STAT37710/CMSC35400 — Homework 5

University of Chicago, Spring 2018 — due Friday, June 1 at 11:59pm

1. Let $\mathcal{G}$ be a directed graphical model (Bayes net) on a graph with n vertices $V = \{1, 2, \ldots, n\}$ and corresponding random variables $X_1, \ldots, X_n$. Assume for simplicity that each X_i can take on k different values, $\{1, 2, \ldots, k\}$, Let $\mathrm{pa}(i)$ denote the set of parents of vertex i in the graph.

   As usual, for any subset U of V, say, $U = \{v_1, \ldots, v_m\}$, $\boldsymbol{X}_U$ will denote the vector $(X_{v_1}, \ldots, X_{v_m})$ of corresponding variables, and when talking about the probability of this these variables jointly taking on the values $\mathbf{x}_U = (x_{v_1}, \ldots, x_{v_m})$ (where naturally $x_{v_i} \in \{1, 2, \ldots, k\}$), instead of writing $\mathbb{P}(\boldsymbol{X}_U = \mathbf{x}_U)$, we will use the shorthand $p(\mathbf{x}_U)$. For the probability of all the variables together instead of $p(\mathbf{x}_V)$ we will just use $p(\mathbf{x})$.

   Recall that the general form of the joint distribution corresponding to this model is

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | x_1^{(i)}, \ldots, x_{p_i}^{(i)}), \tag{1}$$

   where $x_1^{(i)}, x_2^{(i)}, \ldots x_{p_i}^{(i)}$ are values of the parents of X_i. Since all the variables here are in $\{1, 2, \ldots, k\}$, (1) can be equivalently written as

$$p(\mathbf{x}) = \prod_{i=1}^{n} [\theta_i]_{x_i, x_1^{(i)}, x_2^{(i)}, \ldots x_{p_i}^{(i)}}$$

   where each θ_i is a $k \times k \times \ldots \times k$ dimensional array with $[\theta_i]_{x_i, x_1^{(i)}, x_2^{(i)}, \ldots x_{p_i}^{(i)}} = p(x_i | x_1^{(i)}, \ldots, x_{p_i}^{(i)})$. Such multidimensional arrays are often called tensors. These arrays are the parameters of the model. Let $\boldsymbol{\Theta}$ denote the tensors $\theta_1, \ldots, \theta_n$ tensors collectively.

   (a) Write down the log-likelihood $\ell(\boldsymbol{\Theta}) = \log p(\mathbf{x})$ of a single training example (i.e., sample) $\mathbf{x}$ from this model. Now assume that we have a training set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ of m different independent samples from our model and that $[N_i]_{x_i, x_1^{(i)}, \ldots, x_{p_i}^{(i)}}$ is the number of examples in the training data where X_i took on the value x_i, its first parent took on the value $x_1^{(i)}$, etc.. Write down the log-likelihood of the training set in terms of these counts.

   (b) Derive the maximum likelihood estimator of the $\theta_1, \ldots, \theta_n$ parameter tensors. (The final answer in the special case where every node has two parents can be found in the slides, but the derivation cannot.) Explain what might be a problem with using maximum likelihood estimation in this context if some of the $[N_i]_{x_i, x_1^{(i)}, \ldots, x_{p_i}^{(i)}}$ counts are zero.

   (c) As we have mentioned in class, the Dirichlet distribution over k random variables $0 \leq \nu_i \leq 1$ is

$$p_{\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)}(\nu_1, \ldots, \nu_k) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} \nu_1^{\alpha_1 - 1} \nu_2^{\alpha_2 - 1} \ldots \nu_k^{\alpha_k - 1}.$$

   Here $\alpha_1, \ldots, \alpha_k$ are the parameters of the distrbution, and $\Gamma(\cdot)$ is a special function, but you do not need to worry about its form for this exercise.

   The Dirichlet distribution is the favorite prior that Bayesians like to use when approximating the parameters of directed graphical models. Show, in particular, that if node i in our model has just a single parent and the prior distribution on the corresponding θ_i tensor is that

$$p([\theta_i]_{1,b}, \ldots, [\theta_i]_{k,b}) = \mathrm{Dir}(\alpha, \alpha, \ldots, \alpha)$$

   for any value b of the parent, then the posterior distribution of these parameters is

$$p([\theta_i]_{1,b}, \ldots, [\theta_i]_{k,b} | \, \mathbf{x}_1, \ldots, \mathbf{x}_m) = \mathrm{Dir}([N_i]_{1,b} + \alpha, [N_i]_{2,b} + \alpha, \ldots, [N_i]_{k,b} + \alpha).$$

Show that simiarly when i has two parents, and the prior on the entries of θ_i (which is now a $k \times k \times k$ array) is

$$p([\theta_i]_{1,b,c}, \dots, [\theta_i]_{k,b,c}) = \mathrm{Dir}(\alpha, \alpha, \dots, \alpha)$$

for any values b and c of the parents, then the posterior distribution of these parameters is

$$p([\theta_i]_{1,b,c}, \dots, [\theta_i]_{k,b,c} | \, \mathbf{x}_1, \dots, \mathbf{x}_m) = \mathrm{Dir}([N_i]_{1,b,c} + \alpha, [N_i]_{2,b,c} + \alpha, \dots, [N_i]_{k,b,c} + \alpha).$$

The generalization to p_i parents is obvious, but we will save ourselves the work of having to write it down.

(d) The mean of the i'th component of a Dirichlet distribution $p(\nu_1, \dots, \nu_k) = \mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ is

$$\overline{\nu_i} = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}.$$

Explain why the Bayesian strategy avoids the problem with zero counts.

2. Recall that the probability distribution associated with a Hidden Markov Model of length $T+1$ is

$$p(x_0, \dots, x_T, y_0, \dots, y_T) = p(x_0) \left(\prod_{t=1}^{T} p(x_t | x_{t-1}) \right) \left(\prod_{t=0}^{T} p(y_t | x_t) \right).$$

If the model is stationary, and each hidden state can take on one of the values $\{1, 2, \dots, N_h\}$, while each of the observed states can take on one of the values $\{1, 2, \dots, N_o\}$, then

$$p(x_0) = \pi_i \tag{2}$$

$$p(x_t | x_{t-1}) = \theta_{x_{t-1}, x_t} \qquad t = 1, 2, \dots T \tag{3}$$

$$p(y_t | x_t) = \omega_{x_t, y_t} \qquad t = 0, 1, \dots T \tag{4}$$

$$\tag{5}$$

for some parameter vector $\pi \in \mathbb{R}^{N_h}$ and parameter matrices $\theta \in \mathbb{R}^{N_h \times N_h}$ and $\omega \in \mathbb{R}^{N_h \times N_o}$.

As we have seen in class, given these parameters and an observation sequence $y_0, y_1, \dots, y_T$, the marginal distribution of the hidden nodes can then be computed with the forward-backward algorithm as

$$p(x_t | y_0, \dots, y_T) = \frac{\alpha_t(x_t) \, \beta_t(x_t)}{\sum_{x_t'} \alpha_t(x_t') \, \beta_t(x_t')} =: \gamma_t(x_t).$$

(a) Now imagine that a certain observation y_m is missing, i.e., that it is now a hidden (latent) variable of the model. Hence one needs to marginalize over it, i.e., sum (**??**) over its possible values. Derive the joint distribution

$$p(x_0, \dots, x_T, y_0, \dots, y_{m-1}, y_{m+1}, \dots, y_T)$$

of this modified model. Generalizing from the standard case where all y_t's are observed, define

$$\alpha_t(x_t) = \begin{cases} p(x_t, y_0, \dots, y_t) & t < m \\ p(x_t, y_0, \dots, y_{m-1}, y_{m+1}, \dots, y_t) & t \geq m \end{cases}$$

and

$$\beta_t(x_t) = \begin{cases} p(y_{t+1}, \dots, y_{m-1}, y_{m+1}, \dots, y_T | x_t) & t+1 \leq m \\ p(y_{t+1}, \dots, y_T | x_t) & t+1 > m. \end{cases}$$

Derive recursion relations for $\alpha_t(x_t)$ and $\beta_t(x_t)$ that allow these quantities to be computed fast. Derive a formula for $p(y_m | y_0, \dots, y_{m-1}, y_{m+1}, \dots, y_T)$ in terms of the α_t's and β_t's and explain how this can be used to fill in missing observations.

(b) Show that

$$p(x_t, x_{t+1}|y_0, \ldots, y_T) = \frac{\gamma_t(x_t)\,\beta_{t+1}(x_{t+1})}{\beta_t(x_t)}\,p(x_{t+1}|x_t)\,p(y_{t+1}|x_{t+1}) =: \xi_t(x_t, x_{t+1})$$

(c) Express the expected log–likelihood

$$\mathbb{E}(\log \ell(\pi, \theta, \omega)) = \sum_{x_0} \sum_{x_1} \cdots \sum_{x_T} \left[\log \pi_{x_0} + \sum_{t=1}^{T} \log \theta_{x_{t-1}, x_t} + \sum_{t=0}^{T} \log \omega_{x_t, y_t} \right] p(x_0, \ldots, x_T | y_0, \ldots, y_T)$$

in terms of γ_t and ξ_t and show that it is maximized by updating π, θ and ω to

$$\pi_i^{\text{new}} = \gamma_0(i), \qquad \omega_{i,j}^{\text{new}} = \frac{\sum_{t\,:\,y_t=j} \gamma_t(i)}{\sum_{t=0}^{T} \gamma_t(i)}, \qquad \theta_{i,j}^{\text{new}} = \frac{\sum_{t=0}^{T-1} \xi_t(i,j)}{\sum_{t=0}^{T-1} \gamma_t(i)}. \tag{6}$$

Explain the intuitive meaning of these expressions.