CSci 8980: Advanced Topics in Graphical Models

Expectation Propagation

Instructor: Arindam Banerjee

October 26, 2007
Consider a Bayesian model

Posterior Estimation

Latent variable $u$ with prior $P(u)$,

Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest
- Posterior over latent variable $P(u|D)$
- Likelihood of observation $P(D)$

For conjugate priors, posterior is in the same family

In general, it can be intractable

What is the best approximation in the (prior) family?
Consider a Bayesian model

- Latent variable $\mathbf{u}$ with prior $P^{(0)}(\mathbf{u})$
Consider a Bayesian model
- Latent variable $\mathbf{u}$ with prior $P^{(0)}(\mathbf{u})$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$
Consider a Bayesian model
- Latent variable $u$ with prior $P^{(0)}(u)$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest
Consider a Bayesian model
- Latent variable $u$ with prior $P^{(0)}(u)$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest
- Posterior over latent variable $P^{(0)}(u|D)$
Consider a Bayesian model

- Latent variable $\mathbf{u}$ with prior $P^{(0)}(\mathbf{u})$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest

- Posterior over latent variable $P^{(0)}(\mathbf{u}|D)$
- Likelihood of observation $P(D)$
Consider a Bayesian model

- Latent variable $u$ with prior $P^{(0)}(u)$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest

- Posterior over latent variable $P^{(0)}(u|D)$
- Likelihood of observation $P(D)$

For conjugate priors, posterior is in the same family
Consider a Bayesian model

- Latent variable $u$ with prior $P(0)(u)$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest

- Posterior over latent variable $P(0)(u|D)$
- Likelihood of observation $P(D)$

For conjugate priors, posterior is in the same family

In general, it can be intractable
Consider a Bayesian model

- Latent variable $\mathbf{u}$ with prior $P^{(0)}(\mathbf{u})$
- Observable $D$, such as $\{x_1, \ldots, x_m\}$

Quantities of interest

- Posterior over latent variable $P^{(0)}(\mathbf{u}|D)$
- Likelihood of observation $P(D)$

For conjugate priors, posterior is in the same family

In general, it can be intractable

What is the best approximation in the (prior) family?
The likelihood function often factorizes

\[ P(D|u) = \prod_{i=1}^{n} t_i(u) \]
Posterior Estimation (Contd.)

- The likelihood function often factorizes
  \[ P(D|u) = \prod_{i=1}^{n} t_i(u) \]

- The true posterior may be intractable
  \[ P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u) \]
The likelihood function often factorizes

\[ P(D|u) = \prod_{i=1}^{n} t_i(u) \]

The true posterior may be intractable

\[ P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u) \]

The normalizer \( Z \) is the same as the data likelihood, i.e.,

\[ Z = \int_u P^{(0)}(u) \prod_{i=1}^{n} t_i(u) du = \int_u P(D|u)P^{(0)}(u) du = P(D) \]
The likelihood function often factorizes

\[
P(D|u) = \prod_{i=1}^{n} t_i(u)
\]

The true posterior may be intractable

\[
P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u)
\]

The normalizer \(Z\) is the same as the data likelihood, i.e.,

\[
Z = \int_{u} P^{(0)}(u) \prod_{i=1}^{n} t_i(u) du = \int_{u} P(D|u)P^{(0)}(u) du = P(D)
\]

The two problems are closely related.
Approximating the Posterior

- Assume prior $P^{(0)}(\mathbf{u})$ belongs to exponential family $\mathcal{F}$

$$P^{(0)}(\mathbf{u}) = \exp(\langle \theta_0, s(\mathbf{u}) \rangle - \psi(\theta))$$
Approximating the Posterior

- Assume prior $P^{(0)}(u)$ belongs to exponential family $\mathcal{F}$

$$P^{(0)}(u) = \exp(\langle \theta_0, s(u) \rangle - \psi(\theta))$$

- Let $Q(u) \in \mathcal{F}$ be the best approximation to $P(u|D)$
Approximating the Posterior

- Assume prior $P^{(0)}(u)$ belongs to exponential family $\mathcal{F}$

  $$P^{(0)}(u) = \exp(\langle \theta_0, s(u) \rangle - \psi(\theta))$$

- Let $Q(u) \in \mathcal{F}$ be the best approximation to $P(u|D)$
- Tractably compute $Q(u)$ when $P(u|D)$ is hard to compute
Approximating the Posterior

- Assume prior $P^{(0)}(u)$ belongs to exponential family $\mathcal{F}$
  
  $$P^{(0)}(u) = \exp(\langle \theta_0, s(u) \rangle - \psi(\theta))$$

- Let $Q(u) \in \mathcal{F}$ be the best approximation to $P(u|D)$
- Tractably compute $Q(u)$ when $P(u|D)$ is hard to compute
  - Approach 1: Assumed density filtering, online Bayesian learning
Approximating the Posterior

- Assume prior $P^{(0)}(\mathbf{u})$ belongs to exponential family $\mathcal{F}$

  $$P^{(0)}(\mathbf{u}) = \exp(\langle \theta_0, s(\mathbf{u}) \rangle - \psi(\theta))$$

- Let $Q(\mathbf{u}) \in \mathcal{F}$ be the best approximation to $P(\mathbf{u}|D)$
- Tractably compute $Q(\mathbf{u})$ when $P(\mathbf{u}|D)$ is hard to compute
  - Approach 1: Assumed density filtering, online Bayesian learning
  - Approach 2: Expectation propagation
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
- Recall that

$$P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u)$$
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
- Recall that
  \[
P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u)
  \]
- At each step, update $Q$ to incorporate one $t_i(u)$
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
- Recall that
  \[ P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u) \]
- At each step, update $Q$ to incorporate one $t_i(u)$
  - Compute the true Bayesian update
    \[ \hat{P}(u) = \frac{t_i(u)Q(u)}{\int_{z} t_i(z)Q(z)dz} \]
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
- Recall that

\[ P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u) \]

- At each step, update $Q$ to incorporate one $t_i(u)$
  - Compute the true Bayesian update

\[ \hat{P}(u) = \frac{t_i(u)Q(u)}{\int_z t_i(z)Q(z)dz} \]

- Find $Q^{new} \in \mathcal{F}$ such that

\[ Q^{new}(u) = \arg\min_{\tilde{Q} \in \mathcal{F}} KL(\hat{P}(u)\|\tilde{Q}(u)) \]
Assumed Density Filtering

- Start with an initial guess $Q(u) = P^{(0)}(u)$
- Recall that
  \[ P(u|D) \propto P^{(0)}(u) \prod_{i=1}^{n} t_i(u) \]
- At each step, update $Q$ to incorporate one $t_i(u)$
  - Compute the true Bayesian update
    \[ \hat{P}(u) = \frac{t_i(u)Q(u)}{\int_z t_i(z)Q(z)dz} \]
  - Find $Q^{new} \in \mathcal{F}$ such that
    \[ Q^{new}(u) = \arg\min_{\tilde{Q} \in \mathcal{F}} KL(\hat{P}(u)\|\tilde{Q}(u)) \]
- Maximum likelihood estimate with $\hat{P}$ as the true distribution
To obtain $Q^{new}$ it is sufficient to do moment matching

$$\mu_{new} = E_{\hat{P}}[s(u)]$$
To obtain $Q^{new}$ it is sufficient to do moment matching

$$\mu_{new} = E_{\hat{P}}[s(u)]$$

For each factor $t_i(u)$
To obtain $Q^{new}$ it is sufficient to do moment matching

$$\mu_{new} = E_{\hat{P}}[s(u)]$$

For each factor $t_i(u)$

- Compute the means (moments) of $\hat{P}(u) \propto t_i(u)Q(u)$
To obtain $Q^{new}$ it is sufficient to do moment matching

$$\mu_{new} = E_{\hat{P}}[s(u)]$$

For each factor $t_i(u)$

- Compute the means (moments) of $\hat{P}(u) \propto t_i(u)Q(u)$
- Pick $Q^{new} \in \mathcal{F}$ with these mean parameters
For a single factor $t_i(u)$
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
  - The approximate posterior $Q^{new}(u) \propto \tilde{t}_i(u)Q(u)$
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
  - The approximate posterior $Q^{new}(u) \propto \tilde{t}_i(u)Q(u)$
  - The factor $\tilde{t}_i(u) \propto Q^{new}(u)/Q(u)$
ADF: An Alternative Viewpoint

- For a single factor \( t_i(\mathbf{u}) \)
  - The true posterior: \( \hat{P}(\mathbf{u}) \propto t_i(\mathbf{u})Q(\mathbf{u}) \)
  - The approximate posterior: \( Q^{\text{new}}(\mathbf{u}) \propto \tilde{t}_i(\mathbf{u})Q(\mathbf{u}) \)
  - The factor: \( \tilde{t}_i(\mathbf{u}) \propto Q^{\text{new}}(\mathbf{u})/Q(\mathbf{u}) \)

- In general, after a pass through all factors

\[
Q(\mathbf{u}) \propto P^{(0)}(\mathbf{u}) \prod_{i=1}^{n} \tilde{t}_i(\mathbf{u})
\]
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
  - The approximate posterior $Q^{\text{new}}(u) \propto \tilde{t}_i(u)Q(u)$
  - The factor $\tilde{t}_i(u) \propto Q^{\text{new}}(u)/Q(u)$

- In general, after a pass through all factors

  $$Q(u) \propto P^{(0)}(u) \prod_{i=1}^{n} \tilde{t}_i(u)$$

- Algo: Set $\tilde{t}_i = 1, \forall i$. For each factor $t_i(u)$
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
  - The approximate posterior $Q^{\text{new}}(u) \propto \tilde{t}_i(u)Q(u)$
  - The factor $\tilde{t}_i(u) \propto Q^{\text{new}}(u)/Q(u)$

- In general, after a pass through all factors

$$Q(u) \propto P^{(0)}(u) \prod_{i=1}^{n} \tilde{t}_i(u)$$

- Algo: Set $\tilde{t}_i = 1$, $\forall i$. For each factor $t_i(u)$
  - Compute $Q^{\text{new}}$ with $\mu^{\text{new}} = E_{\hat{P}}[s(u)]$
ADF: An Alternative Viewpoint

- For a single factor $t_i(u)$
  - The true posterior $\hat{P}(u) \propto t_i(u)Q(u)$
  - The approximate posterior $Q^{new}(u) \propto \tilde{t}_i(u)Q(u)$
  - The factor $\tilde{t}_i(u) \propto Q^{new}(u)/Q(u)$

- In general, after a pass through all factors

$$Q(u) \propto P^{(0)}(u) \prod_{i=1}^{n} \tilde{t}_i(u)$$

- Algo: Set $\tilde{t}_i = 1, \forall i$. For each factor $t_i(u)$
  - Compute $Q^{new}$ with $\mu^{new} = E_{\hat{P}}[s(u)]$
  - Set $\tilde{t}_i^{new}(u) \propto Q^{new}(u)/Q(u)$
Issues with ADF

- ADF makes one pass through the data
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once
- Issues with ADF
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once
- Issues with ADF
  - Earlier approximations $\tilde{t}_i(u)$ may be poor
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once

- Issues with ADF
  - Earlier approximations $\tilde{t}_i(u)$ may be poor
  - No way of going back and fixing the earlier approximations
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once
- Issues with ADF
  - Earlier approximations $\tilde{t}_i(u)$ may be poor
  - No way of going back and fixing the earlier approximations
  - Depends on the order in which data is processed
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once

- Issues with ADF
  - Earlier approximations $\tilde{t}_i(u)$ may be poor
  - No way of going back and fixing the earlier approximations
  - Depends on the order in which data is processed

- In principle, $\tilde{t}_i$ can be updated multiple times
Issues with ADF

- ADF makes one pass through the data
  - $Q$ is updated once for each factor
  - Equivalently, $\tilde{t}_i(u)$ is updated once

- Issues with ADF
  - Earlier approximations $\tilde{t}_i(u)$ may be poor
  - No way of going back and fixing the earlier approximations
  - Depends on the order in which data is processed

- In principle, $\tilde{t}_i$ can be updated multiple times

- EP effectively extends ADF allowing multiple passes
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^n \tilde{t}_i(u)}{\int_z \prod_{i=1}^n \tilde{t}_i(z) dz}$$
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior
  \[
  Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_{z} \prod_{i=1}^{n} \tilde{t}_i(z) dz}
  \]
- Until all $\tilde{t}_i$ converge
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_{z} \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_z \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
  - Remove $\tilde{t}_i(u)$ from $Q(u)$ to get ‘old’ posterior

$$Q^i(u) \propto Q(u)/\tilde{t}_i(u)$$
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_z \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
  - Remove $\tilde{t}_i(u)$ from $Q(u)$ to get ‘old’ posterior

$$Q^i(u) \propto Q(u)/\tilde{t}_i(u)$$

- Construct $\hat{P} \propto t_i(u)Q^i(u)$
**Expectation Propagation**

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_z \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
  - Remove $\tilde{t}_i(u)$ from $Q(u)$ to get ‘old’ posterior
    $$Q^i(u) \propto Q(u)/\tilde{t}_i(u)$$

- Construct $\hat{P} \propto t_i(u)Q^i(u)$
- Get $Q^{new}$ with $\mu^{new} = E_P[s(u)]$ and normalizer $Z_i$
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_z \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
  - Remove $\tilde{t}_i(u)$ from $Q(u)$ to get ‘old’ posterior

$$Q^i(u) \propto Q(u)/\tilde{t}_i(u)$$

- Construct $\hat{P} \propto t_i(u)Q^i(u)$
- Get $Q^{new}$ with $\mu^{new} = E_{\hat{P}}[s(u)]$ and normalizer $Z_i$
- Update $\tilde{t}_i(u) = Z_i Q^{new}(u)/Q^i(u)$
Expectation Propagation

- Initialize the term approximations $\tilde{t}_i(u)$
- Compute the posterior

$$Q(u) = \frac{\prod_{i=1}^{n} \tilde{t}_i(u)}{\int_{z} \prod_{i=1}^{n} \tilde{t}_i(z) dz}$$

- Until all $\tilde{t}_i$ converge
  - Choose a $\tilde{t}_i(u)$ to refine
  - Remove $\tilde{t}_i(u)$ from $Q(u)$ to get ‘old’ posterior
    $$Q^i(u) \propto Q(u)/\tilde{t}_i(u)$$

  - Construct $\hat{P} \propto t_i(u) Q^i(u)$
  - Get $Q^{new}$ with $\mu^{new} = E_{\hat{P}}[s(u)]$ and normalizer $Z_i$
  - Update $\tilde{t}_i(u) = Z_i Q^{new}(u)/Q^i(u)$

- Estimate the data likelihood as

$$P(D) \approx \int_{z} \prod_{i=1}^{n} \tilde{t}_i(z) dz$$
Experiments

- The clutter problem:

\[
\begin{align*}
p(u) &= \mathcal{N}(u; 0, 100\mathbb{I}) \\
p(x|u) &= (1 - w)\mathcal{N}(x; u, \mathbb{I}) + w\mathcal{N}(x; 0, 100\mathbb{I})
\end{align*}
\]

Experiments

- The clutter problem:

\[
p(u) = \mathcal{N}(u; 0, 100I) \\
p(x|u) = (1 - w)\mathcal{N}(x; u, I) + w\mathcal{N}(x; 0, 100I)
\]

- For a set of observations \(D = \{x_1, \ldots, x_n\}\)

\[
p(u, D) = p(u) \prod_{j=1}^{n} p(x_j|u)
\]
Experiments

- The clutter problem:

\[
p(u) = \mathcal{N}(u; 0, 100I)
\]
\[
p(x|u) = (1 - w)\mathcal{N}(x; u, I) + w\mathcal{N}(x; 0, 100I)
\]

- For a set of observations \( D = \{x_1, \ldots, x_n\} \)

\[
p(u, D) = p(u) \prod_{j=1}^{n} p(x_j|u)
\]

- Evaluation
The clutter problem:

\[
p(u) = \mathcal{N}(u; 0, 100I)
\]

\[
p(x|u) = (1 - w)\mathcal{N}(x; u, I) + w\mathcal{N}(x; 0, 100I)
\]

For a set of observations \( D = \{x_1, \ldots, x_n\} \)

\[
p(u, D) = p(u) \prod_{j=1}^{n} p(x_j|u)
\]

Evaluation

- Evidence/likelihood \( p(D) = \int_u p(u, D)du \)
Experiments

- The clutter problem:

\[
p(u) = \mathcal{N}(u; 0, 100\mathbb{I})
\]

\[
p(x|u) = (1 - w)\mathcal{N}(x; u, \mathbb{I}) + w\mathcal{N}(x; 0, 100\mathbb{I})
\]

- For a set of observations \( D = \{x_1, \ldots, x_n\} \)

\[
p(u, D) = p(u) \prod_{j=1}^{n} p(x_j|u)
\]

- Evaluation
  - Evidence/likelihood \( p(D) = \int_u p(u, D) du \)
  - Posterior mean \( E[u|D] = \int_u u p(u|D) du \)
Results: Likelihood $P(D)$

$n = 20$

$n = 200$
Results: Posterior Mean $E[u|D]$

$n = 20$

$n = 200$
Results: Complex Posterior

![Graph showing posterior distributions]

- **Exact**
- **EP**
- **VB**
- **Laplace**

![Graph showing FLOPS vs Error]

- **Laplace**
- **VB**
- **EP**
- **Gibbs**
- **Importance**