Text Mining for the Social Sciences
Lecture 8: Variational inference

Stephen Hansen

Introduction

We have seen that directly computing the posterior for LDA is intractable.

First option is to stochastically approximate posterior by repeatedly sampling from a Markov chain formed by draws from conditional distributions (Gibbs sampling).

We now cover a more recently popularized approach in Bayesian statistics called variational inference.

As with MCMC, many applications, but we focus on LDA for concreteness. Original article used variational approach.

General Idea

Approximate the true posterior distribution with a simpler functional form that depends on a set of variational parameters.

Then optimize the approximate posterior with respect to the variational parameters so that it lies "close to" the true posterior.

The inference problem becomes an optimization problem.

But note that the family of distributions used to approximate the posterior typically does not include the true posterior.

True and Approximate Distributions

Suppose we have observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$ (treat any parameters as fixed for now).

Let $p(\mathbf{x}, \mathbf{z})$ be their joint distribution.

Assume that $p(\mathbf{z} \mid \mathbf{x})$ is intractable to compute, for example because the latent space is too high-dimensional.

Let $q(\mathbf{z})$ be an approximate distribution over the latent variables. It will depend on variational parameters we suppress for now.

Kullback-Leibler Divergence

To measure the closeness of $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$, we can use the Kullback-Leibler divergence:

$$\mathbb{KL}(p \parallel q) = \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}) \log \left[\frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right] \qquad \text{(forwards KL)}$$

$$\mathbb{KL}(q \parallel p) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(reverse KL)}$$

Forwards KL:

1. "Zero-avoiding"
2. Used in expectation propagation

Reverse KL:

1. "Zero-forcing" $\rightarrow$ better when multi-modal posterior
2. Used in variational inference

Expressing KL Divergence

We can express reverse KL as:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x})} \right] =$$

$$\underbrace{\log \left[p(\mathbf{x}) \right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

Expressing KL Divergence

We can express reverse KL as:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x})} \right] =$$

$$\underbrace{\log \left[p(\mathbf{x}) \right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

$\log [p(\mathbf{x})]$ is hard to compute, but does not depend on $q(\mathbf{z})$.

Minimize KL divergence = maximize ELB, which we can usually compute.

ELB is expected complete data log-likelihood plus entropy of approximating distribution.

Comparison to EM Algorithm

In the EM algorithm, we take the expectation of the complete data log-likelihood with respect to the posterior distribution over $\mathbf{z}$ given fixed parameter values.

We use the true $p(\mathbf{z} \mid \mathbf{x})$ rather than the approximation $q(\mathbf{z})$, so the KL divergence is zero.

The ELB computed using true $p(\mathbf{z} \mid \mathbf{x})$ equals $\log [p(\mathbf{x})]$.

By contrast, with variational inference the ELB is not tight, but we want to make it as tight as possible.

Mean Field Approximation

The space of potential approximating distributions is large, so in practice some restrictions are made.

In mean-field approximation, we assume that q factorizes into M blocks (where M can be less than the dimensionality of $\mathbf{z}$):

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i).$$

Each $q_i(\mathbf{z}_i)$ will have an associated variational parameter(s).

Independence assumptions implicit in mean field approximation generally not present in true posterior.

ELB with Mean Field

Consider set of latent variables $\mathbf{z}_i$, and denote the others by $\mathbf{z}_{-i}$.

Consider dependence of ELB just on $q_i(\mathbf{z}_i)$:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log\left[p(\mathbf{z}, \mathbf{x})\right] = \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \mathbb{E}_{\mathbf{z}_{-i}} \left(\log\left[p(\mathbf{z}, \mathbf{x})\right]\right)$$

$$\begin{aligned}\sum_{\mathbf{z}} q(\mathbf{z}) \log\left[q(\mathbf{z})\right] &= \sum_{\mathbf{z}} \prod_{i=1}^{M} q_i(z_i) \left(\sum_{i=1}^{M} \log[q_i(\mathbf{z}_i)]\right) \\ &= \sum_{\mathbf{z}_i} q_i(\mathbf{z}_i) \log\left[q_i(\mathbf{z}_i)\right] + \text{constant}\end{aligned}$$

Optimal update of $q_i(\mathbf{z}_i)$ satisfies $q_i^*(\mathbf{z}_i) \propto \exp\left[\mathbb{E}_{\mathbf{z}_{-i}} \left(\log\left[p(\mathbf{z}, \mathbf{x})\right]\right)\right]$.

Can also update with $q_i^*(\mathbf{z}_i) \propto \exp\left[\mathbb{E}_{\mathbf{z}_{-i}} \left(\log\left[p(\mathbf{z}_i \mid \mathbf{z}_{-i}, \mathbf{x})\right]\right)\right]$.

Inference

The optimal update equation is a function of $q_j(\mathbf{z}_j)$ for $j \neq i$.

Coordinate ascent algorithm: update each q_i term holding constant the current values of q_{-i}.

Use optimized q_i to approximate posterior distribution.

Interpretation

Form true conditional posterior $p(\mathbf{z}_i \mid \mathbf{z}_{-i}, \mathbf{x})$ (only need to consider Markov blanket of $\mathbf{z}_i$), then take expectation with respect to approximate distribution over the conditioning variables.

Close relationship to Gibbs sampling:

- In Gibbs sampling, we repeatedly sample values from $p(\mathbf{z}_i \mid \mathbf{z}_{-i}, \mathbf{x})$ to simulate true joint distribution.
- In variational inference, we instead average over $p(\mathbf{z}_i \mid \mathbf{z}_{-i}, \mathbf{x})$ rather than take samples.
- Benefit is that analytical averaging "stands in" for collecting many samples.
- But when $\mathbf{z}_i$ is strongly correlated with neighboring nodes, averaging distorts the estimated marginal $q_i(\mathbf{z}_i)$.

Gibbs Sampling / Variational Inference

Advantages of sampling:

1. Typically easier to derive sampling algorithms
2. More accurate, especially for approximating features of posterior distribution beyond the mode

Advantages of variational inference:

1. Faster, especially when optimized (coordinate ascent not the only algorithm)
2. Deterministic
3. Convergence easy to assess

Variational Bayes

Now suppose we wish to approximate posterior over both latent variables $\mathbf{z}$ and parameters $\boldsymbol{\theta}$ given data $\mathbf{x}$.

Mean field assumption is to approximate $p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{x})$ with $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{z}}(\mathbf{z})$, or $q(\boldsymbol{\theta}) \prod_i q_i(\mathbf{z}_i)$ given conditional independence of latent variables.

Can implement VBEM algorithm by alternating between updating $q_i(\mathbf{z}_i)$ given $q(\boldsymbol{\theta})$ (VB E-step), and updating $q(\boldsymbol{\theta})$ given $q_i(\mathbf{z}_i)$ (VB M-step).

Distinction between latent variables and parameters becomes rather artificial, both are treated as unknown quantities and iteratively updated.

Variational Bayes and LDA

We can estimate LDA via Variational Bayes using the mean field approximation

$$p(\Theta, B, \mathbf{z} \mid \mathbf{w}) \approx \prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{d=1}^{D} \left[q(\theta_d \mid \gamma_d) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}) \right]$$

where:

- β_k is Dirichlet with parameters λ_k
- θ_d is Dirichlet with parameters γ_d
- $z_{d,n}$ is multinomial with parameters $\phi_{d,n}$

λ_k, γ_d, and $\phi_{d,n}$ are variational parameters we iteratively update according to the mean-field formula above.

Placing variational distributions in the same family as their priors is without loss of generality within exponential family (see Wainwright and Jordan 2008).

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \text{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^K \right)$$

so

$$\log\left[p(\theta_d \mid \cdot)\right] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k) \right] \log\left(\theta_{d,k}\right) + \text{constant}.$$

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \text{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^K \right)$$

so

$$\log\left[p(\theta_d \mid \cdot)\right] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k) \right] \log\left(\theta_{d,k}\right) + \text{constant}.$$

Taking expectations (ignoring constant) gives

$$\mathbb{E}_{\mathbf{z}_d}[\log\left[p(\theta_d \mid \cdot)\right]] = \sum_k \left[\alpha - 1 + \sum_n \phi_{d,n,k} \right] \log\left(\theta_{d,k}\right)$$

so optimal update is

$$\gamma^*_{d,k} = \alpha + \sum_n \phi_{d,n,k}.$$

Update for λ_k

Recall from Gibbs sampling slides that

$$\beta_k \mid \mathbf{z}, \mathbf{w} \sim \text{Dir}\left([\eta + \sum_d \sum_n \mathbb{1}(z_{d,n} = k)\mathbb{1}(w_{d,n} = v)]_{k=1,v=1}^{K,V} \right)$$

Again taking expectations of log probability, optimal update is

$$\lambda_{k,v}^* = \eta + \sum_d \sum_n \phi_{d,n,k} \mathbb{1}(w_{d,n} = v).$$

Updates for both θ_d and β_k very similar to those for Gibbs sampling, but replacing actual with expected counts.

Update for $\phi_{d,n}$

From the mean field formula and previous results on Gibbs sampling,

$$\phi_{d,n,k} \propto \exp\left(\mathbb{E}\left[\log(\beta_{k,v_{d,n}}\theta_{d,k})\right]\right).$$

Result on Dirichlet: $\mathbb{E}[\log(\theta_i)] = \Psi(\alpha_i) - \Psi(\sum_i \alpha_i)$, so

$$\phi^*_{d,n,k} \propto \exp\left(\Psi\left(\lambda_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda_{k,v}\right) + \Psi\left(\gamma_{d,k}\right)\right).$$

(Ψ function is derivative of $\log(\Gamma)$, implemented in many scientific computing packages).

Overall Algorithm

Seed $\phi^1_{d,n,k} = 1/k$. Then at iteration s:

1. For each topic k (or randomly seed if $s = 1$)

$$\lambda^{s+1}_{k,v} = \eta + \sum_d \sum_n \phi^s_{d,n,k} \mathbb{1}(w_{d,n} = v)$$

2. For each document d

   2.1 $\gamma^{s+1}_{d,k} = \alpha + \sum_n \phi^s_{d,n,k}$

   2.2 For each word n in document d

$$\phi^{s+1}_{d,n,k} \propto \exp\left(\Psi\left(\lambda^{s+1}_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda^{s+1}_{k,v}\right) + \Psi\left(\gamma^{s+1}_{d,k}\right)\right)$$

3. Check convergence of ELB, if not then proceed to iteration $s + 1$

Model Selection

Another advantage of variational inference over MCMC is the optimized ELB provides an estimate of the log evidence $\log [p(\mathbf{x} \mid K)]$, which we can use for model selection.

We can run the above algorithm for different values of K and compare the bound across them.

We need to add a $\log(K!)$ term to the optimized bound to account for multiple modes.

Conclusion

Within context of LDA, we have discussed approximate posterior inference in graphical models using both MCMC and variational approaches.

Neither approach is "better" than another, although for big data applications in text mining, variational methods have a distinct scalability advantage.