Recent Advances in Bayesian Inference Techniques

Christopher M. Bishop
Microsoft Research, Cambridge, U.K.
research.microsoft.com/~cmbishop

SIAM Conference on Data Mining, April 2004
Abstract

Bayesian methods offer significant advantages over many conventional techniques such as maximum likelihood. However, their practicality has traditionally been limited by the computational cost of implementing them, which has often been done using Monte Carlo methods. In recent years, however, the applicability of Bayesian methods has been greatly extended through the development of fast analytical techniques such as variational inference. In this talk I will give a tutorial introduction to variational methods and will demonstrate their applicability in both supervised and unsupervised learning domains. I will also discuss techniques for automating variational inference, allowing rapid prototyping and testing of new probabilistic models.
Overview

• What is “Bayesian Inference”?
• Variational methods
• Example 1: uni-variate Gaussian
• Example 2: mixture of Gaussians
• Example 3: sparse kernel machines (RVM)
• Example 4: latent Dirichlet allocation
• Automatic variational inference
Maximum Likelihood

- Parametric model $p(x|\theta)$
- Data set (i.i.d.) $D = \{x_n\}$ where $n = 1, \ldots, N$
- Likelihood function
  \[
  L(\theta) = p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta)
  \]
- Maximize (log) likelihood
  \[
  \theta_{ML} = \arg \max_{\theta} \ln L(\theta)
  \]
- Predictive distribution
  \[
  p(x|D) \sim p(x|\theta_{ML})
  \]
Regularized Maximum Likelihood

- Prior $p(\theta)$, posterior $p(\theta|D) \propto p(D|\theta)p(\theta)$
- MAP (maximum posterior)
  \[ \theta_{\text{MAP}} = \arg \max_{\theta} \ln p(\theta|D) \]
- Predictive distribution
  \[ p(x|D) \simeq p(x|\theta_{\text{MAP}}) \]
- For example, if data model is Gaussian with unknown mean and if prior over mean is Gaussian
  \[ \ln p(\mu|D) = -\frac{\beta}{2} \sum_{n=1}^{N} ||x_n - \mu||^2 - \frac{\alpha}{2}||\mu||^2 + \text{const.} \]
  which is least squares with quadratic regularizer
Bayesian Inference

• Key idea is to *marginalize* over unknown parameters, rather than make point estimates

\[
p(x|D) = \int p(x|\theta)p(\theta|D) \, d\theta
\]

– avoids severe over-fitting of ML/MAP
– allows direct model comparison

• Most interesting probabilistic models also have hidden (latent) variables – we should marginalize over these too

• Such integrations (summations) are generally intractable

• Traditional approach: Markov chain Monte Carlo
  – computationally very expensive
  – limited to small data sets
This Talk

• Variational methods extend the practicality of Bayesian inference to “medium sized” data sets
Data Set Size

• Problem 1: learn the function $y(x) = \sin(2\pi x)$ for $x \in (0, 2\pi)$ from 100 (slightly) noisy examples
  – data set is computationally small but statistically large
• Problem 2: learn to recognize 1,000 everyday objects from 5,000,000 natural images
  – data set is computationally large but statistically small
• Bayesian inference
  – computationally more demanding than ML or MAP
  – significant benefit for statistically small data sets
Model Complexity

• A central issue in statistical inference is the choice of model complexity:
  – too simple ⇒ poor predictions
  – too complex ⇒ poor predictions (and slow on test)
• Maximum likelihood always favours more complex models – over-fitting
• It is usual to resort to cross-validation
  – computationally expensive
  – limited to one or two complexity parameters
• Bayesian inference can determine model complexity from training data – even with many complexity parameters
• Still a good idea to test final model on independent data
Variational Inference

- Goal: approximate posterior $p(\theta | D)$ by a simpler distribution $q(\theta)$ for which marginalization is tractable.
- Posterior related to joint $p(\theta, D)$ by marginal likelihood

$$p(D) = \int p(D | \theta) p(\theta) d\theta$$

- also a key quantity for model comparison
Variational Inference

- For an arbitrary $q(\theta)$ we have

$$\ln p(D) = \mathcal{L}(q) + \text{KL}(q \| p)$$

where

$$\mathcal{L}(q) = \int q(\theta) \ln \left( \frac{p(D, \theta)}{q(\theta)} \right) d\theta$$

$$\text{KL}(q \| p) = -\int q(\theta) \ln \left( \frac{p(\theta|D)}{q(\theta)} \right) d\theta$$

- Kullback-Leibler divergence satisfies

$$\text{KL}(q \| p) \geq 0$$
Variational Inference

• Choose $q(\theta)$ to maximize the lower bound

\[ \ln p(D) \]

\[ KL(q||p) \]

\[ \mathcal{L}(q) \]
Variational Inference

- Free-form optimization over $q(\theta)$ would give the true posterior distribution – but this is intractable by definition
- One approach would be to consider a parametric family of distributions and choose the best member
- Here we consider factorized approximations

$$q(\theta) = \prod_i q_i(\theta_i)$$

with free-form optimization of the factors $q_i(\theta_i)$
- A few lines of algebra shows that the optimum factors are

$$q_i(\theta_i) \propto \exp \left\{ \ln p(D, \theta) \right\}_{j \neq i}$$

- These are coupled so we need to iterate
$\mathsf{KL}(q\|p)$
Lower Bound

• Can also be evaluated
• Useful for maths + code verification
• Also useful for model comparison

\[ p(\mathcal{M}|D) \propto p(D|\mathcal{M})p(\mathcal{M}) \]

and hence

\[ \ln p(\mathcal{M}|D) \geq \mathcal{L}(q) + \ln p(\mathcal{M}) + \text{const}. \]
Example 1: Simple Gaussian

- Likelihood function

\[ p(D|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp \left\{ -\frac{\tau}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} \]

- Conjugate priors

\[ p(\mu|\mu_0, \lambda_0) = \mathcal{N}(\mu|\mu_0, \lambda_0^{-1}) \]
\[ p(\tau|a_0, b_0) = \mathcal{G}(\tau|a_0, b_0) \]

- Factorized variational distribution

\[ q(\mu, \tau) = q(\mu)q(\tau) \]
Variational Posterior Distribution

\[ q(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}) \]

\[ q(\tau) = \mathcal{G}(\tau|a_N, b_N) \]

where

\[ \mu_N = \frac{\lambda_0 \mu_0 + \langle \tau \rangle N \bar{x}}{\lambda_0 + N \langle \tau \rangle} \]

\[ \lambda_N = \lambda_0 + N \langle \tau \rangle \]

\[ a_N = a_0 + \frac{N}{2} \]

\[ b_N = b_0 + \frac{1}{2} \left( \sum_n (x_n - \mu)^2 \right)_\mu \]
Initial Configuration
After Updating $q(\mu)$
After Updating $q(\tau)$
Converged Solution

(d)
Applications of Variational Inference

- Hidden Markov models (MacKay)
- Neural networks (Hinton)
- Bayesian PCA (Bishop)
- Independent Component Analysis (Attias)
- Mixtures of Gaussians (Attias; Ghahramani and Beal)
- Mixtures of Bayesian PCA (Bishop and Winn)
- Flexible video sprites (Frey et al.)
- Audio-video fusion for tracking (Attias et al.)
- Latent Dirichlet Allocation (Jordan et al.)
- Relevance Vector Machine (Tipping and Bishop)
- Object recognition in images (Li et al.)
- …
Example 2: Gaussian Mixture Model

• Linear superposition of Gaussians

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \]

• Conventional maximum likelihood solution using EM:
  – E-step: evaluate responsibilities

\[ \gamma_k(x) \equiv p(k|x) = \frac{\pi_k \mathcal{N}(x | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x | \mu_j, \Sigma_j)} \]
Gaussian Mixture Model

- M-step: re-estimate parameters

\[
\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n)
\]

\[
\mu_j = \frac{1}{N\pi_j} \sum_{n=1}^{N} \gamma_j(x_n)x_n
\]

\[
\Sigma_j = \frac{1}{N\pi_j} \sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu_j)(x_n - \mu_j)^T
\]

- Problems:
  - singularities
  - how to choose \( K \)?
Bayesian Mixture of Gaussians

- Conjugate priors for the parameters:
  - Dirichlet prior for mixing coefficients
    \[
p(\pi) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}
    \]
  - normal-Wishart prior for means and precisions
    \[
p(\mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, \beta_0^{-1} \Lambda_k^{-1}) \mathcal{W}(\Lambda_k | W_0, \nu_0)
    \]

where the Wishart distribution is given by

\[
\mathcal{W}(\Lambda | W, \nu) \propto |\Lambda|^{(\nu - d - 1)/2} \exp \left( -\frac{1}{2} \text{Tr}(W^{-1} \Lambda) \right)
\]
Graphical Representation

- Parameters and latent variables appear on equal footing
Variational Mixture of Gaussians

• Assume factorized posterior distribution

\[ q(z, \pi, \mu, \Lambda) = q(z)q(\pi, \mu, \Lambda) \]

• No other assumptions!
• Gives optimal solution in the form

\[
q(\pi, \mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k) \\
q(z) = \prod_{n=1}^{N} q(z_n)
\]

where \( q(\pi) \) is a Dirichlet, and \( q(\mu_k, \Lambda_k) \) is a Normal-Wishart; \( q(z_n) \) is multinomial
Sufficient Statistics

\[
N_k = \sum_{n=1}^{N} \langle z_{nk} \rangle \\
\bar{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} \langle z_{nk} \rangle x_n \\
S_k = \frac{1}{N_k} \sum_{n=1}^{N} \langle z_{nk} \rangle (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T
\]

- Similar computational cost to maximum likelihood EM
- No singularities!
- Predictive distribution is mixture of Student-t distributions
Variational Equations for GMM

\[ q^*(z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \quad \langle z_{nk} \rangle = r_{nk} \]

\[ r_{nk} \propto \tilde{\pi}_k \tilde{\Lambda}_k^{1/2} \exp \left\{ -\frac{d}{2\beta_k} - \frac{\nu_k}{2} (x_n - m_k)^T W_k (x_n - m_k) \right\} \]

\[ \ln \tilde{\Lambda}_k = \sum_{i=1}^{d} \psi \left( \frac{\nu_k + 1 - i}{2} \right) + d \ln 2 - \ln |W_k| \]

\[ \ln \tilde{\pi}_k = \psi(\alpha_k) - \psi(\bar{\alpha}) \]

\[ q^*(\pi) = \mathcal{D}(\pi | \alpha) \quad \alpha_k = \alpha_0 + N_k \quad \nu_k = \nu_0 + N_k \]

\[ q^*(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k | m_k, \beta_k^{-1} \Lambda_k^{-1}) \mathcal{W}(\Lambda_k | W_k, \nu_k) \]

\[ m_k = \frac{1}{\beta_k} (\beta_0 m_0 + N_k \bar{x}_k) \quad \beta_k = \beta_0 + N_k \]

\[ W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T \]
Old Faithful Data Set

Time between eruptions (minutes)

Duration of eruption (minutes)
Bound vs. K for Old Faithful Data

Christopher M. Bishop

SIAM Conference on Data Mining, April 2004
Bayesian Model Complexity

\[ p(D|K) \]

Christopher M. Bishop  
SIAM Conference on Data Mining, April 2004
Sparse Gaussian Mixtures

- Instead of comparing different values of $K$, start with a large value and prune out excess components.
- Achieved by treating mixing coefficients as parameters, and maximizing marginal likelihood (Corduneanu and Bishop, AI Stats 2001).
- Gives simple re-estimation equations for the mixing coefficients – interleave with variational updates.
Example 3: RVM

- Bayesian alternative to support vector machine (SVM)
- Limitations of the SVM:
  - two classes
  - large number of kernels (in spite of sparsity)
  - kernels must satisfy Mercer criterion
  - cross-validation to set parameters $C$ (and $\epsilon$)
  - decisions at outputs instead of probabilities
Relevance Vector Machine

- Linear model as for SVM
  \[ y(x, w) = \sum_{m=0}^{M} w_m \phi_m(x) = w^T \phi \]

- Input vectors \( \{x_n\}_{n=1}^{N} \) and targets \( \{t_n\}_{n=1}^{N} \)

- Regression
  \[ p(t|x, w, \tau) = \mathcal{N}(t|y(x, w), \tau^{-1}) \]

- Classification
  \[ p(t|x, w) = \sigma(y)^t (1 - \sigma(y))^{1-t} \]
Relevance Vector Machine

- Gaussian prior for $w_m$ with hyper-parameters $\alpha_m$

$$p(w|\alpha) = \prod_{m=0}^{N} \mathcal{N}(w_m|0, \alpha_m^{-1})$$

- Hyper-priors over $\alpha$ (and $\tau$ if regression)
- A high proportion of $\alpha_m$ are driven to large values in the posterior distribution, and corresponding $w_m$ are driven to zero, giving a sparse model
Relevance Vector Machine

• Graphical model representation (regression)

• For classification use sigmoid (or softmax for multi-class) outputs and omit noise node
Relevance Vector Machine

• Regression: synthetic data
Relevance Vector Machine

- Regression: synthetic data

\[ \log_{10} \langle \alpha \rangle \]
Relevance Vector Machine

- Classification: SVM
Relevance Vector Machine

- Classification: VRVM
Relevance Vector Machine

- Results on Ripley regression data (Bayes error rate 8%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Error</th>
<th>No. kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>10.6%</td>
<td>38</td>
</tr>
<tr>
<td>VRVM</td>
<td>9.2%</td>
<td>4</td>
</tr>
</tbody>
</table>

- Results on classification benchmark data

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>GP</th>
<th>RVM</th>
<th>SVM</th>
<th>GP</th>
<th>RVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pima Indians</td>
<td>67</td>
<td>68</td>
<td>65</td>
<td>109</td>
<td>200</td>
<td>4</td>
</tr>
<tr>
<td>U.S.P.S.</td>
<td>4.4%</td>
<td>-</td>
<td>5.1%</td>
<td>2540</td>
<td>-</td>
<td>316</td>
</tr>
</tbody>
</table>
Relevance Vector Machine

• Properties
  – comparable error rates to SVM on new data
  – no cross-validation to set complexity parameters
  – applicable to wide choice of basis function
  – multi-class classification
  – probabilistic outputs
  – dramatically fewer kernels (by an order of magnitude)
  – but, slower to train than SVM
Fast RVM Training

- Tipping and Faul (2003)
- Analytic treatment of each hyper-parameter in turn
- Applied to 250,000 image patches (face/non-face)
- Recently applied to regression with $10^6$ data points
Face Detection
Face Detection
Face Detection
Face Detection
Face Detection
Example 4: Latent Dirichlet Allocation

- Blei, Jordan and Ng (2003)
- Generative model of documents (but broadly applicable e.g. collaborative filtering, image retrieval, bioinformatics)
- Generative model:
  - choose $\theta \sim \text{Dir}(\alpha)$
  - choose topic $z_n \sim \text{Mult}(\theta)$
  - choose word $w_n \sim p(w_n|z_n, \beta)$
Latent Dirichlet Allocation

- Variational approximation

\[ q(\theta, z) = q_\theta(\theta)q_z(z) \]

- Data set:
  - 15,000 documents
  - 90,000 terms
  - 2.1 million words

- Model:
  - 100 factors
  - 9 million parameters

- MCMC totally infeasible for this problem
Automatic Variational Inference

• Currently for each new model we have to
  1. derive the variational update equations
  2. write application-specific code to find the solution
• Each can be time consuming and error prone
• Can we build a general-purpose inference engine which automates these procedures?
VIBES

- Variational Inference for Bayesian Networks
- A general inference engine using variational methods
- Analogous to BUGS for MCMC
- VIBES is available on the web

http://vibes.sourceforge.net/index.shtml
VIBES (cont’d)

• A key observation is that in the general solution

\[ q_i(\theta_i) \propto \exp \langle \ln p(D, \theta) \rangle_{j \neq i} \]

the update for a particular node (or group of nodes) depends only on other nodes in the Markov blanket

• Permits a local message-passing framework which is independent of the particular graph structure
VIBES (cont’d)

[Diagram of a Bayesian network with nodes labeled pi, s, x, alpha, t, W, mu, and an edge labeled x.W+mu connecting x and W]
VIBES (cont’d)
VIBES (cont’d)
New Book

- *Pattern Recognition and Machine Learning*
  - Springer (2005)
  - 600 pages, hardback, four colour, maximum $75
  - Graduate level text book
  - Worked solutions to all 250 exercises
  - Complete lectures on www
  - Companion software text with Ian Nabney
  - Matlab software on www
Conclusions

• Variational inference: a broad class of new semi-analytical algorithms for Bayesian inference
• Applicable to much larger data sets than MCMC
• Can be automated for rapid prototyping
Viewgraphs and tutorials available from:

research.microsoft.com/~cmbishop