Variational Algorithms for Approximate Bayesian Inference

PhD Thesis by Matthew J. Beal
2003

Presented by: Bernhard Nessler
Bayesian Occam’s razor

• Overfitting in traditional ML can be solved by cross-validation, etc.
• Bayesian Model Learning intrinsically avoids overfitting

\[ p(m | y) = \frac{p(m)p(y | m)}{p(y)} \]
post dist over parameter:
\[ p(\theta | y, m) = \frac{p(y | \theta, m)p(\theta | m)}{p(y | m)} \quad \text{given one model!} \]
predictive distribution:
\[ p(y' | y) = \sum_m \int d\theta \ p(y' | \theta, m, y)p(\theta | m, y)p(m | y) \]
marginal likelihood:
\[ p(y | m) = \int d\theta \ p(y | \theta, m)p(\theta | m) \]

Overcomplex models are implicitly penalized since they will give a lower marginal likelihood.
Model selection

- Structure learning
- Input dependence (input selection)
- Cardinality (e.g. number of mixture components)
- Dimensionality (real valued latent variables)
Variational approach for EM

- ML/MAP learning of parameters (point estimates)
- Lower bound improvements
- Works with or without mean-field approximation

\[
\mathcal{L}(\theta) = \sum_{i=1}^{n} \ln \left( \frac{\ln p(y_i | \theta)}{\ln q_{x_i}(x_i)} \right) \\
= \sum_{i=1}^{n} \ln \left( \frac{\ln p(y_i | \theta)}{\ln q_{x_i}(x_i)} \right) \\
\geq \sum_{i=1}^{n} \ln \left( \frac{\ln p(y_i | \theta)}{\ln q_{x_i}(x_i)} \right) \\
= \sum_{i=1}^{n} \ln q_{x_i}(x_i) \ln \left( \frac{\ln p(y_i | \theta)}{\ln q_{x_i}(x_i)} \right) \\
\equiv \mathcal{F}(q_{x_1}(x_1), \ldots, q_{x_n}(x_n), \theta)
\]
Expectation maximization

E step: \( q_{x_i}^{(t+1)} \leftarrow \arg \max_{q_{x_i}} \mathcal{F}(q_{x_i}(x), \theta^{(t)}), \ \forall \ i \in \{1, \ldots, n\} \),

M step: \( \theta^{(t+1)} \leftarrow \arg \max_\theta \mathcal{F}(q_{x_i}^{(t+1)}(x), \theta) \).

Exact solution for E-Step: 
\[
q_{x_i}^{(t+1)}(x_i) = p(x_i \mid y_i, \theta^{(t)}), \ \forall \ i
\]

\[
\mathcal{F}(q_{x_i}^{(t+1)}(x), \theta^{(t)}) = \sum_i \int dx_i q_{x_i}^{(t+1)}(x_i) \ln \frac{p(x_i, y_i \mid \theta^{(t)})}{q_{x_i}^{(t+1)}(x_i)}
\]

\[
= \sum_i \int dx_i p(x_i \mid y_i, \theta^{(t)}) \ln \frac{p(x_i, y_i \mid \theta^{(t)})}{p(x_i \mid y_i, \theta^{(t)})}
\]

\[
= \sum_i \int dx_i p(x_i \mid y_i, \theta^{(t)}) \ln \frac{p(y_i \mid \theta^{(t)}) p(x_i \mid y_i, \theta^{(t)})}{p(x_i \mid y_i, \theta^{(t)})}
\]

\[
= \sum_i \ln p(y_i \mid \theta^{(t)}) = \mathcal{L}(\theta^{(t)})
\]

M-step solution by setting differentives to zero

M step: \( \theta^{(t+1)} \leftarrow \arg \max_\theta \sum_i \int dx_i p(x_i \mid y_i, \theta^{(t)}) \ln p(x_i, y_i \mid \theta) \)
Figure 2.1: The variational interpretation of EM for maximum likelihood learning. In the E step the hidden variable variational posterior is set to the exact posterior $p(x \mid y, \theta^{(t)})$, making the bound tight. In the M step the parameters are set to maximise the lower bound $\mathcal{F}(q^{(t+1)}_x, \theta)$ while holding the distribution over hidden variables $q^{(t+1)}_x(x)$ fixed.
Constrained distribution for $q_x(x)$

- intractable posterior for $x \mid y, \text{parameters}$
- Choose parameterized of mean field approximation

\[
q_{x_i}(x_i) = q_{x_i}(x_i \mid \lambda_i) \quad q_{x_i}(x_i) = \prod_{j=1}^{x_i} q_{x_{ij}}(x_{ij})
\]

E-step: minimize following KL

\[
\int d x_i \ q_{x_i}(x_i) \ln \frac{q_{x_i}(x_i)}{p(x_i \mid y_i, \theta)} \equiv \text{KL } [q_{x_i}(x_i) \parallel p(x_i \mid y_i, \theta)]
\]

Solution with mean field approximation:

\[
q_{x_{ij}}(x_{ij}) = \frac{1}{Z_{ij}} \exp \left[ \int d x_i \ j \prod_{j'} q_{x_{ij'}}(x_{ij'}) \ln p(x_i, y_i \mid \theta) \right]
\]
Approximate lower bound

Figure 2.2: The variational interpretation of constrained EM for maximum likelihood learning. In the E step the hidden variable variational posterior is set to that which minimises \( \text{KL} \left[ q_x(x) \| p(x \mid y, \theta^{(t)}) \right] \), subject to \( q_x(x) \) lying in the family of constrained distributions. In the M step the parameters are set to maximise the lower bound \( \mathcal{F}(q_x^{(t+1)}, \theta) \) given the current distribution over hidden variables.
Model learning

- approximate $p(y|m)$ for model comparison
- Approximate the distribution $p(\theta|y,m)$ for prediction
- Parameters are equivalent to hidden variables
- caveat: correlations in posterior: $p(x,\theta|y,m)$

$$\ln p(y|m) = \ln \int d\theta \, dx \, p(x, y, \theta | m)$$
$$= \ln \int d\theta \, dx \, q(x, \theta) \frac{p(x, y, \theta | m)}{q(x, \theta)}$$
$$\geq \int d\theta \, dx \, q(x, \theta) \ln \frac{p(x, y, \theta | m)}{q(x, \theta)}$$

Approximation by constraining to factorized distributions:

$$q(x, \theta) \approx q_x(x)q_\theta(\theta)$$

$$\ln p(y|m) \geq \int d\theta \, dx \, q_x(x)q_\theta(\theta) \ln \frac{p(x, y, \theta | m)}{q_x(x)q_\theta(\theta)}$$
$$= \int d\theta \, q_\theta(\theta) \left[ \int dx \, q_x(x) \ln \frac{p(x, y | \theta, m)}{q_x(x)} + \ln \frac{p(\theta | m)}{q_\theta(\theta)} \right]$$
$$= \mathcal{F}_m(q_x(x), q_\theta(\theta))$$
$$= \mathcal{F}_m(q_{x_1}(x_1), \ldots, q_{x_n}(x_n), q_\theta(\theta)),$$
Variational Bayesian EM (VBEM)

- Iteratively maximize the functional $\mathcal{F}_m$ with respect to the distributions $q_{x_i}(x)$ and $q_\theta(\theta)$.
- Coordinate ascent in function space of variational distributions
- Solutions from variational calculus using Lagrange multipliers

\[
\text{VBE step: } q_{x_i}^{(t+1)}(x_i) = \frac{1}{Z_{x_i}} \exp \left[ \int d\theta \, q_\theta^{(t)}(\theta) \ln p(x_i, y_i | \theta, m) \right] \quad \forall \ i
\]

\[
\text{VBM step: } q_\theta^{(t+1)}(\theta) = \frac{1}{Z_\theta} p(\theta | m) \exp \left[ \int dx \, q_{x}^{(t+1)}(x) \ln p(x, y | \theta, m) \right]
\]

\[
\ln p(y | m) - \mathcal{F}_m(q_x(x), q_\theta(\theta)) = \int d\theta \, dx \, q_x(x) \, q_\theta(\theta) \ln \frac{q_x(x) \, q_\theta(\theta)}{p(x, \theta | y, m)} = \text{KL} [q_x(x) \, q_\theta(\theta) \| p(x, \theta | y, m)] \geq 0 .
\]
log marginal likelihood

Figure 2.3: The variational Bayesian EM (VBEM) algorithm. In the VBE step, the variational posterior over hidden variables \( q_x(x) \) is set according to (2.60). In the VBM step, the variational posterior over parameters is set according to (2.56). Each step is guaranteed to increase (or leave unchanged) the lower bound on the marginal likelihood. (Note that the exact log marginal likelihood is a fixed quantity, and does not change with VBE or VBM steps — it is only the lower bound which increases.)
Posterior factorisation over hiddens and parameters

\[ q(x, \theta) \approx q_x(x)q_\theta(\theta) \]

In the VBEM framework, any factorisation works. But the quality of the lower bound depends on the quality of the factorisation. The goal is always to remove arcs, in order to achieve e.g. tree structures or the like (=structured approximation).
Model selection

- Comparing lower bounds ?? Bounds can be differently tight.
- KL is sum of KL’s for every component
- Increase of components (=complexity) increases KL

\[
\ln \frac{p(m \mid y)}{p(m' \mid y)} = + \ln p(m) + p(y \mid m) - \ln p(m') - \ln p(y \mid m') \\
= + \ln p(m) + \mathcal{F}(q_{x,\theta}) + \text{KL} [q(x, \theta) \| p(x, \theta \mid y, m)] \\
- \ln p(m') - \mathcal{F}'(q'_{x,\theta}) - \text{KL} [q'(x, \theta) \| p(x, \theta \mid y, m')] \\
\]

- Bias towards simple models
Hyperparameter optimization

\[ a^{(t+1)} = \arg \max_a \mathcal{F}_m(q_x(x), q_{\theta}(\theta), y, a) \]

**Figure 2.5:** The variational Bayesian EM algorithm with hyperparameter optimisation. The VBEM step consists of VBE and VBM steps, as shown in figure 2.3. The hyperparameter optimisation increases the lower bound and also improves the marginal likelihood.
Conjugate Exponential models

- complete data likelihood is from exponential family
- parameter prior is conjugate to complete data likelihood

\[
p(x_i, y_i \mid \theta) = g(\theta) f(x_i, y_i) e^{\phi(\theta)^\top u(x_i, y_i)}
\]

\[
p(\theta \mid \eta, \nu) = h(\eta, \nu) g(\theta)^\eta e^{\phi(\theta)^\top \nu}
\]

\[
p(\theta \mid \eta', \nu') \propto p(\theta \mid \eta, \nu)p(x, y \mid \theta)
\]
VBEM for CE-Models

**VBE Step:** Compute the expected sufficient statistics $\{\overline{u}(y_i)\}_{i=1}^n$ under the hidden variable distributions $q_{x_i}(x_i)$, for all $i$.

$$ q_{x_i}(x_i) \propto f(x_i, y_i) e^{\overline{\phi}^T u(x_i, y_i)} = p(x_i | y_i, \overline{\phi}) $$

$$ \overline{\phi} = \int d\theta \ q_\theta(\theta) \phi(\theta) \equiv \langle \phi(\theta) \rangle_{q_\theta(\theta)} $$

**VBM Step:** Compute the expected natural parameters $\overline{\phi} = \langle \phi(\theta) \rangle$ under the parameter distribution given by $\tilde{\eta}$ and $\tilde{\nu}$.

$$ q_\theta(\theta) = h(\tilde{\eta}, \tilde{\nu}) g(\theta) \tilde{\eta} e^{\phi(\theta)^T \tilde{\nu}} $$

$$ \overline{u}(y_i) = \langle u(x_i, y_i) \rangle_{q_{x_i}(x_i)} $$

$$ \tilde{\eta} = \eta + n $$

$$ \tilde{\nu} = \nu + \sum_{i=1}^n \overline{u}(y_i) $$
Experiment for structure learning

• 136 possible graph alternatives
That’s it.