E6892 Bayesian Models for Machine Learning

Columbia University, Fall 2015

Lecture 8, 11/5/2015

Instructor: John Paisley

- We next look at scalable variational inference in the context of conjugate exponential family models. We first review these types of models.

  Conjugate exponential family (CEF) models

- General setup: We have a model of data X with variables $\theta_1, \dots, \theta_m$. In combination with their priors, this defines a joint likelihood $p(X, \theta_1, \dots, \theta_m)$.

- A major class of models are called "conjugate exponential family" (CEF) models. The features of this type of model are:

  1. All distributions on variables and data defined in the model are in the exponential family.

  2. If we pick one variable, θ_i, it's conditional posterior $p(\theta_i | X, \theta_{-i})$ is in the same family as the prior distribution on θ_i. That is, all conditional posteriors are conjugate.

- Example: Let's return the the earlier toy model,

$$y_i \sim N(x_i^T w, \alpha^{-1}), \quad w \sim N(0, \lambda^{-1} I), \quad \alpha \sim Gam(a,b), \quad \lambda \sim Gam(c,d)$$

  Checks:

  1. The Gaussian and gamma distributions are exponential family distributions

  2. $p(w|-)$ is Gaussian; $p(\alpha|-)$ is gamma; $p(\lambda|-)$ is gamma

  So it is a CEF model.

- The reason we care about CEF models is that (1) They are pervasive, (2) We can make general statements about variational inference for these types of models.

- From Bishop (slightly modified), we have a conjugate likelihood-prior pair if they can be written as follows:

$$p(x|\eta) = h(x) g(\eta) e^{\eta^T t(x)}, \quad p(\eta | \xi, \nu) = f(\xi, \nu) g(\eta) \mathrm{e}^{\eta^T \xi} \tag{1}$$

Because we gave names to the terms in $p(x|\eta)$ last time, and we are claiming that both $p(x|\eta)$ and $p(\eta|\xi,\nu)$ are exponential family distributions we should map the names to $p(\eta|\xi,\nu)$. In these two distributions, the base measures are $h(x)$ and $g(\nu)$ respectively. The natural parameters are η and ξ respectively. The log normalizers are $\ln g(\eta)$ and $\ln f(\xi,\nu)$ respectively. And the sufficient statistics are $t(x)$ and η respectively. However, this isn't the perspective we take with the prior.

Without belaboring the point too much further. Notice that the log normalizer of $p(x|\eta)$ is the same as the log base measure of $p(\eta|\xi,\nu)$. We'll move on to an example now.

- To give a simple, more general example, let

$$x \sim p(x|\theta_1,\theta_2), \quad \theta_1 \sim p(\theta_1|\theta_3), \quad \theta_2 \sim p(\theta_2), \quad \theta_3 \sim p(\theta_3) \tag{2}$$

Let's focus on θ_1. Using a new notation that can be mapped to Bishop (above), we have

Likelihood:

$$p(x|\theta_1,\theta_2) = h(x,\theta_2)\mathrm{e}^{\eta(\theta_1)^T t(x,\theta_2) - nA(\eta(\theta_1))}$$

Conjugate prior:

$$p(\theta_1|\theta_3) = f(\xi(\theta_3),\nu(\theta_3))\mathrm{e}^{\eta(\theta_1)^T \xi(\theta_3) - \nu(\theta_3)A(\eta(\theta_1))}$$

Posterior:

$$\begin{aligned} p(\theta_1|x,\theta_2,\theta_3) &\propto p(x|\theta_1,\theta_2)p(\theta_1|\theta_3) \\ &\propto \mathrm{e}^{\eta(\theta_1)^T(t(x,\theta_2)+\xi(\theta_3)) - (n+\nu(\theta_3))A(\eta(\theta_1))} \end{aligned} \tag{3}$$

- However, notice that if we want to make this a distribution on θ_1, it's the same form as the prior. That is, we can say that the posterior distribution

$$p(\theta_1|x,\theta_2,\theta_3) = f(\xi',\eta')\mathrm{e}^{\eta(\theta_1)^T\xi' - \nu' A(\eta(\theta_1))} \tag{4}$$

$$\xi' = t(x,\theta_2) + \xi(\theta_3), \quad \nu' = n + \nu(\theta_3)$$

- How does this relate to variational inference? Recall that the optimal $q(\theta_1)$ is found by:

  1. Taking the log of the joint likelihood
  2. Taking the expectation wrt all q distributions except for $q(\theta_1)$
  3. Exponentiating the result
  4. Normalizing it as a function of θ_1

- In a CEF model this optimal q distribution always is in the same family as the prior and has expectations over the terms in the exponent. In the context of the above abstract model, we can already see that

$$q(\theta_1) \propto \exp\{\eta(\theta_1)^T(\mathbb{E}[t(x,\theta_2)] + \mathbb{E}[\xi(\theta_3)]) - (n + \mathbb{E}[\eta(\theta_3)])A(\eta(\theta_1))\} \tag{5}$$

That is,

$$q(\theta_1) = f(\xi', \eta')e^{\eta(\theta_1)^T \xi' - \nu' A(\eta(\theta_1))} \tag{6}$$

$$\xi' = \mathbb{E}[t(x, \theta_2)] + \mathbb{E}[\xi(\theta_3)], \quad \nu' = n + \mathbb{E}[\nu(\theta_3)]$$

- Let's be a little less abstract (but not totally concrete) and look at this in the context of a model *framework*.

Mixed membership models

- Mixed membership models can be broken down into the following components

  1. Grouped data $x_1, \ldots, x_D$

     The data is naturally grouped, and modeled as a group. Here, group is not precise, but can be thought of as being more complex than a simple point in $\mathbb{R}^d$. For example, each group could be a person, and the data could be various vital statistics about that person. The group could also be an audio signal from which is extracted a large set of characterizing features. In topic modeling, the group is the document, and the data is the set of words making up that document.

  2. Global variables β

     These are the model variables that directly impact every group of data. For example, in LDA they are the set of topics from which every word in the data set is drawn.

  3. Local variables $z_1, \ldots, z_D$

     These mirror the grouped data: Model variables in z_i only interact with data in group x_i. Given the global variables, z_i in no way impacts $x_{i'}$ for $i' \neq i$. The variables in z_i can be complicated and have their own hierarchical dependencies. For example, in LDA $z_i = \{\theta_i, c_{i1}, \ldots, c_{in_i}\}$, where θ_i is the distribution on topics and c_{ij} is the indicator of the topic for the jth word in x_i. The c's are generated from a distribution parameterized by θ.

- Mixed membership models also differ from other models in the way z and β combine to generate x probabilistically, but these details aren't necessary for the following discussion. LDA is the most common mixed membership model for discrete data.

- An important property of this type of model is that the joint likelihood factorizes in a nice way,

$$p(x, z, \beta) = p(\beta) \prod_{i=1}^{D} p(x_i, z_i | \beta) \quad \Big(= p(\beta) \prod_{i=1}^{D} p(x_i | z_i, \beta) p(z_i) \Big) \tag{7}$$

For example, for LDA this is

$$p(x, z, \beta) = \Big(\prod_{k} p(\beta_k) \Big) \prod_{i=1}^{D} p(\theta_i) \textstyle\prod_{j=1}^{n_i} p(x_{ij} | \beta_{c_{ij}}) p(c_{ij} | \theta_i)$$

- For the generic model, it follows that the log joint likelihood is

$$\ln p(x, z, \beta) = \sum_{i=1}^{D} \ln p(x_i, z_i|\beta) \; + \; \ln p(\beta) \tag{8}$$

- Since we can't find the exact posterior distribution of all the model variables, we approximate with $q(z, \beta)$ using variational inference. We pick the factorization

$$q(z, \beta) = q(\beta) \prod_{i=1}^{D} q(z_i) \tag{9}$$

  These factorizations can then sub-factorize. For example, with LDA, $q(z_i) = q(\theta_i) \prod_j q(c_{ij})$ (using the definition of z_i from earlier). However, at the level at which we are talking about this model framework, we must leave the factorization as written above.

- The variational objective function then is computed,

$$\mathcal{L} = \sum_{i=1}^{D} \mathbb{E}\left[\ln \frac{p(x_i, z_i|\beta)}{q(z_i)}\right] + \mathbb{E}\left[\ln \frac{p(\beta)}{q(\beta)}\right] \tag{10}$$

- One typical way for optimizing this objective function is called "batch" variational inference. This algorithm is as follows:

Batch variational inference

```
1. For $i = 1, \ldots, D$, optimize $q(z_i)$
2. Optimize $q(\beta)$
3. Repeat
```

- The general coding structure therefore involves an inner loop to optimize the q distribution for each group of data. Then the q distribution for the global variables is optimized, usually very quickly using the statistics collected during the inner loop. This is done in an outer loop, which counts how many iterations we run.

- For LDA, since $q(z_i) = q(\theta_i) \prod_j q(c_{ij})$, the inner loop itself requires an algorithm where we iterate back and forth several times between updating $q(\theta_i)$ and updating each $q(c_{ij})$.

- This can lead to scalability issues. For example, what if D is massive? What if we have millions of documents we want to do topic modeling on? Even if learning each $q(z_i)$ is fast (e.g., a fraction of a second), that fraction multiplied several million times can lead to an algorithm that takes over 24 hours to run a single iteration!

- Looking at the structure of the model, we can obviously speed this up with parallelization. That is, in Step 1 of the algorithm, we can break the groups into subsets and send that information out to different processors. For example, if we have 100 computers, each computer can be responsible for optimizing 1% of the $q(z_i)$. This can make the algorithm run about 100 times faster.

- Another way to speed up variational inference in this context is to use techniques from optimization. Notice that $\mathcal{L}$ is an objective function like any other, just with a Bayesian interpretation of the result. Since $\mathcal{L}$ for mixed membership models such as LDA can be written as a sum over groups, we can use stochastic optimization to optimize it.

- We will talk about stochastic optimization next at a very high level. It's the type of technique that needs to be studies more fully in an optimization class, since there is a lot that is known about it. We will just state the procedure here and analyze it in the context of variational inference for CEF models.

- One thing to notice in the following is that it is not an either/or choice between stochastic inference and parallelization. The stochastic algorithm described below can be parallelized very easily to provide a massive speed-up of inference.

Stochastically optimizing $\mathcal{L}$

- Since the local variables factorize, we get a sum over groups in the variational objective. Again, this means we can generically write

$$\mathcal{L} = \sum_{i=1}^{D} \mathbb{E}\left[\ln \frac{p(x_i, z_i|\beta)}{q(z_i)}\right] + \mathbb{E}\left[\ln \frac{p(\beta)}{q(\beta)}\right] \tag{11}$$

- This type of objective function is perfect for "stochastic optimization." Operationally speaking, to stochastically optimize $\mathcal{L}$, we do the following at each iteration:

Stochastic variational inference

1. Randomly select a subset of local data, $S_t \subset \{1, \ldots, D\}$

2. Construct the scaled variational objective function

$$\mathcal{L}_t = \frac{D}{|S_t|} \sum_{i \in S_t} \mathbb{E}_q\left[\ln \frac{p(x_i, z_i|\beta)}{q(z_i)}\right] + \mathbb{E}_q\left[\ln \frac{p(\beta)}{q(\beta)}\right] \tag{12}$$

3. Optimize each $q(z_i)$ in $\mathcal{L}_t$ *only*

4. Update the parameters of $q(\beta)$ using a gradient step of the form

$$q(\beta|\psi) \quad \rightarrow \quad \psi_t = \psi_{t-1} + \rho_t M_t \nabla_\psi \mathcal{L}_t \tag{13}$$

5. Repeat

- Comments:

  1. We can show that $\mathbb{E}[\mathcal{L}_t] = \mathcal{L}$. This expectation is over the randomness of S_t. Therefore, we sum over all $\binom{D}{|S_t|}$ possible subsets S_t and multiply with the probability $\binom{D}{|S_t|}^{-1}$ since each subset is equally probable. This is significant since theory shows that we are therefore optimizing $\mathcal{L}$ with this method (subject to the next constraint).

  2. We must have $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$ for this method to provably converge to a local optimal solution of $\mathcal{L}$. Often, people will choose $\rho_t = \frac{1}{(t_0+t)^\kappa}$ with $\kappa \in (\frac{1}{2}, 1]$, since this can be shown to satisfy these requirements.

Example: Stochastic inference for LDA

- Before we look more in detail into $\psi_t = \psi_{t-1} + \rho_t M_t \nabla_\psi \mathcal{L}_t$ in general for CEF models, let's look at the final algorithm we will get for LDA.

- The local variables are:

  1. $q(\theta_d)$: Distribution on topics for document d

  2. $q(c_{dj})$: Word allocation for word j in document d

- The global variables are:

  1. $q(\beta_k)$: Topics (distributions on words) for $k = 1, \ldots, K$. We pick

$$q(\beta_k) = \text{Dirichlet}(\gamma_{k1}, \ldots, \gamma_{kV})$$

SVI for LDA

1. Pick a random subset of documents

2. Learn $q(c_{dj})$ an $q(\theta_d)$ for each by iterating between these q

3. Update γ_k as follows:

   (a) Define $\lambda_d^{(k)} = \sum_{j=1}^{n_d} \mathbb{E}_q[\mathbb{1}(c_{di} = k)] \mathrm{e}_{x_{di}}$ (e_i is vector of all zeros except 1 in dim i)

   (b) Set $\gamma_k^{(t)} = (1-\rho_t)\gamma_k^{(t-1)} + \rho_t \left(\gamma + \frac{D}{|S_t|} \sum_{d \in S_t} \lambda_d^{(k)}\right)$

- Notice that SVI for LDA takes the old parameters and averages them with the new parameters calculated only over the subset chosen for the current iteration. As the iterations increase, the new information is weighted less and less since ρ_t is decreasing to zero.

- Contrast this with the "batch" inference algorithm

  1. In Step 1, we instead picked all the documents

2. In Step 3b, we instead set $\gamma_k^{(t)} = \gamma + \sum_{d=1}^{D} \lambda_d^{(k)}$. Equivalently, we just set $\rho_t = 1$ because we fully weight all the new information.

- Let's return to the generic CEF hierarchy to see more of the difference between batch and stochastic variational inference.

Abstract stochastic variational inference

- We're in the general mixed membership setting described above. Also, let β be a natural parameter for notation convenience.
- Likelihood:

$$p(x, z|\beta) = \prod_{i=1}^{D} p(x_i, z_i|\beta) = \left[\prod_{i=1}^{D} h(x_i, z_i)\right] \mathrm{e}^{\beta^T \sum_{i=1}^{D} t(x_i,z_i) - DA(\beta)} \tag{14}$$

- Prior:

$$p(\beta) = f(\xi, \nu)\mathrm{e}^{\beta^T \xi - \nu A(\beta)} \tag{15}$$

- Approximate posterior chosen in same family as prior:

$$q(\beta) = f(\xi', \nu')\mathrm{e}^{\beta^T \xi' - \nu' A(\beta)} \tag{16}$$

- Next, we specifically work with the variational objective function. Before, we didn't bother with this because we had the "trick" which allowed us to go straight to the answer. In other words, we could find (ξ', ν') such that $\nabla\mathcal{L} = 0$ without explicitly calculating $\mathcal{L}$.
- Now, in the stochastic setting we want to set

$$\begin{bmatrix} \xi' \\ \nu' \end{bmatrix} \leftarrow \begin{bmatrix} \xi' \\ \nu' \end{bmatrix} + \rho_t M_t \nabla_{(\xi',\nu')}\mathcal{L} \tag{17}$$

Therefore, we need to explicitly calculate $\mathcal{L}$.

- If we only focus on the terms in $\mathcal{L}$ involving β, we can say that the full variational objective function is

$$\mathcal{L}_\beta = \sum_{i=1}^{D} \mathbb{E}[\ln p(x_i, z_i|\beta)] + \mathbb{E}[\ln p(\beta)] - \mathbb{E}[\ln q(\beta)] \tag{18}$$

- And plugging in the distributions above, we have

$$\mathcal{L}_\beta = \mathbb{E}_q[\beta]^T \left(\sum_{i=1}^{D} \mathbb{E}[t(x_i, z_i)] + \xi - \xi'\right) - \mathbb{E}_q[A(\beta)](D + \nu - \nu') + \ln f(\xi', \nu') + \text{const.} \tag{19}$$

The constant is with respect to ξ' and ν'.

- Next we need to calculate $\mathbb{E}_q[\beta]$ and $\mathbb{E}_q[A(\beta)]$ using the q distribution of β. We have already seen how we can do this with exponential family distributions. That is, if we solve the equalities

$$\int \nabla_{\xi'} q(\beta)d\beta = 0, \qquad \int \frac{\partial}{\partial \nu'} q(\beta)d\beta = 0 \tag{20}$$

we will find that

$$\mathbb{E}_q[\beta] = -\nabla_{\xi'} \ln f(\xi', \nu'), \qquad \mathbb{E}_q[A(\beta)] = \frac{\partial \ln f(\xi', \nu')}{\partial \nu'} \tag{21}$$

- Therefore,

$$\begin{aligned} \mathcal{L}_\beta &= -\left[\nabla_{\xi'} \ln f(\xi', \nu')\right]^T \left(\sum_{i=1}^D \mathbb{E}[t(x_i, z_i)] + \xi - \xi'\right) - \frac{\partial \ln f(\xi', \nu')}{\partial \nu'}(D + \nu - \nu') \\ &\quad - \ln f(\xi', \nu') \; + \; \text{const.} \end{aligned} \tag{22}$$

- Recall that we want to set

$$\nabla_{(\xi', \nu')} \mathcal{L}_\beta = \left[\begin{array}{c} \nabla_{\xi'} \mathcal{L}_\beta \\ \frac{\partial}{\partial \nu'} \mathcal{L}_\beta \end{array}\right] = 0$$

- Using the equation for $\mathcal{L}_\beta$ above, we can calculate that

$$\begin{aligned} \nabla_{\xi'} \mathcal{L}_\beta &= -\nabla^2_{\xi'} \ln f(\xi', \nu') \left(\sum_{i=1}^D \mathbb{E}[t(x_i, z_i)] + \xi - \xi'\right) - \nabla_{\xi'} \ln f(\xi', \nu') \\ &\quad \frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu' \partial \xi}(D + \nu - \nu') + \nabla_{\xi'} \ln f(\xi', \nu') \end{aligned} \tag{23}$$

$$\begin{aligned} \frac{\partial}{\partial \nu'} \mathcal{L}_\beta &= -\frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu' \partial \xi'^T} \left(\sum_{i=1}^D \mathbb{E}[t(x_i, z_i)] + \xi - \xi'\right) - \frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu'^2}(D + \nu - \nu') \\ &\quad + \frac{\partial \ln f(\xi', \nu')}{\partial \nu'} - \frac{\partial \ln f(\xi', \nu')}{\partial \nu'} \end{aligned} \tag{24}$$

- Notice that in both equations, there are two terms that cancel out. As written, this is a two-equations two-unknowns situation. However, it makes the problem much clearer to write it as follows

$$\nabla_{(\xi', \nu')} \mathcal{L}_\beta = \left[\begin{array}{c} \nabla_{\xi'} \mathcal{L}_\beta \\ \frac{\partial}{\partial \nu'} \mathcal{L}_\beta \end{array}\right] = -\left[\begin{array}{cc} \nabla^2_{\xi'} \ln f(\xi', \nu') & \frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu' \partial \xi} \\ \frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu' \partial \xi'^T} & \frac{\partial^2 \ln f(\xi', \nu')}{\partial \nu'^2} \end{array}\right] \left[\begin{array}{c} \sum_{i=1}^D \mathbb{E}[t(x_i, z_i)] + \xi - \xi' \\ D + \nu - \nu' \end{array}\right]$$

- Again, all we've done is write the previous two equations in matrix vector form.

- The goal is to set the resulting vector to zero. However, notice that, since the preconditioning matrix is negative definite, the only way we can make this matrix-vector product equal zero is by setting the right vector to zero. This means we want to find values of ξ' and ν' such that the right vector is zero. We can simply read this from the vector:

$$\xi' = \xi + \sum_{i=1}^{D} \mathbb{E}[t(x_i, z_i)], \qquad \nu' = \nu + D \tag{25}$$

- This is exactly the update for these parameters that we derived before by following the "log–expectation–exponential–normalization" rule. It is good to see via another route that this is the correct updates for these parameters. However, the reason we are calculating these things now is to see how stochastic variational inference modifies this.

- Recall that for a parameter to a variational q distribution, ψ, we want to set $\psi_t = \psi_{t-1} + \rho_t M_t \nabla_\psi \mathcal{L}_t$, where $\mathcal{L}_t$ is calculated using the sub-sampled set of groups. Using the above notation, that means

$$\begin{bmatrix} \xi'_t \\ \nu'_t \end{bmatrix} = \begin{bmatrix} \xi'_{t-1} \\ \nu'_{t-1} \end{bmatrix} - \rho_t M_t \begin{bmatrix} \nabla^2_{\xi'} \ln f(\xi', \nu') & \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu' \partial \xi} \\ \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu' \partial \xi'^T} & \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu'^2} \end{bmatrix} \begin{bmatrix} \frac{D}{|S_t|} \sum_{i \in S_t} \mathbb{E}[t(x_i, z_i)] + \xi - \xi'_{t-1} \\ D + \nu - \nu'_{t-1} \end{bmatrix} \tag{26}$$

- We could simply leave it here and pick $M_t = I$. However, if we pick M_t in a clever way, we can get a very "clean" update. We need to pick M_t to be some positive definite matrix. For example, steepest ascent uses $M_t = I$. Newton's method sets $M_t = -(\nabla^2 \mathcal{L}_t)^{-1}$. For stochastic variational inference, we set

$$M_t = - \begin{bmatrix} \nabla^2_{\xi'} \ln f(\xi', \nu') & \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu' \partial \xi} \\ \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu' \partial \xi'^T} & \frac{\partial^2 \ln f(\xi',\nu')}{\partial \nu'^2} \end{bmatrix}^{-1} \tag{27}$$

- This preconditioning matrix can be shown to be equivalent to $M_t = \mathbb{E}[\nabla^2 \ln q(\beta)]$. This gradient is known as the *natural gradient*. This is a "good" gradient direction for reasons that have been analyzed (and are beyond the scope of this class). Therefore, we pick this M_t not only for convenience.

- When we use this value of M_t, we see that the update for ν'_t is always to set it equal to $D + \nu$. For ξ'_t it is

$$\xi'_t = (1 - \rho_t)\xi'_{t-1} + \rho_t \Big(\xi + \frac{D}{|S_t|} \sum_{i \in S_t} \mathbb{E}[t(x_i, z_i)] \Big) \tag{28}$$

- We see that the update is a weighted average of the new sufficient statistics with the old value. Compare this with the update for stochastic LDA above to see that this result generalizes the pattern we observed there.

- Also notice that the function of the scaling parameter is to treat each group of data as though it appears $D/|S_t|$ times in the data set. This is because we only look at $|S_t|/D$ fraction of data—if we look at 1% of the data, we treat each observation (group) as if it appeared 100 times to keep the size of the data set constant.