score of the network $X \rightarrow Y$ differs from the score of $X \leftarrow Y$.

2 [15 pts] Scoring functions

Consider the problem of learning a Bayesian network structure over two random variables X and Y.

1. Show a data set – an empirical distribution and a number of samples N will suffice – where the optimal network structure according to the BIC scoring function is different from the optimal network structure according to the ML scoring function.

2. Assume that we continue to get more samples that exhibit precisely the same empirical distribution. (For simplicity, we restrict attention to values of N that allow that empirical distribution to be achieved; e.g., an empirical distribution of 50% heads and 50% tails can only be achieved for an even number of samples.) At what value of N will the optimum BIC network structure become the same as the optimum ML network structure?

[1] Please contact Monica Hopes(meh@cs)if you need a copy of a handout.

3 [25 pts] Harmonium Model

Consider a Markov random field model defined on a bipartite graph consisting of a hidden layer of node $\{h_j\}$, and an observed layer of nodes $\{x_i\}$. For each node, we define a set of singleton features for the local potentials, $\{f_{ia}(x_i)\}$ for node x_i and $\{g_{jb}(h_j)\}$ for node h_j. For every pair of connected nodes x_i, h_j in the bipartite graph, we introduce a set of simple coupling features, $\{f_{ia}(x_i)g_{jb}(h_j)\}$, for the pairwise potentials to capture quadratic interactions. Thus we have the following partially observed Markov random field:

$$p(\{x_i, h_j\}) \propto \exp\big\{ \sum_{i,a} \theta_{ia} f_{ia}(x_i) + \sum_{j,b} \lambda_{jb} g_{jb}(h_j) + \sum_{i,j,a,b} W_{ia}^{jb} f_{ia}(x_i) g_{jb}(h_j) \big\}. \tag{1}$$

This model is known as a *harmonium*.

3.1 Conditionals under exponential family harmonium.

Show that the conditional distribution $p(\{x_i\}|\{h_j\})$, and $p(\{h_j\}|\{x_i\})$ can be represented exactly by a product of independent distributions in the exponential family. That is,

$$p(\{h_j\}|\{x_i\}) = \prod_j p_j(h_j) = \prod_j \exp\{\sum_b \lambda'_{jb} g_{jb}(h_j) - B_j(\{\lambda'_{jb}\})\},$$

and

$$p(\{x_i\}|\{h_j\}) = \prod_i \exp\{\sum_a \theta'_{ia} f_{ia}(x_i) - A_i(\{\theta'_{ia}\})\}.$$

You need to show how you would get a posterior of this form, and what θ'_{ia}, λ'_{jb}, $A(\cdot)$ and $B(\cdot)$ are. Discuss the benefit of this deposition of the posterior to the solution of the inference task.

3.2 Marginals under exponential family harmonium.

Derive the marginals of the observed and latent variables, $p(\{x_i\})$ and $p(\{h_j\})$. (Hint, use the partition functions you just derived, i.e., $A_i(\cdot)$ and $B_j(\cdot)$.) You should show that the the marginals have a nice compact form, i.e.,

$$p(\{h_j\}) = \prod_j p_j(h_j) = \prod_j \exp\{\sum_b \lambda'_{jb} g_{jb}(h_j) - B_j(\{\lambda'_{jb}\})\}.$$

Are the observed nodes marginally independent?

3.3 Constructing a harmonium from bottom up.

Now, assume that

$$p(h_j|\{x_i\}) = \text{Normal}(\sum_i x_i W_i^j, 1), \tag{2}$$

and

$$\begin{aligned} p(x_i|\{h_j\}) &= \text{Binomial}(N, \text{logistic}(\alpha_i + \sum_j h_j W_i^j)), \\ &\propto \exp\big\{ -\log\Gamma(x_i+1) - \log\Gamma(N-x_i+1) + \alpha_i x_i + \sum_j h_j W_i^j x_i \big\}. \end{aligned} \tag{3}$$

Note that here we assume x is discrete and h is continuous. This definition assumes that each x_i has only one local feature, and each h_i has two local features, why? Inspecting the conditional you derived in 3.1, and the canonical harmonium defined by Eq. 1, write out the MRF induced by the two conditionals given above. (Hint, you want to represent all conditionals in exponential family representations, as is already done in Eq. 3). Using your results in 3.2, derive the marginal distribution $p(\{x_i\})$ under this harmonium. Is the resulting marginal always well-defined, i.e., normalizable?

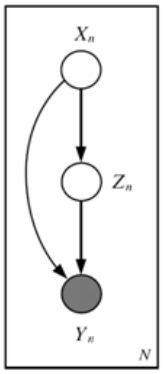

Figure 1: Bayesian network for Problem 4

3.4 Normalizability.

Now assume x is continuous, and let $p(x_i|\{h_j\}) = \text{Normal}(\sum_j h_j W_i^j, 1)$. Using this new conditionals, write the marginals as you did in 3.3. Are the marginals always normalizable in this case? Why or why not?

4 [20 pts] EM for Conditional Mixture Model

Consider the Conditional Mixture (or Mixture of Experts) model discussed in class. The model is an example of a three node graphical model with one latent(unobserved) variable as shown in Fig. 1. The latent variable z_n chooses an expert, usually with a softmax function, i.e $P(z^k = 1|x_n) \propto e^{\zeta_k^T x_n}$ while each expert can be a linear regression model $P(y|x, z^k = 1) = Normal(y; \theta_k^T x, \sigma_k^2)$.

The expected log-likelihood of the data under this model is given by

$$\sum_n < \log(p(z_n|x_n, \zeta)) >_{p(z|x,y)} + \sum_n < \log(p(y_n|x_n, z_n, \theta, \sigma)) >_{p(z|x,y)}$$

Derive the update equations for the E and M steps for this conditional mixture model.

5 [30 pts] EM for Continuous Variables

Consider a HMM with Gaussian emission probabilities $p(y_t|x_t)$ where y_t is a two dimensional real vector of observed variables and x_t represents the hidden variables.

1. Consider the problem of parameter estimation in this model. Derive the E and M update equations for this model. Restrict the covariance matrices to be isotropic: $\Sigma = \sigma^2 I$.

2. In Matlab, use these update equations to fit a HMM (with 4 states) to the two dimensional data(containing the y_t) in em-train.dat. Report the log-likelihood of this data under the model you just learnt. Then, evalutate the log likelihood on the test data(em-test.dat). Plot the data along with the means of the 4 component densities. Include this plot in your writeup.

3. Fit a Gaussian mixture model with 4 states to the same data(again with isotropic covariance matrices). Compare the performance with that of the HMM. Submit whatever code you have written for this question, along with a README briefly describing your code and submit them to the homework directory.