Probablistic Graphical Models, Spring 2007
Homework 4

Due at the beginning of class on 11/26/07

Instructions

There are four questions in this homework. The last question involves some programming which should be done in MATLAB. Do *not* attach your code to the writeup. Make a tarball of your code and output, name it <userid>.tgz where your userid is your CS or Andrew id, and copy it to /afs/cs/academic/class/10708-f07/hw4/<userid> .

If you are not submitting this from an SCS machine, you might need to authenticate yourself first. See `http://www.cs.cmu.edu/~help/afs/cross_realm.html` for instructions. If you are not a CMU student and don't have a CS or Andrew id, email your submission to 10708-07-instr@cs.cmu.edu.

You are allowed to use any of (and only) the material distributed in class for the homeworks. This includes the slides and the handouts given in the class[1]. Refer to the web page for policies regarding collaboration, due dates, and extensions.

1 [20 pts]Importance Sampling

To do this question, you will need to read (Koller& Friedman, 10.2.2). The likelihood weighting of an importance sampler is defined $w(x) = P(x)/Q(x)$ where P is the distribution we want to sample from and Q is a proposal distribution.

1. Why is computing the probability of a complete instantiation of the variables in a Markov Random Field computationally intractable?
2. Given a chordal graph, describe how to compute the likelihood weighting for an importance sampler(*Hint:What is the relationship between chordal graphs and junction trees?*)
3. Given a non-chordal graph, describe how to compute the likelihood weighting for an importance sampler.
4. Briefly comment on why it is not useful to use importance sampling for approximate inference on MRFs.

2 [20 pts] BP in sigmoid networks

Consider a three-layer sigmoid network ($X_1, \ldots X_n, Y_1, \ldots, Y_n, Z_1 \ldots Z_n$). All the variables are binary; the variables $X_1, \ldots X_n$ in the top layer are all independent of each other; the variables in the second and third layers depend on the variables in their previous layer CPT:

$$P(y_j|x_1 \ldots x_n) = sigmoid(\sum_i w1_{i,j} x_i + w1_{0,j})$$

$$P(z_j|y_1 \ldots y_n) = sigmoid(\sum_i w2_{i,j} y_i + w2_{0,j})$$

where $sigmoid(x) = 1/(1+\exp^{-t})$.

[1] Please contact Monica Hopes(meh@cs)if you need a copy of a handout.

1. Write down the belief propagation updates for the network.

2. Is the above graph loopy? Are the updates in the previous section guaranteed to converge?

3 [30 pts] Mean Field Variational Inference for Admixture Models

Let us consider a specific case of the Admixture model for text, that uses a Dirichlet prior. This model is commonly known as the Latent Dirichlet Allocation (LDA) model. The graphical representation of the model is shown in figure 1(a). The complete data log-likelihood is given as follows:

$$\begin{aligned} & P(\{\mathbf{w}_d\}_{d=1}^M, \{\{\mathbf{z}_{dn(.)}\}_{n=1}^{N_d}\}_{d=1}^M, \{\theta_d\}_{d=1}^M | \{\beta_k\}_{k=1}^K, \alpha) \\ = & \prod_{d=1}^M \left\{ P(\theta_d|\alpha) \prod_{n=1}^{N_d} \prod_{k=1}^K (P(z_{dnk}=1|\theta_d) P(w_{dn}|\beta_k))^{z_{dnk}} \right\} \\ = & \prod_{d=1}^M \left\{ \left(\frac{\Gamma(K\alpha)}{(\Gamma(\alpha))^K} \prod_{k=1}^K (\theta_{dk})^{\alpha-1} \right) \prod_{n=1}^{N_d} \prod_{k=1}^K (\theta_{dk} \beta_{kw_{dn}})^{z_{dnk}} \right\} \end{aligned} \tag{1}$$

where M is the number of documents, N_d is the length of document d, K is the number of topics, n is the position in the document and k is the topic index. $\mathbf{w}_d = (w_{d1}, \cdots, w_{dN_d})$ is the text of the document, where $w_{dn} \in \{1, \cdots, V\}$ is the word at the n^{th} position in document d, where V is the vocabulary size.

- The vector $\mathbf{z}_{dn(.)} = (z_{dn1}, \cdots, z_{dnK})$ is a topic indicator vector for position n in the document d, such that $z_{dnk} \in \{0,1\}$ and $\sum_k z_{dnk} = 1$,
- $\theta_d = (\theta_{d1}, \cdots, \theta_{dK}) | \sum_{k=1}^K \theta_{dk} = 1; \theta_{dk} \geq 0$ is the multinomial topic mixing proportions for document d,
- α is the parameter of the symmetric Dirichlet distribution that generates θ_d for any document d,
- $\beta_k = (\beta_{k1}, \cdots, \beta_{kV}) | \sum_{w=1}^V \beta_{kw} = 1; \beta_{kw} \geq 0$ is a multinomial distribution over the vocabulary for topic k, and
- $\Gamma()$ is the Gamma function.

1. This model is intractable for exact inference of $P(\theta_d, \mathbf{z}_{dn(.)} | \mathbf{w}, \beta, \alpha)$. Let us actually understand and verify this fact: construct a junction-tree for a single document d with $N_d = 3$. In other words, assume the document is (w_1, w_2, w_3). What is the maximum tree-width? Does the tree-width grow with increasing N_d? Write an expression for the marginal for $P(z_1 = k | \beta, \alpha)$ and $P(\theta_d | \beta, \alpha)$ in terms of the messages on the junction tree. Is it possible to compute the marginals? what is the main reason for intractability of exact inference?

2. Let us use mean-field variational technique for approximate inference. In particular, let us approximate the posterior by the following distribution:

$$Q(\theta_d, \{\mathbf{z}_{dn}\}_{n=1}^{N_d} | \gamma_d, \{\phi_{dn}\}_{n=1}^{N_d}) = \left\{ P(\theta_d|\gamma_d) \prod_{n=1}^{N_d} (\phi_{dnk})^{z_{dnk}} \right\} \tag{2}$$

where $\phi_{dn} = (\phi_{dn1}, \cdots, \phi_{dnK})$ is a variational multinomial distribution over topics for position n in the document d and $\gamma_d = (\gamma_{d1}, \cdots, \gamma_{dK})$ are the parameters of a variational Dirichlet distribution given by:

$$P(\theta_d|\gamma_d) = \frac{\Gamma(\sum_k \gamma_{dk})}{\prod_k \Gamma(\gamma_{dk})} \prod_{k=1}^K (\theta_{dk})^{\gamma_{dk}-1} \tag{3}$$

The graphical representation for the variational distribution is given in figure 1(b).

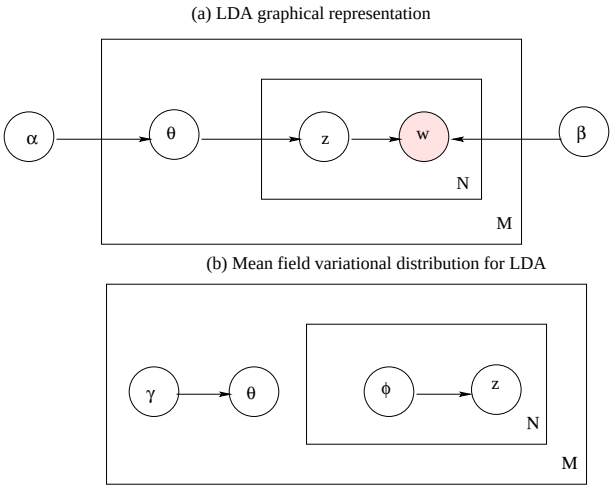

Figure 1: Admixture model: graphical representation

Starting with the results of the mean field variational inference given in Eq. (4) below, derive closed form expressions for the variational parameters γ_{dk} and ϕ_{dnk}.

$$Q(x) = \frac{1}{Z}\exp(\sum_{\Phi:X\in \text{scope}(\Phi)} E_{Q(U)}[\ln \Phi(U_\Phi, x)]) \tag{4}$$

where Φ is a factor in the graphical model and $U_\Phi = \text{scope}(\Phi) - X$. In your derivation, you will use the following properties of the Dirichlet distribution:

$$E_{Q(\theta_d|\gamma_d)}[\log \theta_{dk}] = \Psi(\gamma_{dk}) - \Psi(\sum_{k'=1}^{K} \gamma_{dk'}) \tag{5}$$

where $\Psi(\gamma_{dk}) = \frac{d}{d\gamma_{dk}} \log\Gamma(\gamma_{dk})$ is the digamma function. Prove the result above using the properties of the exponential family we discussed in class. (Hint: convert Dirichlet to its natural parametrization and use the property that the expected value of a sufficient statistic is equal to the first derivative of the log-partition function w.r.t the corresponding natural parameter.)

3. Now, let us consider the Logistic Normal admixture model, known in common parlance as the Correlated Topic Model. It differs from LDA in only that the symmetric Dirichlet prior with parameter α is replaced by a Logistic normal distribution, which is given as follows:

$$P(\eta_d|\mu,\Sigma) \quad = \quad \mathcal{N}(\mu,\Sigma) \tag{6}$$

where $\eta_d = (\eta_{d1}, \cdots, \eta_{dK})$ with each $\eta_{dk} \in \mathbf{R}$. Each θ_{dk} is a simple logistic transformation of η_{dk} given by

$$\theta_{dk} = \frac{\exp(\eta_{dk})}{\sum_{k'=1}^{K}\exp(\eta_{dk'})} \tag{7}$$

Using a logistic normal distribution as described above, allows us to capture correlations in topics, given by the matrix Σ.

Note that the description of logistic normal distribution above is actually an overcomplete representation since θ_d has only $(K-1)$ free parameters, but η_d has K free parameters.

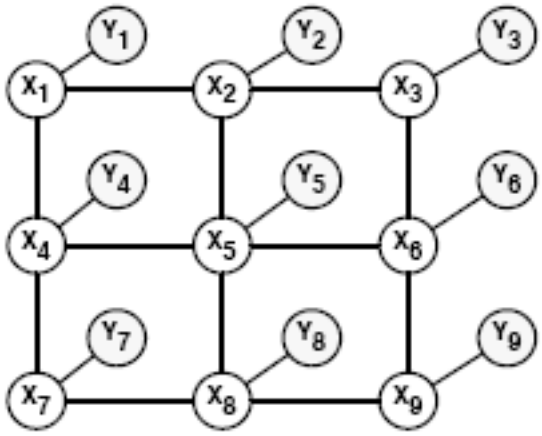

Figure 2: Graphical model for Segmentation

(a) Suggest a simple transformation to solve this problem.

(b) Let us use the following mean-field variational distribution for approximate inference:

$$Q(\eta_d, \{\mathbf{z}_{dn}\}_{n=1}^{N_d} | \lambda_d, \gamma_d, \{\phi_{dn}\}_{n=1}^{N_d}) = \prod_{k=1}^{K} \left\{ \mathcal{N}(\eta_{dk} | \lambda_{dk}, \nu_{dk}) \right\} \prod_{n=1}^{N_d} (\phi_{dnk})^{z_{dnk}} \tag{8}$$

where we chose to use independent normal distributions for each topic in each document. Are you still able to derive a closed-form relation for the variational parameters? If yes, why? If not, why not? (justify your answer in terms of the relationship between the distributions used in the model.)

4 [25 pts] Image Segmentation

Given an image, a K-ary segmentation is a clustering that assigns each pixel to one of K-classes, typically under the assumption that neighbouring pixels are more likely to belong to the same class. As discussed in class, we could represent this as a pairwise Markov Random Field where each node corresponds to a pixel. Note that the value of a node is the cluster it belongs to.

Formally, the observed image is denoted $Y = \{Y_i\}$ and $X = \{X_i\}$, $X_i \in \{1 \dots K\}$ is the segmentation. The markov random field has distribution

$$P(X,Y) = \frac{1}{Z} \prod_i \phi(x_i, y_i) \prod_{(i,j) \in E|} \psi(x_i, x_j)$$

where ϕ is the node potential, the effect y_i has on the label of x_i; ψ is the edge potential, how the label of x_i is influenced by the labels of its neighbors. Let

$$\phi(x_i, y_i) = exp(\frac{-(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2})$$

$$\psi(x_i, x_j) = exp(-\beta(x_i - x_j)^2)$$

Consider the image in Figure 2. We want to produce a binary segmentation (K=2). You are given the following parameters: $\beta = 20, \mu_1 = 147, \sigma_1^2 = 1/2, \mu_2 = 150, \sigma_2^2 = 1/2$. `img.dat` in hw4data.tar.gz contains this image. You may use the helper functions provided in hw4data.tar.gz to plot the image and segmentations.

1. Perform a binary segmentation using naive mean field approximation.

2. Perform a binary segmentation using loopy belief propagation.

Produce plots of your segmentations which must be included in your writeup. Initialize $m_{ij} = 1 \forall i \neq j$. Stop running your algorithm when the absolute difference between your old message and new message is less then 10^{-5}. Show a separate plot of the original image(ie plot y_i's).

Remember to compute the messages in log space, for numerical stability. You don't need to dampen messages in your implementation of loopy BP.

Submit a code along with a README describing your code briefly, along with instructions on how to run it.