Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: April 20, 2019

QUESTION 1

BBVI Paper

Many models are non conjugate and inferencing on these complex and interesting models are generally intractable or very slow using MCMC methods. BBVI tries to solve this by leveraging gradient based optimization techniques.

For any variational inference problem we can write the Evidence Lower Bound (ELBO) as:

$$ELBO = \mathbb{E}_q\{\log p(X,Z) - \log q(Z)\}$$

To maximize this lower bound we can do gradient ascent under the dominant convergence constraint and estimate the gradient using monte carlo samples as follows:

$$\nabla_\lambda ELBO = \mathbb{E}_q[\nabla_\lambda \log q(Z|\lambda)(\log p(X,Z) - \log q(Z|\lambda)]$$

Sometimes when number of variables is large the monte carlo estimate may introduce large variance which is not useful, so paper proposes two methods to solve these problem.

Controlling Variance

Rao-Blackwellized
It tries ti reduce the variance of gradient by replacing monte carlo estimate with its conditional expectation with respect to a subset of variables.

$$\hat{J}(x) = \mathbb{E}[J(X,Y)|X]$$

Its easy to show that $\mathbb{E}[\hat{J}(x)] = \mathbb{E}[J(X,Y)]$, also if we look at the variance of this estimate:

$$var(\hat{J}(X) = var(J(X,Y)) - \mathbb{E}[(J(X,Y) - \hat{J}(X))^2]$$

and so variance of this estimate is lower than the real one without change in expected value.

Control Variates

It replaces the function having higher variance with function having lower variance but same expectation.
Consider,

$$\hat{f} = f - a(h(Z) - \mathbb{E}(h(Z))$$

We can easily show that the expectation of this estimate is same as that of real one. To minimize the variance of this estimate we differentiating the variance of this estimate and equating to zero to get optimal value for **a** as:

$$a^* = Cov(f, h)/Var(h)$$

In our specific example above given methods can be used under mean field variational assumption with rao-blackwellized included as follows:

$$\begin{aligned} h_i &= \nabla_{\lambda_i} \log q_i(z|\lambda) \\ f_i &= \nabla_{\lambda_i} \log q_i(z|\lambda_i)(\log p_i(X, z) - \log q_i(z|\lambda_i)) \\ a_i* &= \frac{\sum_{d=1}^{n_i} Cov(f_i^d, h_i^d)}{\sum_{d=1}^{n_i} Var(h_i^d)} \end{aligned}$$

and gradients becomes :

$$\nabla_{\lambda_i} L = \frac{1}{S} \sum_s \nabla_{\lambda_i} \log q_i(z_s|\lambda_i)(\log p_i(X, z_s) - \log q_i(z_s|\lambda_i) - a_i^*)$$

Note that the expected gradient remains same here but with lower variance.

Above given methods can be extended to Adagrad for adaptive gradients in each dimesion, and can be done stochastically as given in paper.

This concludes the answer to this question.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: April 20, 2019

QUESTION

2

Relational LDA

Given,
LDA model with No. Documents : D
A : doc2doc binary link matrix.

We need to model for A as well here. So for each cell in A we can sample values from a bernoulli distribution where the bernnoulli parameter can be represented as ψ.

Generative Story

1. For document d in corpus:
   (a) Draw Topic proportion $\theta_d \sim Dir(\alpha)$
   (b) For each word in document d:
      i. Draw topic label : $z_{d,n} \sim Mult(\theta_d)$
      ii. Draw word : $w_{d,n} \sim Mult(\phi_{z_{n,d}})$
2. For a pair of documents d1,d2:
   (a) sample the link as $A_{d1,d2} \sim Bern(\psi(z1, z2))$

Now we need to define ψ which follows the constrained of being a values from 0 to 1. We can first take the aggregates of Z matrices to get mixture compoents of topics in each documents as $\bar{z}_d$ (a Kx1 vector).
No we can define ψ in following way:

$$\psi(z1, z2) = \sigma(\eta^T(\bar{z}_{d1} \circ \bar{z}_{d2}) + \nu)$$

or

$$\psi(z1, z2) = \exp(-\beta||\bar{z}_{d1} - \bar{z}_{d2}||_2)$$

This concludes the answer to this question.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: April 20, 2019

QUESTION

3

Stochastic Block Model

Given generative story,

1. For observation n:
   (a) $z_n \sim Mult(\pi)$
2. For each n in 1:N,
   (a) For each m in 1:N-1,
      i. $A_{nm} \sim Bern(\eta_{z_n}, \eta_{z_m})$

given these priors,

$$\eta_{kl} \sim Beta(a, b)$$
$$\pi \sim Dir(\alpha, ..., \alpha)$$

Let's find out joint distribution (ignoring the global parameters in notation).

$$\begin{aligned} p(A, Z, \eta, \pi) &= p(A|Z, \eta, \pi)p(Z|\pi)p(\pi)p(\eta) \\ &= [\prod_{n=1}^{N}\prod_{m=1}^{M} p(A_{nm}|z_n, z_m, \eta)] \times [\prod_{n=1}^{N}\prod_{k=1}^{K} p(z_n|\pi)] \times p(\pi) \times [\prod_{l=1}^{K}\prod_{k=1}^{K} p(\eta_{lk})] \end{aligned}$$

We can look into the log joint probability as follows:

$$\begin{aligned} L = &\sum_{n=1}^{N}\sum_{m=1}^{M} A_{nm}\log(\eta_{z_n,z_m}) + (1 - A_{nm})\log(1 - \eta_{z_n,z_m}) \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\log(\pi_k) + \sum_{k=1}^{K}(\alpha - 1)\log(\pi_k) \\ &+ \sum_{l=1}^{K}\sum_{k=1}^{K}(\alpha - 1)\log(\eta_{lk}) + (\beta - 1)\log(1 - \eta_{lk}) \end{aligned}$$

By ignoring the terms that do not contain the respective terms in log joint we can find posteriors distribution as follows (Skipping derivation steps):

$$p(\pi|Z) = Dir(\alpha + n_1, ..., \alpha + n_K)$$
$$p(\eta_{kl}|A, Z) = Beta(a + \Omega_{kl}, b + \rho_{kl})$$

$$p(z_p = k|A, \eta, \pi) = \frac{\pi_{z_p} \times [\prod_{m \neq p} \eta_{k,z_m}^{A_{pm}} (1 - \eta_{k,z_m})^{1-A_{pm}}] \times [\prod_{n \neq p} \eta_{z_n,k}^{A_{np}} (1 - \eta_{z_n,k})^{1-A_{np}}]}{\sum_{l=1}^{K} \pi_{z_p} \times [\prod_{m \neq p} \eta_{l,z_m}^{A_{pm}} (1 - \eta_{l,z_m})^{1-A_{pm}}] \times [\prod_{n \neq p} \eta_{z_n,l}^{A_{np}} (1 - \eta_{z_n,l})^{1-A_{np}}]}$$

where,
$n_l = \sum_{n=1}^{N} z_{nl}$,
$\Omega_{kl} = \sum_n \sum_m I(z_n = k, z_m = l) A_{nm}$,
$\rho_{kl} = \sum_n \sum_m I(z_n = k, z_m = l)(1 - A_{nm})$.

Now we can have gibbs sampler using these posteriors.
This concludes the answer to above question.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: April 20, 2019

QUESTION

4

Dirichlet Process Conjugacy

Given a Dirichlet process we know, for any finite measurable sets this relation follows:

$$G \sim DP(\alpha_0, G_0)$$

$$[G(A_1), G(A_2), ...G(A_r)] \sim Dir(\alpha_0 G_0(A_1), .., \alpha_0 G_0(A_r))$$

which can be seen as infinite dimensional Dirichlet distribution with other partition having zero probability. So we can use this representation to get posterior over G finite dimensional representation as follows:

Let's consider a single point θ_1 instead of a sequence..

$$\begin{aligned} p(G|\theta_1) &\propto p(\theta_1|G)p(G) \\ &\propto \prod_{k=1}^{K} \pi_k^{\delta_{\theta_1}(A(k))} \times \frac{1}{\beta(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_0 G_0(\phi_k)-1} \\ &\propto \prod_{k=1}^{K} \pi_k^{\alpha_0 G_0(\phi_k)+\delta_{\theta_1}(A(k))-1} \\ &= Dir((\alpha_1)G_1(\phi_1), ..., (\alpha_1)G_1(\phi_K)) \end{aligned}$$

where, $\delta_{\theta_1}(A(k)) = 1$, if $\theta_1 \in A(k)$ otherwise 0

$$\alpha_1 = \alpha_0 + 1$$

$$G_1 = \frac{\alpha_0}{\alpha_0 + 1} G_0 + \frac{1}{\alpha_0 + 1} \sum_{k}^{K} \delta_{\theta_1}(A(k))$$

Similarly we can do it for infinite dimensional Dirichlet distribution (DP) where K is ∞ and we will still obtain a similar result in infinite dimensions so we can say,

$$G|\theta_1 \sim DP(\alpha_1, G_1)$$

Note: Now we are again where we started but with a new Dirichlet process so we can add other θ's one by one sequentially in similar fashion under exchangeable assumptions to generalize above equation as:

$$G|\theta_1, ..., \theta_N \sim DP(\alpha_N, G_N)$$

where,

$$\alpha_N = \alpha_0 + N$$

$$
\begin{aligned}
G_N &= \frac{\alpha_0}{\alpha_0 + N} G_0 + \frac{1}{\alpha_0 + N} \sum_{k}^{K} \sum_{n}^{N} \delta_{\theta_n}(A(k)) \\
&= \frac{\alpha_0}{\alpha_0 + N} G_0 + \frac{1}{\alpha_0 + N} \sum_{k}^{K} n_k \delta_{\phi_k}
\end{aligned}
$$

Where,

$$
n_k = \sum_{n}^{N} \delta_{\theta_n}(A(k))
$$

This concludes the answer to this question.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Abhishek Kumar
Roll Number: 18111002
Date: April 20, 2019

QUESTION

5

HDP paper

HDP(Heirarchical Dirichlet Process) is an extension of Dirichlet process.

A Dirichlet process is a distribution of a random probability measure G. such that for any finite measurable partition of domain $(A_1, A_2, ...A_r)$ the random vector $(G(A_1), G(A_2), ...G(A_r))$ is dirichlet distributed.

$$[G(A_1), G(A_2), ...G(A_r)] \sim Dir(\alpha_0 G_0(A_1), .., \alpha_0 G_0(A_r))$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

so, we say :

$$G|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

where ϕ_k is a sample from G_0 base distribution. On other hand
A Hierarchical Dirichlet Process is distribution over a set of random probability measures G_j's where the base distribution itself is sampled from another dirichlet process based on some base distribution H. i.e:

$$G_0|\gamma, H \sim DP(\gamma, H)$$

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \ \ \forall j$$

Key Differences

1. G_0 is discrete in HDP and was unrestricted in DP. so samples from G_0 are not atomic.

2. HDP due to atomic property of samples from G_0 exhibits a shared property based grouping or clustering(i.e any new sampled cluster in any group may not be new in G_0). Whereas in DP all the new samples were non redundant since the base distribution was non-atomic with probability 1.

3. HDP is like Admixture model (i.e mixture of mixture model with shared mixtuer components) whereas DP is only a mixture model.

Stick Breaking DP

We can see DP as a stick breaking process in which we given a unit lenght stick break it into pieces with a GEM distribution and sample points from base distribution G_0 accordingly. Where the GEM distribution uses following beta distribution to sample proportion of remaining stick(length of stick remained after k-1 such breaks) to break.

$$\pi'_k | \alpha_0, G_0 \sim Beta(1, \alpha_0)$$

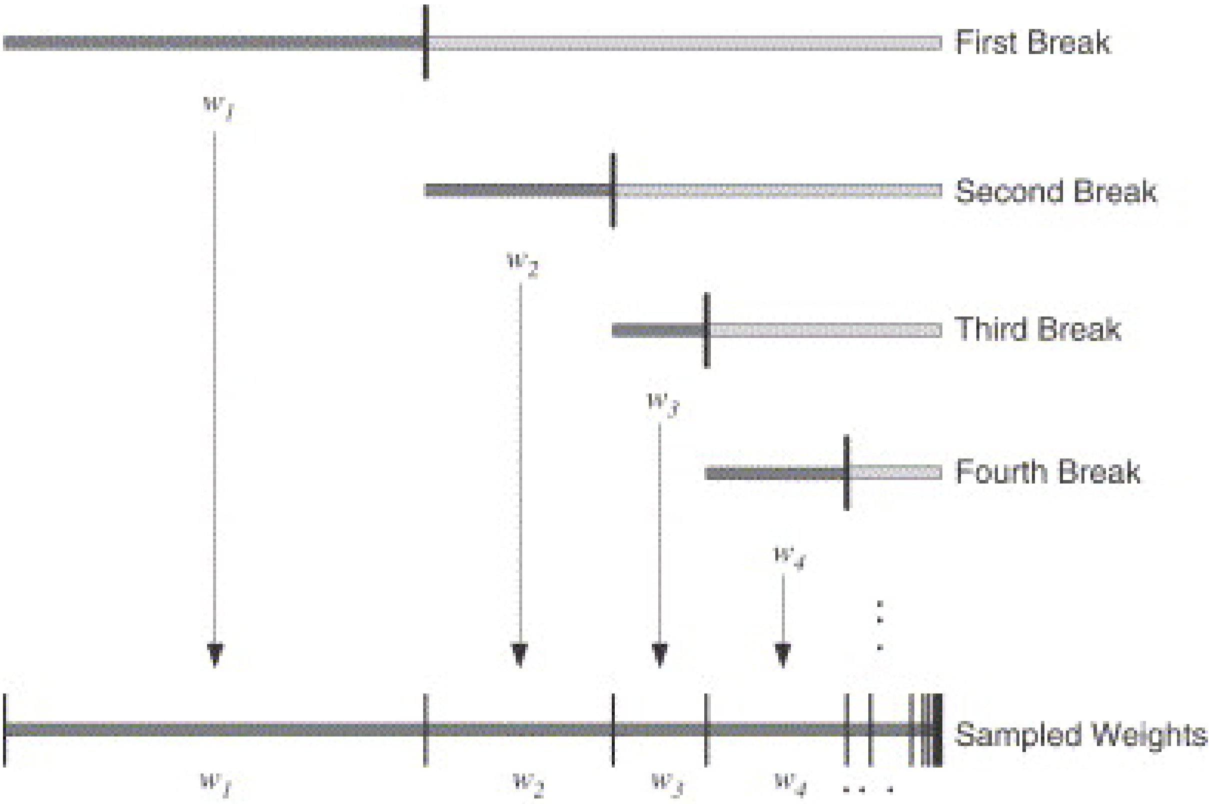

Figure 1: Stick Breaking process Example

Stick Breaking HDP

In HDP similar to DP we have stick breaking on the first level for β with similar GEM distribution as used in DP with samples from H_0 as base distribution. The beta distribution this GEM uses is:

$$\beta'_k | \gamma, H_0 \sim Beta(1, \gamma)$$

Also on the second level of hierarchy considering sampled G_0 as a new base distribution for another stick breaking process over π_k. These π_k's are drawn from a $DP(\alpha_o, \beta)$ which makes sense since G_0 is discrete so samples should be from those atoms only. This formulation can also be seen as a GEM distribution with following beta distribution.

$$\pi'_{jk} \sim Beta(\alpha_0 \beta_k, \alpha_0(1 - \sum_{l=1}^{k} \beta_l))$$

Chinese Restaurant Franchise (CRF) It's an extension of CRP(Chinese restaurant Processes), it includes the possibilities of multiple restaurants with infinte number of tables where each table is serving a particular dish. And since it's a franchise the menu of all the restaurants are shared (i,e all the restaurants have same menu).

CRP does not have this concept of multiple restaurants and particular shared dishes.
CRF can be represented as :

$$\theta_{ji} \rightarrow i^{th} \text{ Customer in } j^{th} \text{ restaurant}$$

$$\psi_{jt} \rightarrow \text{dish of } j^{th} \text{ restaurant at } t^{th} \text{ table}$$

then a new customer can select a table to eat a dish as follows:

$$\theta_{ji}|\theta_{j1},...,\theta_{j,i-1},\alpha_0,G_0 \sim \sum_{t=1}^{m_{j.}} \frac{n_{jt.}}{i-1+\alpha_0}\delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0}G_0$$

and

$$\psi_{jt}|\psi_{11},\psi_{12},..,\psi_{21},..,\psi_{jt-1},\gamma,H \sim \sum_{k=1}^{K} \frac{m_{.k}}{m_{..}+\gamma}\delta_{\phi_k} + \frac{\gamma}{\gamma+m_{..}}H$$

Where, $m_{j.}$ is number of tables in j^{th} restaurant and $n_{jt.}$ is number of customer on t^{th} table of j^{th} restaurant, $m_{.k}$ is number of tables serving k^{th} dish and $m_{..}$ is total number of tables occupied.

This two level hierarchy was not present in CRP. Infact there was no concept of dishes which is intuitive option to have since the restaurants are related so sharing of dishes makes sense.

Note: In first expression the sub groups G_j's are integrated out and similarly in second G_0 is integrated out. And also G_0 is discrete distribution so many θ's in different or groups will points to same ψ (dish).

Intuitive Sense
Multiple datasets that are related have some common/shared properties(mixture components) which needs to be modeled for better optimization. They can be represented as mixture of groups or clusters that share certain properties. For example consider topic modelling in which each document shared same set of topics among them, so in such cases it's better to allow sharing. Which can be naturally obtained using hierarchical formulation like HDP and so joint mixture modelling is useful.

This concludes the answer to this question.