Expectation Propagation

Thomas P Minka

Kuhwan Jeong [1]

[1]Department of Statistics, Seoul National University, South Korea

December, 2016

Example 1

- Given a parameter θ, observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently sampled from

$$p(\mathbf{x}|\theta) = (1-w)\mathcal{N}(\mathbf{x}; \theta, \mathbf{I}) + w\mathcal{N}(\mathbf{x}; \mathbf{0}, 10\mathbf{I}).$$

- We use a Gaussian prior for θ :

$$\theta \sim \mathcal{N}(\theta; \mathbf{0}, 100\mathbf{I}_d).$$

- The joint distribution of θ and $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is

$$p(D, \theta) = p(\theta) \prod_i p(\mathbf{x}_i|\theta).$$

- We want to know $p(\theta|D)$.

Assumed-density filtering

$$
\begin{aligned}
p(\theta|D) &\propto p(\theta)p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)\cdots p(x_n|\theta) \\
&\propto p(\theta|x_1)p(x_2|\theta)p(x_3|\theta)\cdots p(x_n|\theta) \\
&\propto p(\theta|x_1,x_2)p(x_3|\theta)\cdots p(x_n|\theta) \\
&\vdots \\
&\propto p(\theta|x_1,x_2,\ldots,x_{n-1})p(x_n|\theta)
\end{aligned}
$$

$$
\begin{aligned}
p(\theta|D) &\approx q_1(\theta)p(x_2|\theta)p(x_3|\theta)\cdots p(x_n|\theta) \\
&\approx q_2(\theta)p(x_3|\theta)\cdots p(x_n|\theta) \\
&\vdots \\
&\approx q_{n-1}(\theta)p(x_n|\theta)
\end{aligned}
$$

Assumed-density filtering

1. Initialize with $q(\theta) = p(\theta)$.

2. For $i = 1, \ldots, n$,
   1. Define $\hat{q}(\theta) = \frac{p(\mathbf{x}_i|\theta)q(\theta)}{\int p(\mathbf{x}_i|\theta)q(\theta)d\theta}$.
   2. Find $q^{new}(\theta) \in \mathcal{H}$ minimizing the KL-divergence

$$D(\hat{q}(\theta)||q^{new}(\theta)).$$

- Here $\mathcal{H}$ is a class of distributions.
- After the i^{th} iteration, $q^{new}(\theta)$ approximates the posterior distribution given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$ with a prior $p(\theta)$.

Assumed-density filtering

- In the example, let

$$q(\theta) = \mathcal{N}(\theta; \mathbf{m}_\theta, v_\theta \mathbf{I}),$$
$$q^{new}(\theta) = \mathcal{N}(\theta; \mathbf{m}_\theta^{new}, v_\theta^{new} \mathbf{I}).$$

- Zeroing the gradient of $D(\hat{p}(\theta)||q^{new}(\theta))$ gives the conditions

$$\mathbf{m}_\theta^{new} = \int \hat{p}(\theta)\theta d\theta,$$
$$v_\theta^{new} d + (\mathbf{m}_\theta^{new})^T(\mathbf{m}_\theta^{new}) = \int \hat{p}(\theta)\theta^T\theta d\theta,$$

  or in other words, expectation constraints:

$$\mathbb{E}_{q^{new}}[\theta] = \mathbb{E}_{\hat{p}}[\theta],$$
$$\mathbb{E}_{q^{new}}[\theta^T\theta] = \mathbb{E}_{\hat{p}}[\theta^T\theta].$$

Assumed-density filtering

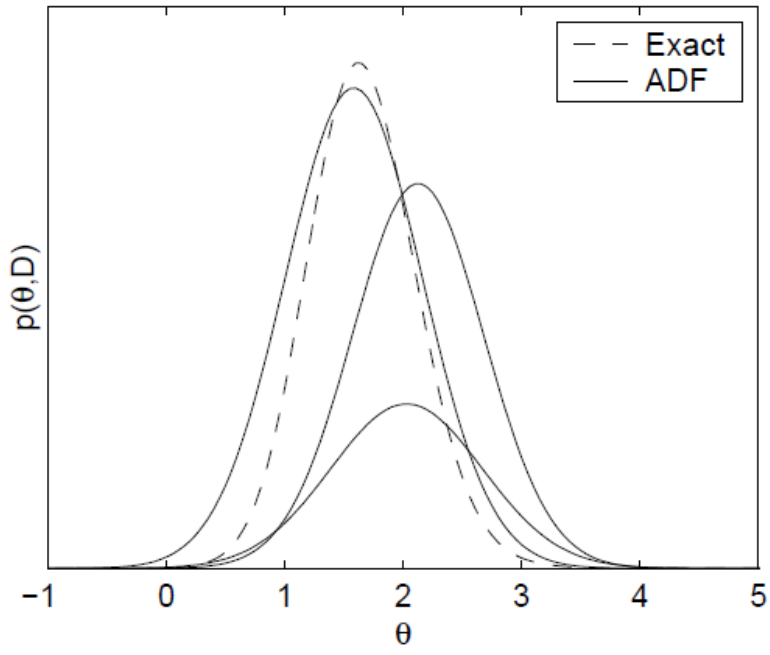
1. Initialize $\mathbf{m}_\theta = \mathbf{0}, v_\theta = 100$ (the prior).
2. For $i = 1, \ldots, n$,

$$r_i = \frac{(1-w)\mathcal{N}(\mathbf{x}_i; \mathbf{m}_\theta, (v_\theta+1)\mathbf{I})}{(1-w)\mathcal{N}(\mathbf{x}_i; \mathbf{m}_\theta, (v_\theta+1)\mathbf{I}) + w\mathcal{N}(\mathbf{x}_i; \mathbf{0}, 10\mathbf{I})},$$

$$\mathbf{m}_\theta^{new} = \mathbf{m}_\theta + v_\theta r_i \frac{\mathbf{x}_i - \mathbf{m}_\theta}{v_\theta + 1},$$

$$v_\theta^{new} = v_\theta - r_i \frac{v_\theta^2}{v_\theta + 1} + r_i(1-r_i)\frac{v_\theta^2 (\mathbf{x}_i - \mathbf{m}_\theta)^T(\mathbf{x}_i - \mathbf{m}_\theta)}{d(v_\theta+1)^2}.$$

Assumed-density filtering



- ADF depends on the order in which data is processed.
- No theory is available for how ADF varies with ordering.
- The error increases whenever similar data points are processed together.
- Processing the data in sorted order is especially bad.

Assumed-density filtering

1. Initialize with $q(\theta) = p(\theta)$.

2. For $i = 1, \ldots, n$,
   1. Define $\hat{q}(\theta) = \frac{p(\mathbf{x}_i|\theta)q(\theta)}{\int p(\mathbf{x}_i|\theta)q(\theta)d\theta}$.
   2. Find $q^{new}(\theta) \in \mathcal{H}$ minimizing the KL-divergence
   $$D(\hat{q}(\theta)||q^{new}(\theta)).$$
   3. Define
   $$\tilde{p}(\mathbf{x}_i|\theta) \propto \frac{q^{new}(\theta)}{q(\theta)}.$$

3. Then
$$q(\theta) \propto p(\theta) \prod_i \tilde{p}(\mathbf{x}_i|\theta).$$

Expectation Propagation

1. Initialize with $q(\theta) = p(\theta)$ and $\tilde{p}(\mathbf{x}_i|\theta) = 1$.
2. Until all $\tilde{p}(\mathbf{x}_i|\theta)$ converge :
   1. Choose a $\tilde{p}(\mathbf{x}_i|\theta)$ to refine.
   2. Define
   $$q^{-i}(\theta) \propto q(\theta)/\tilde{p}(\mathbf{x}_i|\theta)$$
   and
   $$\hat{q}(\theta) = \frac{p(\mathbf{x}_i|\theta)q^{-i}(\theta)}{\int p(\mathbf{x}_i|\theta)q^{-i}(\theta)d\theta}.$$
   3. Find $q^{new}(\theta) \in \mathcal{H}$ minimizing the KL-divergence
   $$D(\hat{q}(\theta)||q^{new}(\theta)).$$
   4. Set
   $$\tilde{p}(\mathbf{x}_i|\theta) \propto \frac{q^{new}(\theta)}{q^{-i}(\theta)}.$$

- Here $q^{-i}(\theta)$ approximates the posterior distribution given data $D \setminus \{\mathbf{x}_i\}$ with a prior $p(\theta)$.

Example 2

- There are n documents.
- There are W distinct words $(1, \ldots, W)$.
- n_{dw} is the number of word w in a document d.
- There are K topics $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K$ where $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kW})$.
- $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})$ is a topic proportion of a document d.

Example 2

- The probability of a document d is

$$p(d|\boldsymbol{\alpha},\boldsymbol{\phi}) = \int \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{w=1}^{W} \Big(\sum_k \theta_k \phi_{kw}\Big)^{n_{dw}} d\boldsymbol{\theta},$$
$$= \int \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{w=1}^{W} p(w|\boldsymbol{\theta})^{n_{dw}} d\boldsymbol{\theta}.$$

- EP approximates each term $p(w|\boldsymbol{\theta})$ by a simpler term

$$\tilde{p}(w|\boldsymbol{\theta}) = \prod_k \theta_k^{\beta_{dwk}},$$

giving

$$q_d(\boldsymbol{\theta}) = \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{w=1}^{W} \tilde{p}(w|\boldsymbol{\theta})^{n_{dw}} = \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\gamma}_d)$$

where $\gamma_{dk} = \alpha_k + \sum_w n_{dw}\beta_{dwk}$.

Expectation Propagation

1. Initialize with $\boldsymbol{\gamma}_d = \boldsymbol{\alpha}$ and $\beta_{dwk} = 0$
   $(q_d(\boldsymbol{\theta}) = \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\alpha}), \tilde{p}(w|\boldsymbol{\theta}) = 1)$.

2. Until all $\tilde{p}(w|\boldsymbol{\theta})$ converge :
   1. Choose a $\tilde{p}(w|\boldsymbol{\theta})$ to refine.
   2. Define
      $$\gamma_{dk}^{-w} = \gamma_{dk} - n_{dw}\beta_{dwk}.$$
      and
      $$\hat{q}(\boldsymbol{\theta}) = \frac{\left(\sum_k \theta_k \phi_{kw}\right)^{n_{dw}} \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\gamma}_d^{-w})}{\int \left(\sum_k \theta_k \phi_{kw}\right)^{n_{dw}} \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\gamma}_d^{-w}) d\boldsymbol{\theta}}.$$
   3. Find $q_d^{new}(\boldsymbol{\theta}) = \mathcal{D}(\boldsymbol{\theta}|\boldsymbol{\gamma}_d)$ by matching the mean and variance of $q_d^{new}(\boldsymbol{\theta})$ against those of $\hat{q}(\boldsymbol{\theta})$.
   4. Set
      $$\beta_{dwk} = \frac{\gamma_{dk} - \gamma_{dk}^{-w}}{n_{dw}}.$$

Estimation of ϕ

- Maximize the following lower bound to the log-likelihood:

$$\sum_{d=1}^{n} \log p(d|\alpha,\phi) \geq \sum_{d} \int q_d(\theta) \log \left\{ \mathcal{D}(\theta|\alpha) \prod_{w} \left(\sum_{k} \theta_k \phi_{kw} \right)^{n_{dw}} \right\} d\theta$$

$$- \sum_{d} \int q_d(\theta) \log q_d(\theta) d\theta$$

$$= \sum_{dw} n_{dw} \int q_d(\theta) \log \left(\sum_{k} \theta_k \phi_{kw} \right) d\theta + \text{const.}$$

- By zeroing the derivative with respect to ϕ_{kw}, we obtain

$$\phi_{kw}^{new} \propto \sum_{d} n_{dw} \int q_d(\theta) \frac{\theta_k \phi_{kw}}{\sum_k \theta_k \phi_{kw}} d\theta.$$

Experimental Results

- The models are compared quantitatively using test perplexity

$$\exp\left(-\frac{\sum_d \log p(d)}{\sum_d n_d}\right).$$