Homework 3: Part I

Statistical Approaches to Learning and Discovery

Zoubin Ghahramani & Teddy Seidenfeld

Due: Mon Apr 22, 2002

On this assignment, you are welcome to discuss the homework with each other, but you are encouraged to work independently. Please hand in your own work—identical answers will be frowned upon.

1. Gaussian Markov Models

   (a) Consider a multivariate Gaussian variable $(x_1, \ldots, x_n)$ with given mean vector μ and covariance matrix Σ. Write out the probability density function for this vector. How can we define a Markov network that captures the conditional independencies between the x_i?

   (b) Let $n = 4$, $\mu = (0, 1, 1, 0)$ and

$$\Sigma = \frac{1}{6} \begin{pmatrix} 7 & -2 & -2 & 1 \\ -2 & 7 & 1 & -2 \\ -2 & 1 & 7 & -2 \\ 1 & -2 & -2 & 7 \end{pmatrix},$$

   draw the corresponding Markov network and define clique potentials consistent with the above Gaussian.

   (c) State a general theorem relating the zeros in the inverse covariance matrix of a multivariate Gaussian and conditional independence between the variables. Prove your theorem.

2. Download the data file called `geyser.txt` from the course web site. This is a sequence of 295 consecutive measurements of two variables from Old Faithful geyser in Yellowstone National Park: the duration of the current eruption in minutes (to nearest 0.1 minute), and the waiting time until the next eruption in minutes (to nearest minute). Examine the data by plotting the variables within and between consecutive time steps. E.g. `plot(geyser(1:end-1,1),geyser(2:end,1),'o');`. Discuss and justify based on your observations what kind of model might be most appropriate for this data set: e.g. a mixture of Gaussians, a hidden Markov model, a linear dynamical system, etc.

3. Consider the following two HMMs:

$$P_1(\mathbf{y}_{1:T}, \mathbf{s}_{1:T}) = P(\mathbf{s}_1)P(\mathbf{y}_1|\mathbf{s}_1)\prod_{t=2}^{T} P(\mathbf{y}_t|\mathbf{s}_t)P(\mathbf{s}_t|\mathbf{s}_{t-1})$$

   and

$$P_2(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = P(\mathbf{z}_1)P(\mathbf{y}_1|\mathbf{z}_1)\prod_{t=2}^{T} P(\mathbf{y}_t|\mathbf{z}_t)P(\mathbf{z}_t|\mathbf{z}_{t-1})$$

where $x_{1:T}$ denotes the sequence $x_1 \ldots x_T$, $\mathbf{y}_t$ is the observation at time t and $\mathbf{s}$ and $\mathbf{z}$ are the hidden state variables for each HMM, respectively. Now form a new model for the data by multiplying these two models and renormalizing:

$$P_3(\mathbf{y}_{1:T}, \mathbf{s}_{1:T}, \mathbf{z}_{1:T}) = \frac{1}{Z} P_1(\mathbf{y}_{1:T}, \mathbf{s}_{1:T}) P_2(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})$$

(a) Draw an graphical model—with a node for each variable $\mathbf{y}_t$, $\mathbf{s}_t$, and $\mathbf{z}_t$—representing the conditional independence relationships in this new model, P_3.

(b) Given a sequence $\mathbf{y}_{1:T}$, describe how you would compute $P(\mathbf{s}_t, \mathbf{z}_t | \mathbf{y}_{1:T})$. What is the time complexity of your algorithm?

(c) If $\mathbf{s}_t$ and $\mathbf{z}_t$ are both discrete, taking on at most K states, is this model equivalent to an HMM with K^2 states? Why or why not?

(d) Assume you want to learn the parameters of this model from data. Let's re-write the model more explicitly to make it clear:

$$P_3(\mathbf{y}_{1:T}, \mathbf{s}_{1:T}, \mathbf{z}_{1:T} | \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{Z} P_1(\mathbf{y}_{1:T}, \mathbf{s}_{1:T} | \boldsymbol{\theta}) P_2(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \boldsymbol{\phi})$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the usual transition, emission, and initial state HMM parameters for HMM 1 and 2, respectively. What is the derivative of the log likelihood of P_3 with respect to the transition parameter, $\boldsymbol{\theta}_{ij} = P(\mathbf{s}_{t+1} = j | \mathbf{s}_t = i)$, (for all t)?

(e) Assume there are L possible symbols: $\mathbf{y}_t \in \{1, \ldots, L\}$ and each HMM has K states. What is the maximum mutual information between $\mathbf{y}_t$ and $\mathbf{y}_{t+1}$ in this model, maximizing over all parameters?

4. In the automatic speech recognition community, HMMs are sometimes trained by using the Viterbi algorithm instead of the forward–backward algorithm. In other words, in the E step of EM (Baum–Welch), instead of computing the expected sufficient statistics from the posterior distribution over hidden states: $P(\mathbf{s}_{1:T} | \mathbf{y}_{1:T}, \boldsymbol{\theta})$, the sufficient statistics are computed using the single *most probable* hidden state sequence: $\mathbf{s}^*_{1:T} = \arg\max P(\mathbf{s}_{1:T} | \mathbf{y}_{1:T}, \boldsymbol{\theta})$. Is this algorithm guaranteed to converge? Is so, will it converge to the a maximum of the likelihood? If not, will it oscillate? Support your arguments.