Problem Set 3
Parameter Learning and EM in Bayesian Networks

Prof. Jim Rehg
CS 7636 Computational Perception
College of Computing
Georgia Institute of Technology

February 20, 2004

Solutions must be submitted at beginning of class on Friday, February 27, 2004. Please submit printouts of your code and results for problems 1-3. Handwritten answers are fine for the discussion questions and derivation in problem 4, as long as they are legible and organized. Please show all work. I’ve tried to leave space on this handout for your solutions. This problem set is worth 8% of your total grade, distributed among the problems as shown.
1-3. [6% of grade]
See the file PS3.zip for sample code, data, and the description of the first three problems.

4. [2% of grade]
The EM algorithm for parameter estimation in a graphical model with unobserved (latent) nodes can be viewed as a coordinate ascent procedure for an auxiliary function $L(q, \theta)$. The coordinates are the distribution $q$ (over the latent variable) and the vector of parameters $\theta$. We have

$$
L(q, \theta) = \sum_z q(z|x) \log P(x, z|\theta) - \sum_z q(z|x) \log q(z|x).
$$

(1)

$$
= l(\theta; x) - D(q(z|x) || p(z|x, \theta))
$$

(2)

(a) Assume that the latent variable $z$ is discrete with $n$ possible values. Using equation (1), show analytically that the choice $q^* = p(z|x, \theta)$ satisfies

$$
q^* = \arg \max_q L(q, \theta)
$$

for all $\theta$. In particular, you must show that $q^*$ is a stationary point of the auxiliary function. (Hint: Use a Lagrange multiplier $\lambda$ to augment $L$ with the constraint $\sum_i q_i = 1$, obtaining $H(q, \theta, \lambda)$. Take derivatives of $H$ with respect to $q_i$ and set them equal to zero. Use the constraint equation to solve for $\lambda$.)
(b) Using equation (2), show that $\mathcal{L}(q^*, \theta) = l(\theta; x)$. Draw a picture which illustrates this condition.

(c). [Extra Credit: +0.2% of grade]
In many problems of interest, such as fitting mixture densities, the M-step can be solved in closed form. In other words, one can derive an explicit formula for $\theta^* = \arg \max_\theta \mathcal{L}(q, \theta)$. Explain the fallacy in the following (incorrect) argument:

"We have from part (b) that $\mathcal{L}(q^*, \theta) = l(\theta; x)$ for all possible $\theta$. It follows in particular that $\mathcal{L}(q^*, \theta^*) = l(\theta^*; x)$. Since we can solve in closed-form for the $\theta^*$ which is the global maximum of $\mathcal{L}(q^*, \theta)$, only one iteration of the M-step is required. The result is the global maximum of $l(\theta, x)$."