10708 Graphical Models: Homework 3

Due Monday, April 1, beginning of class

March 18, 2013

Instructions: There are five questions on this assignment. There is a problem involves coding. You can program in whatever language you like, although we suggest MATLAB. Do *not* attach your code to the writeup. Instead, put your code in a directory called "andrewid-HW3" and tar it into a tgz named "andrewid-HW3". For example, epxing-HW3.tgz. Email your tgz file ONLY to gunhee@cs.cmu.edu, seunghak@cs.cmu.edu and kpuniyan@cs.cmu.edu. Refer to the web page for the policies regarding collaboration, due dates, extensions, and late days.

1 Conditional Random Fields [15 points]

[Exercise 5.16, Daphne Koller and Nir Friedman]

In the class, we have learned that the linear-chain CRF is the sequential version of logistic regression. In this problem, we will show that the naive Markov model of CRF corresponds to the logistic regression. Consider a CRF model over the l-valued variables $\mathbf{X} = \{X_1, \dots, X_k\}$ and m-valued $\mathbf{Y} = \{Y\}$, with the pairwise potentials defined via the following log-linear models

$$\phi_i(x_i^l, y^m) = \exp\{w_i^{ml}\mathbf{I}(X_i = x_i^l, Y = y^m)\}.$$

Again, we have a single-node potential $\phi_0 = \exp\{w_0^m \mathbf{I}(Y = y^m)\}$. Here $\mathbf{I}(\cdot)$ is an indicator function. Show that the CPD $P(Y|X_1, \dots, X_k)$ defined by this model is symbolically identical to that of multinomial logistic model. (*i.e.*, they belong to the same general class of models, so find out what is that general class, and what is the mathematical form of it). (*Hint*: You can start from the definition of CRF shown in Eq.(4.11) of Koller &Friedman).

2 Belief Propagation and Bethe Free Energy [15 points]

In the class, we have learned that BP fixed points correspond to the minimum stationary points of the Bethe approximation of the free energy for a factor graph, and the BP algorithm can attain only an approximation (not exact) to the true distribution $P(x_1, \ldots, x_N)$ when the factor graph has cycles. In this problem, we will see a simple example where a distribution that minimizes the Bethe free energy is not even a valid distribution.

Consider a simple factor graph with three binary variable nodes (x_1, x_2, x_3), where each pair of nodes is connected by a factor node. Suppose that one-node and two-node beliefs are given as follows.

$$b_1(x_1) = b_1(x_2) = b_3(x_3) = (0.5\ 0.5)$$

$$b_{12}(x_1, x_2) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}, \quad b_{23}(x_2, x_3) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}, \quad b_{13}(x_1, x_3) = \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}.$$

1. (7 pts) Show that the above belief set satisfies the normalization condition and marginalization conditions.
2. (8 pts) Show that there can be no distribution $P(x_1, x_2, x_3)$ that has the above belief set as its marginals. (Hint: Compute $P(x_1 = 0, x_2 = 0, x_3 = 0), \ldots, P(x_1 = 1, x_2 = 1, x_3 = 1)$ and show that they cannot constitute a valid distribution).

3 The Marginal Polytope [15 points]

Example 3.8 in Wainwright and Jordan illustrates the marginal polytope of an Ising model with two variables. Now let us consider an Ising Model with three variables X_1, X_2, and X_3 (that take on values in $\{0, 1\}$) with edges $\{X_1, X_2\}$ and $\{X_2, X_3\}$.

1. (7 pts) List the points that are the corners of the marginal polytope.
2. (8 pts) List 8 constraints that are part of the half space representation of the marginal polytope.

4 Variational Inference in Latent Dirichlet Allocation (LDA) [30 marks]

The LDA graphical model (Figure 1) was discussed in class. The most popular use for LDA is in modeling a document collection by topics, however, LDA-like models can also be used for various other modeling tasks. In this question, we will apply LDA to the problem of discovering human ancestry.

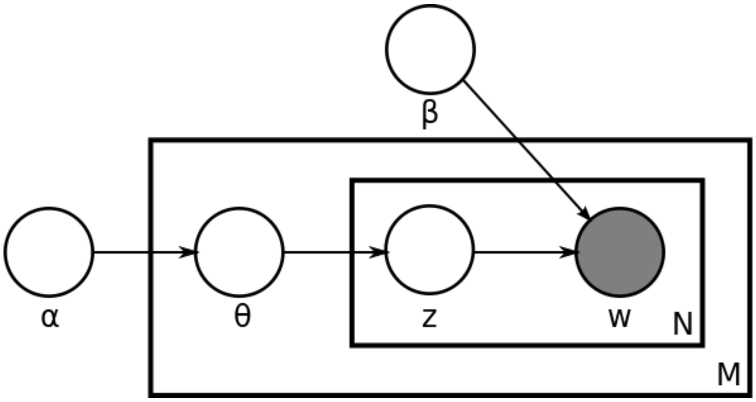

Figure 1: The LDA graphical model

In applications of population genetics, it is often useful to classify individuals in a sample into populations. An underlying assumption is that there are K ancestor populations, and each individual is an admixture of the ancestor populations. For example, in studies of human evolution, the population is often considered to be the unit of interest, and a great deal of work has focused on learning about the evolutionary relationships of modern populations.

For each individual, we measure some genetic data about them, called genotype data. Each genotype is a locus that can take a discrete count value, individuals with similar genotypes are expected to belong to the same ancestor populations. We can derive the admixture coefficients (θ) for each individual by running an LDA model, where the documents are the individuals, and the words are the genotype.

In this question, we will implement variational inference to infer the population mixture (θ) and the genotype ancestry (topic) assignments (z) for any individual. The variational distribution used to approximate the posterior (for a given individual i) is $q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N_i} q(z_n|\phi_n)$, where the Dirichlet parameter γ and the multinomial parameters $(\phi_1, \cdots, \phi_{N_i})$ are the free variational parameters (N_i is the number of non-zero genotype loci for this individual). See Figure 2 for a graphical representation.

The data matrix in **data.mat** provides data about $M = 100$ individuals, each represented by a vocabulary of $N = 200$ genotype loci. This data has been preprocessed into a count matrix D of size $M \times N$. $D_{ij} = 1$ represents the value of genotype j in individual i.

We learnt the LDA topic model over $K = 4$ ancestor populations, and the inferred β matrix of size $N \times K$ has been stored in **beta_matrix** in **data.mat**. The value of α is 0.1.

In the writeup, report the following:

1. Report the variational inference update equations for estimating γ and ϕ (you don't have to derive them).

2. For individual 1, run LDA inference to find ϕ for each genotype locus, store it as a

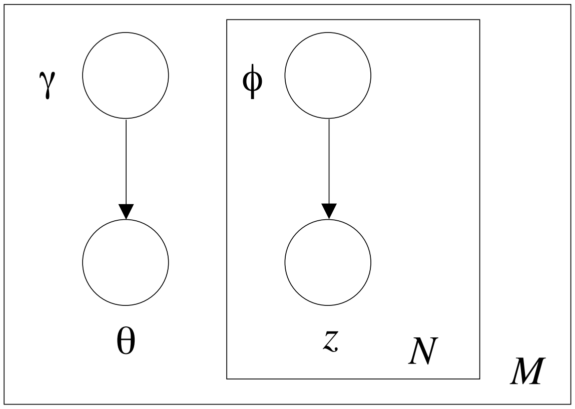

Figure 2: Graphical model representation of the variational distribution used to approximate the posterior in LDA.

matrix of size $n_1 \times K$ (where n_1 is the number of non-zero genotype loci present in individual 1), and plot it as an image in your writeup (Use $imagesc(\phi); colorbar$ in matlab). Don't forget to show the colormap using the *colorbar* function to allow the colors in the image to be mapped to numbers!

3. We will construct a matrix Θ of size $M \times K$ to represent the ancestor assignments for all individuals in the population. For each individual i, run LDA inference to find γ, and store it as row of Θ, i.e. $\Theta_i = \gamma$. Visualize Θ as an image (Use $imagesc(\Theta); colorbar$ in matlab;) and print it in your write up.

4. Report the number of iterations needed to get to convergence for running inference on all M individuals (check the convergence criteria in the "implementation hints" section below).

5. Report the time taken to run inference on all M individuals.

6. Repeat the experiment for $\alpha = 0.01$, $\alpha = 1$, $\alpha = 10$, and for each value of α, visualize the Θ matrix summarizing the ancestor population assignments for all individuals. Discuss the changes in the ancestor population assignments to the individuals as α changes. Does the mean number of iterations required for convergence for inference change as α changes?

Implementation hints:

1. If you use matlab, **beta** is a pre-defined function for the beta function, hence you might want to not use beta as a variable name to avoid overloading.

2. In this assignment, regular updates will most likely work fine, since the vocabulary size (number of genotype loci) is so small, but if you wanted a usable implementation for other problems, updating probabilities would need to be done in log-space to avoid

overflow and underflow issues.

3. Your convergence criteria should be that the absolute change in EACH value of γ AND ϕ is less than ϵ (Use $\epsilon = 1e-3$).

5 Collapsed Gibbs Sampling for LDA [25 marks]

In this problem, we will derive collapsed Gibbs sampling equations for Latent Dirichlet Allocation (LDA) with conditional probabilities:

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}) \tag{1}$$

$$\boldsymbol{\theta}_i \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{2}$$

$$z_{ji}|\theta_i \sim \text{Discrete}(\boldsymbol{\theta}_i) \tag{3}$$

$$d_{ji}|z_{ji}, \boldsymbol{\phi}_{z_{ji}} \sim \text{Discrete}(\boldsymbol{\phi}_{z_{ji}}) \tag{4}$$

Here j is the index for words ($\boldsymbol{d}_i = \{d_{1i}, \ldots, d_{Ni}\}$), i is the index for documents, and k is the index for topics. Also, we use the following notation: $N_{wki} = |\{j : d_{ji} = w, z_{ji} = k\}|$ (total number of times the word w is assigned to the topic k), $N_{ki} = \sum_w N_{wki}$, and $N_{wk} = \sum_i N_{wki}$. We use superscript $(-ji)$ (e.g. $N_{wki}^{(-ji)}$) to indicate that the corresponding word d_{ji} in document i is not counted in N_{wki}.

1. [2 pts] Write down $P(\boldsymbol{d}|\boldsymbol{z})$ and $P(\boldsymbol{z})$ using their conditional probabilities. (Hint: Integrate out $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively)

2. [1 pts] Exact probabilistic inference on $p(\boldsymbol{z}|\boldsymbol{d})$ is infeasible. Explain the reason why the exact inference is infeasible.

3. [2 pts] Gibbs sampling is a particular instance of Metropolis-Hastings algorithm. Show that Gibbs sampling can be viewed as a Metropolis-Hastings algorithm with acceptance probability 1.

4. [10 pts] Since exact inference is infeasible, we will use approximate inference. In particular, in this problem, we are interested in collapsed Gibb's sampling (It is called "collapsed" Gibb's sampling since $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are integrated out in the inference procedure). Prove the following LDA collapsed Gibb's sampling equation:

$$p(z_{ji} = k|\boldsymbol{z}\backslash z_{ji}, \boldsymbol{d}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto (N_{ki}^{(-ji)} + \alpha_k)\frac{N_{wk}^{(-ji)} + \beta_w}{N_k^{(-ji)} + \sum_w \beta},$$

   where $w = d_{ji}$.
   (Hint: $\Gamma(x+1) = x \times \Gamma(x)$)

5. [5 pts] Note that $\boldsymbol{\theta}_i$ (document-topic proportion) and $\boldsymbol{\phi}_k$ (topic-word distribution) can be represented by using only z_{ji} (topic assignment for each word d_{ji} in document i).

Write down θ_{ik} and ϕ_{kj} using only z_{ji}, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

6. [5 pts] Write down pseudo-code for LDA collapsed Gibbs Sampling.