Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Problem Set 9

Issued: Tuesday, November 25, 2014 **Due:** Thursday, December 4, 2014

Suggested Reading: Lecture notes 20-22

Problem 9.1

Alice wants to build an optical character recognition (OCR) system, which scans images of words and recognizes the letters in the words. Each letter x_j takes values in the alphabet $\{A, B, ..., Z\}$. For any fixed word length d, we model the word $(x_1, x_2, ..., x_d)$ as a Markov chain

$$p_{x_1,x_2,...,x_d}(x_1, x_2, ..., x_d) = p_{x_1}(x_1) \prod_{j=2}^{d} p_{x_j|x_{j-1}}(x_j|x_{j-1})$$

(a) Suppose that Alice is given a very large collection of data $D = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$, where each $\mathbf{x}^{(i)}$ is a d-letter word. Give an expression for the maximum likelihood estimates of the parameters (The parameters are the elements of $p_{x_1}(x_1)$ and $p_{x_j|x_{j-1}}(x_j|x_{j-1})$).

(b) Assume that in Alice's dataset D, no word starts with "EE". What is the ML estimate of $P(x_2^{(n+1)} =$ "E" $|x_1^{(n+1)} =$"E", $D)$, i.e., the probability of the second letter in the (n+1)-th word being "E" conditioned on that the first letter in the (n+1)-th word is "E" given the data?

(c) Disappointed that her system cannot recognize the word "EECS", Alice decided to use the Bayesian estimates instead, and assumed that each parameter vector has the Dirichlet prior with all hyperparameters equal to 1. Give an expression for the posterior distribution of the parameters conditioned on data.

(d) What is $P(x_2^{(n+1)} =$ "E" $|x_1^{(n+1)} =$"E", $D)$ based on the Bayesian estimates, assuming that D still does not contain a word starting with "EE"?

Problem 9.2

In this problem, we try to learn undirected graph parameters for joint Gaussian distributions. Consider a joint Gaussian distribution over $x = [x_1, x_2, \ldots, x_6]$, as shown in Figure 1. Each node can be a Gaussian vector.

(a) Suppose that you observe K i.i.d. samples $x^{(1)}, \ldots, x^{(K)}$. Provide the Maximum Likelihood estimator for the covariance matrix $\hat{\Sigma}$ and the mean $\hat{\mu}$.

(b) However, in order to do inference on the graphical model, you need to learn the information matrix J. In this example, assume that you are interested in estimating the block $J_{123,123}$ corresponding to the variables x_1, x_2, x_3.
One approach is to invert the estimated covariance matrix $\hat{\Sigma}$ to obtain an estimation

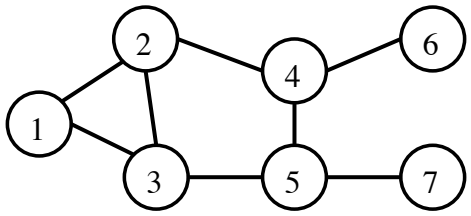


Figure 1

of $\hat{J}$. Another approach is described below. Assume that you have a script of loopy Gaussian BP algorithm, which inputs an information matrix $\widetilde{J}$ for a potentially loopy graph and outputs all the messages $\widetilde{J}_{i\to j}$. Use this script to get an estimation of $\hat{J}_{123,123}$.
(hint: Specify the input and output to the script, and express the estimator $\hat{J}_{123,123}$ in terms of the output as well as $\hat{\Sigma}$)

(c) Comment on the complexity and the accuracy of the two approaches to learn Gaussian graphical models in (b).
((hint: in general, if a matrix A is sparse, the inverse A^{-1} will not have the sparsity pattern. Moreover, A^{-1} may not be sparse at all.))

Problem 9.3
Consider the Naive Bayes model with class variable c and discrete observed variables $\mathsf{x}_1, ..., \mathsf{x}_K$. The conditional probability distributions for the model are parameterized by $p_{\mathsf{c}}(c) = \theta_c$ and $p_{\mathsf{x}_k|\mathsf{c}}(x|c) = \theta_{x_k|c}$ for $k = 1, ..., K$ and for all assignments $x_k \in \mathcal{X}$ and classes $c \in \mathcal{C}$.

Now given a dataset $D = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$, where each $\mathbf{x}^{(n)}$ is a complete sample of the observed variables, $\mathsf{x}_1, ..., \mathsf{x}_K$, we can use the EM to learn the parameters of our model. Note that the class variable, c, is never observed.

(a) Show that if we initialize the parameters uniformly,

$$\theta_c^{(0)} = \frac{1}{L} \qquad \text{and} \qquad \theta_{x_k|c}^{(0)} = \frac{1}{M} \tag{1}$$

for all x_k, c where $L = |\mathcal{C}|$ and $M = |\mathcal{X}|$, then the EM algorithm converges in one iteration, and gives a closed form expression for the parameter values at this convergence point.

(b) Consider a simple example with $K = 2, M = 2, L = 2$ and the true parameters are $\theta_0 = \theta_1 = 1/2$, $\theta_{1|1} = \theta_{0|0} = 1$ for both $k = 1$ and 2. Assume that for half of our dataset, $\mathbf{x}^{(n)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and for the other half $\mathbf{x}^{(n)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Show that if we initialize our parameters to

$$\theta_c^{(0)} = \frac{1}{2} \qquad \text{and} \qquad \theta_{1|1}^{(0)} = \theta_{0|0}^{(0)} = \frac{1}{2} + \epsilon \tag{2}$$

for both $k = 1$ and 2, where ϵ is a small positive number, then the EM algorithm converges to the true parameters.

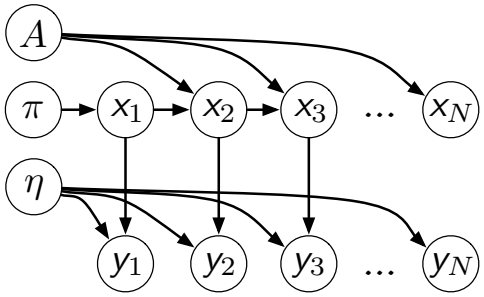
Hint: Show that as long as $0 < \epsilon < 1/2$, $\theta_{1|1}$ and $\theta_{0|0}$ increase at each iteration of the EM and converge to 1.

(c) Explain why the particular initialization in (a) is bad.

Hint: Think about the joint distribution implied by the initial parameters. Does it have any additional independence structure that is not implied by the Naive Bayes model?

Problem 9.4

Consider a "Bayesian" hidden Markov model (HMM) in which the parameters are random variables, as described by the following directed acyclic graph.



In this (homogenous) model, given realizations $\bar{\theta} = (\bar{A}, \bar{\eta}, \bar{\pi})$ of the set of random parameters $\theta = (A, \eta, \pi)$, respectively, the transmission, emission, and initial state distributions take the form

$$\begin{aligned} p_{x_{t+1}|x_t,A}(j|i,\bar{A}) &= \bar{a}_{ij} \quad t = 1, 2, \ldots, N-1 \\ p_{y_t|x_t,\eta}(j|i,\bar{\eta}) &= \bar{\eta}_{ij} \quad t = 1, 2, \ldots, N \\ p_{x_1|\pi}(i|\bar{\pi}) &= \bar{\pi}_i \end{aligned}$$

For convenience, we denote the full hidden state and observation sequences by $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ and $\mathsf{y} = (\mathsf{y}_1, \ldots, \mathsf{y}_N)$, respectively. Moreover, we restrict our attention to the case of binary states and observations, i.e., $\mathcal{X} = \mathcal{Y} = \{1, 2\}$.

We place the associated Dirichlet priors on the parameters θ. For example, the first row of $A = [a_{ij}]$ is distributed as

$$\begin{pmatrix} a_{11} & a_{12} \end{pmatrix} \sim \mathrm{Dir}(\lambda_{11}, \lambda_{12})$$

The priors on η and π are defined similarly.

Recall: For a binary variable z whose parameter $\gamma \triangleq p_{\mathsf{z}}(1)$ has distribution $p(\gamma) = \mathrm{Dir}(\alpha_1, \alpha_2)$, the posterior based on samples $\bar{z} = \{\bar{z}_1, \ldots, \bar{z}_K\}$ of z takes the form $p(\gamma|\bar{z}) = \mathrm{Dir}(\alpha_1 + K(1), \alpha_2 + K(2))$, where

$$K(m) = \sum_{k=1}^{K} \mathbb{1}_{z_k = m}, \quad m = 1, 2, \qquad \text{with } \mathbb{1}_{u=v} \triangleq \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}.$$

Given observations $\bar{y}$, we desire the posterior marginals $p(\theta|\bar{y})$ and $p(x_n|\bar{y})$ for $n = 1, \ldots, N$, which we approximate by particle representations from Gibbs sampling. To apply Gibbs sampling, we first sample θ conditioned on sample values for x (and y), then we resample x conditioned on the sample values of θ (and y). We then repeat this resampling process, iterating until convergence.

(a) As needed for parameter resampling, express $p((a_{11}\ a_{12})|\bar{x}, \bar{y})$ in terms of the appropriate count statistics over $\bar{x}$ and $\bar{y}$.

(b) One approach to state sequence resampling is to choose one x_t variable at a time and resample it conditioned on fixed values for the other state variables (and parameters and observations). Express the distribution $p(x_t|\bar{x}_{\setminus t}, \bar{\theta}, \bar{y})$ in terms of the HMM transition, emission, and initial state distributions, where $\mathsf{x}_{\setminus t}$ denotes $\mathsf{x} \setminus \{\mathsf{x}_t\}$.

(c) A more efficient alternative to state sequence resampling is to resample the entire sequence x at once, a method called *blocked* Gibbs sampling.

The forward-backward inference algorithm can be exploited to efficiently compute the required sample sequence $\bar{x}$ from $p(x|\bar{\theta}, \bar{y})$, which would otherwise require exponential complexity in N. Recall that the backward messages computed by the algorithm are

$$\beta_i(x_i) = p(\bar{y}_{i+1}, \ldots, \bar{y}_N|\mathsf{x}_i = x_i, \theta = \bar{\theta}) \quad i = 1, 2, \ldots, N-1$$

Show how to use the β messages together with the parameters $\bar{\theta}$ and observations $\bar{y}$ to efficiently produce a sample sequence $\bar{x}$.

Hint: To produce a sample from a Markov chain, one can sample the root node and recursively sample the subsequent node given its parent.

6.438 Algorithms for Inference

Fall 2014