Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Problem Set 10

Issued: Thursday, December 4, 2014

Problem 10.1
The Chow-Liu algorithm is a simple and efficient way to learn a tree structure that minimizes the information divergence to the empirical distribution $\tilde{p}$:

$$p_{\text{Chow}-\text{Liu}} = \arg\min_{p \in \mathcal{T}} D(\tilde{p}|p)$$

where $\mathcal{T}$ is the class of distributions defined on trees.

The algorithm first computes empirical mutual information of all pairs of variables using their sample values:

$$I(x_i, x_j) \equiv \sum_{x_i, x_j} \tilde{p}_{x_i,x_j}(x_i, x_j) \log \frac{\tilde{p}_{x_i,x_j}(x_i, x_j)}{\tilde{p}_{x_i}(x_i)\tilde{p}_{x_j}(x_j)}$$

where $\tilde{p}_{x_i,x_j}(x_i, x_j)$ is the empirical marginal distribution of x_i and x_j obtained from samples. Then, it sets each edge weight equal to the empirical mutual information and finds the maximum weight spanning tree, which can be solved using greedy algorithms. In particular, Kruskal's algorithm begins with a graph with no edges, and successively adds maximal-weight edge to the graph while ensuring that the graph remains cycle-free.

(a) Consider four variables with the following empirical mutual information:

$I(x_i, x_j)$	1	2	3	4
1	0.3415	0.2845	0.0003	0.0822
2	0.2845	0.3457	0.0005	0.0726
3	0.0003	0.0005	0.5852	0.0002
4	0.0822	0.0726	0.0002	0.5948

Find the Chow-Liu tree of the four variables.

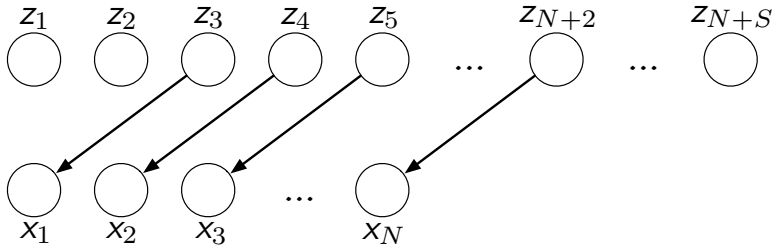
(b) `chowLiuData.mat` contains a 10×5000 matrix, where each column represents binary sample values of $\{x_1, x_2, ...x_{10}\}$. Implement the Chow- Liu algorithm and find three Chow-Liu trees of the 10 variables using the first $100, 3000,$ and 5000 samples, respectively.

You may find the following functions useful:

- `connected.m`: `connected(adjmat,i,j)` returns true if node `i` and `j` are connected in the graph, where `adjmat` is the adjacency matrix of the graph with `adjmat(i,j) = 1` if there is an edge between `i` and `j` and 0 otherwise.
- `draw_graph.m`: `draw_graph(adjmat)` draws a graph with adjacency matrix `adjmat`.

Problem 10.2

Suppose we are given random sequences $z_1, z_2, \ldots, z_{N+S}$ and $x_1, x_2, \ldots, x_N$, where N and S are given, and where $z_n, x_n \in \{1, 2, \ldots, M\}$. The structure in the joint probability distribution for these variables is given by a directed acyclic graph $\mathcal{G}_s$ in which z_n has no parents and z_{n+s} is the unique parent node of x_n, for $n = 1, 2, \ldots, N$. For example, $\mathcal{G}_2$ is as follows.



In the graphical model, $s \in \{0, \ldots, S\}$ is a parameter, as are the distributions $p^s_{x|z}(x|z) = p_{x_n|z_{n+s}}(x|z)$ and $p_z(z) = p_{z_n}(z)$, which do not depend on n (as our notation reflects).

We are given a sample for each variable: $D_x = \{x_1, x_2, \ldots, x_N\}$, $D_z = \{z_1, z_2, \ldots, z_{N+S}\}$, so our complete data is $D = \{D_x, D_z\}$.

(a) For a given s, express the maximum-likelihood (ML) estimates for the parameters $p^s_{x|z}(x|z)$ and $p_z(z)$ in terms of the following empirical distributions computed from D

$$\hat{p}^s_{x,z}(x, z) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{x_n=x} \mathbb{1}_{z_{n+s}=z}, \qquad \hat{p}_z(z) = \frac{1}{N+S} \sum_{n=1}^{N+S} \mathbb{1}_{z_n=z},$$

where we have used the indicator function notation

$$\mathbb{1}_{u=v} = \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}.$$

(b) The model $\mathcal{G}_s$ maximizing the log-likelihood of the data $\ell(\mathcal{G}_s; D) \triangleq \ell((\mathcal{G}_s, \hat{\theta}^{ML}_{\mathcal{G}_s}); D)$ can be expressed as $\mathcal{G}_{s^*}$ where

$$s^* = \arg\max_s f(\hat{p}^s_{x,z}, \hat{p}_z),$$

with $\hat{p}^s_{x,z}(x, z)$ and $\hat{p}_z(z)$ as defined in part (a). Determine the function $f(\cdot)$.

In the following parts (c) and (d), we restrict our attention to the case $M = 2$ (binary variables), and $S = 2$. Suppose our data are $D_x = \{2, 1, 1, 1\}$ and $D_z = \{1, 1, 1, 2, 1, 1\}$. Furthermore, suppose we are deciding between candidate structure $\mathcal{G}_2$ and an alternative structure $\tilde{\mathcal{G}}$ that has no edges, corresponding to a fully-disconnected graph.

(c) Compute $\ell(\mathcal{G}_2, D) - \ell(\tilde{\mathcal{G}}, D)$ and determine which structure has the higher likelihood score. Explain why the likelihood score is not appropriate for choosing between $\mathcal{G}_2$ and $\tilde{\mathcal{G}}$.

Reminder: Via the usual convention, we define $0 \log 0 \triangleq 0$.

(d) Recall that the Bayesian score is defined as $\ell_{\mathrm{B}}(\mathcal{G}, D) = \log p(D|\mathcal{G}) + \log p(\mathcal{G})$ where

$$p(D|\mathcal{G}) = \int p(D|\theta_{\mathcal{G}}, \mathcal{G})\, p(\theta_{\mathcal{G}}|\mathcal{G})\, d\theta_{\mathcal{G}}.$$

Assume that $p(\tilde{\mathcal{G}}) = p(\mathcal{G}_2)$ and both $p(\theta_{\tilde{\mathcal{G}}}|\tilde{\mathcal{G}})$ and $p(\theta_{\mathcal{G}_2}|\mathcal{G}_2)$ are uniform distributions. Compute $\ell_{\mathrm{B}}(\mathcal{G}_2, D) - \ell_{\mathrm{B}}(\tilde{\mathcal{G}}, D)$ and determine which structure has the higher Bayesian score.

Hint: In your analysis, you may find it convenient to use the following parameter notation

$$\begin{aligned} &\text{For } \tilde{\mathcal{G}}: && \tilde{\gamma} = p_{\mathsf{z}}(1), && \beta = p_{\mathsf{x}}(1) \\ &\text{For } \mathcal{G}_2: && \gamma = p_{\mathsf{z}}(1), && \alpha_1 = p_{\mathsf{x}|\mathsf{z}}(1|1), \qquad \alpha_2 = p_{\mathsf{x}|\mathsf{z}}(1|2). \end{aligned} \tag{1}$$

Problem 10.3

Recall Problem 5.6 where you were asked to implement the forward-backward and the Viterbi algorithm to locate genes in a DNA sequence. Now, pretend you do not have the model parameters for the HMM. Implement the Baum-Welch algorithm by augmenting your forward-backward algorithm, then run it on your test sequence, which you generated in Problem 5.6(a) (if you did not store your sequence, then you may re-generate a sequence using the true parameters in 5.6). Again find the region labels as the MAP estimates of the marginals, and compare estimation accuracy to your answer in Problem 5.6(b).

Problem 10.4

(a) A discrete-time Gaussian stochastic process $\mathsf{y}[n]$ is generated by a first-order system driven by white Gaussian noise $\mathsf{w}[n]$:

$$\mathsf{y}[n+1] = \alpha \mathsf{y}[n] + \mathsf{w}[n], \qquad n = 0, 1, 2, \ldots$$

where $\mathsf{y}[0] \sim N(0, \lambda)$ is independent of the noise sequence. Also, $\mathsf{w}[n] \sim N(0, \sigma^2)$, and α is unknown. Find the ML estimate of α based on observation of $\mathsf{y}[0], \ldots, \mathsf{y}[N+1]$, i.e., compute the value of α that maximizes

$$p_{\mathsf{y}[0],\ldots,\mathsf{y}[N+1]|\alpha}(y[0], \ldots, y[N+1]|\alpha).$$

(*Hint:* $\mathsf{y}[n]$ is a Gauss-Markov process.)

(b) Now suppose $\mathsf{y}[n]$ is a Gaussian process generated by an $(M+1)$th-order system driven by white Gaussian noise $\mathsf{w}[n]$:

$$\mathsf{y}[n+1] = a_0 \mathsf{y}[n] + a_1 \mathsf{y}[n-1] + \cdots + a_M \mathsf{y}[n-M] + \mathsf{w}[n], \qquad n = 0, 1, 2, \ldots,$$

where $[\mathsf{y}[-M] \cdots \mathsf{y}[0]]^{\mathrm{T}} \sim N(0, \mathbf{\Lambda})$ is independent of the white noise sequence. Also, $\mathsf{w}[n] \sim N(0, \sigma^2)$, and $\mathbf{a} = [a_0 \cdots a_M]^{\mathrm{T}}$ is unknown. Find the ML estimate of $\mathbf{a}$ based on observation of $\mathsf{y}[-M], \ldots, \mathsf{y}[N+1]$.

(c) Sometimes it is useful to take a discrete-time sequence $y[-M], \ldots, y[N+1]$ and try to model each sample as a linear combination of its past samples. That is, we seek to find the best $\mathbf{a} = [a_0 \cdots a_M]^{\mathrm{T}}$ so that we minimize

$$\sum_{k=1}^{N+1} \left(y[k] - \sum_{i=0}^{M} a_i y[k-1-i] \right)^2 .$$

Find the optimal $\mathbf{a}$, and compare with part (b).

6.438 Algorithms for Inference

Fall 2014