Unsupervised Learning

Bayesian Model Selection

Zoubin Ghahramani
zoubin@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc in Intelligent Systems, Dept Computer Science
University College London

Autumn 2003
Model structure and overfitting: a simple example
Learning Model Structure

How many clusters in the data?

What is the intrinsic dimensionality of the data?

Is this input relevant to predicting that output?

What is the order of a dynamical system?

How many states in a hidden Markov model?

How many auditory sources in the input?
Using Occam’s Razor to Learn Model Structure

Select the model class $\mathcal{M}_i$ with the highest probability given the data:

$$P(\mathcal{M}_i | y) = \frac{P(y | \mathcal{M}_i) P(\mathcal{M}_i)}{P(y)}, \quad P(y | \mathcal{M}_i) = \int_{\Theta_i} P(y | \theta_i, \mathcal{M}_i) P(\theta_i | \mathcal{M}_i) \, d\theta_i$$

**Interpretation of $P(y | \mathcal{M}_i)$**: The probability that randomly selected parameter values from the model class would generate data set $y$.

Model classes that are **too simple** are unlikely to generate the data set.

Model classes that are **too complex** can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.

---

![Graph showing the probability of data sets generated by simple, just right, and complex models.](image)

The graph illustrates the probability $P(Y | \mathcal{M}_i)$ for different model classes. The green peak represents a very simple model, the red peak represents a model that is just right, and the blue peak represents a very complex model. The y-axis represents the probability, and the x-axis represents all possible data sets.
Bayesian Model Selection: Occam’s Razor at Work

e.g. for quadratic (M=2): \( y = a_0 + a_1 x + a_2 x^2 + \epsilon \), where \( \epsilon \sim \mathcal{N}(0, \tau) \) and \( \theta_2 = [a_0 \ a_1 \ a_2 \ \tau] \)

demo: polybayes
Practical Bayesian approaches

• Laplace approximations:
  – Appeals to Central Limit Theorem making a Gaussian approximation about maximum \textit{a posteriori} parameter estimate.

• Large sample approximations (e.g. BIC).

• Markov chain Monte Carlo methods (MCMC):
  – In the limit are guaranteed to converge, but:
    – Many samples required to ensure accuracy.
    – Sometimes hard to assess convergence.

• Variational approximations

Note: other deterministic approximations are also available now: e.g. Bethe approximations and Expectation Propagation
Laplace Approximation

data set $y$, models $\mathcal{M}_1 \ldots, \mathcal{M}_n$, parameter sets $\theta_1 \ldots, \theta_n$

Model Selection: \[ P(\mathcal{M}_i | y) \propto P(\mathcal{M}_i) P(y | \mathcal{M}_i) \]

For large amounts of data (relative to number of parameters, $d$) the parameter posterior is approximately Gaussian around the MAP estimate $\hat{\theta}_i$:

\[
P(\theta_i | y, \mathcal{M}_i) \approx (2\pi)^{-d/2} |A|^{1/2} \exp \left\{ -\frac{1}{2}(\theta_i - \hat{\theta}_i)^\top A(\theta_i - \hat{\theta}_i) \right\}
\]

\[
P(y | \mathcal{M}_i) = \frac{P(\theta_i, y | \mathcal{M}_i)}{P(\theta_i | y, \mathcal{M}_i)}
\]

Evaluating the above expression for $\ln P(y | \mathcal{M}_i)$ at $\hat{\theta}_i$:

\[
\ln P(y | \mathcal{M}_i) \approx \ln P(\hat{\theta}_i | \mathcal{M}_i) + \ln P(y | \hat{\theta}_i, \mathcal{M}_i) + \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |A|
\]

where $A$ is the $d \times d$ negative Hessian matrix of the log posterior,

\[
A_{\ell m} = -\frac{\partial^2}{\partial \theta_{i\ell} \partial \theta_{i m}} \ln P(\theta_i | y, \mathcal{M}_i) |_{\hat{\theta}_i}.
\]

This can be used for model selection.
Bayesian Information Criterion (BIC)

BIC can be obtained from the Laplace approximation:

\[
\ln P(y|M_i) \approx \ln P(\hat{\theta}_i|M_i) + \ln P(y|\hat{\theta}_i, M_i) + \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |A|
\]

in the large sample limit \((N \to \infty)\) where \(N\) is the number of data points, \(A\) grows as \(NA_0\) for some fixed matrix \(A_0\), so \(\ln |A| \to \ln |NA_0| = \ln (N^d|A_0|) = d \ln N + \ln |A_0|\). Retaining only terms that grow in \(N\) we get:

\[
\ln P(y|M_i) \approx \ln P(y|\hat{\theta}_i, M_i) - \frac{d}{2} \ln N
\]

Properties:

- Quick and easy to compute
- It does not depend on the prior
- We can use the ML estimate of \(\theta\) instead of the MAP estimate
- It is equivalent to the MDL criterion
- It assumes that in the large sample limit, all the parameters are well-determined (i.e. the model is identifiable; otherwise, \(d\) should be the number of well-determined parameters)
- **Danger:** counting parameters can be deceiving! (c.f. sinusoid, infinite models)
MCMC Approximations

Let’s consider a non-Markov chain method, **Importance Sampling:**

\[
\ln P(y|M_i) = \ln \int_{\Theta_i} P(y|\theta_i, M_i)P(\theta_i|M_i) \, d\theta_i
\]

\[
= \ln \int_{\Theta_i} P(y|\theta_i, M_i) \frac{P(\theta_i|M_i)}{Q(\theta_i)} Q(\theta_i) \, d\theta_i
\]

\[
\approx \ln \sum_k P(y|\theta^{(k)}_i, M_i) \frac{P(\theta^{(k)}_i|M_i)}{Q(\theta^{(k)}_i)}
\]

where \(\theta^{(k)}_i\) are i.i.d. draws from \(Q(\theta_i)\). Assumes we can sample from and evaluate \(Q(\theta_i)\) (incl. normalization!) and we can compute the likelihood \(P(y|\theta^{(k)}_i, M_i)\).

Although importance sampling does not work well in high dimensions, it inspires the following approach: Create a **Markov chain**, \(Q_k \rightarrow Q_{k+1} \ldots\) for which:

- \(Q_k(\theta)\) can be evaluated including normalization
- \(\lim_{k \to \infty} Q_k(\theta) = P(\theta|y, M_i)\)
Variational Bayesian Learning
Lower Bounding the Evidence

Let the hidden latent variables be $x$, data $y$ and the parameters $\theta$. We can lower bound the evidence (Jensen’s inequality):

$$\ln P(y|\mathcal{M}) = \ln \int dx \, d\theta \, P(y, x, \theta|\mathcal{M})$$

$$= \ln \int dx \, d\theta \, Q(x, \theta) \frac{P(y, x, \theta)}{Q(x, \theta)}$$

$$\geq \int dx \, d\theta \, Q(x, \theta) \ln \frac{P(y, x, \theta)}{Q(x, \theta)}.$$

Use a simpler, factorised approximation to $Q(x, \theta)$:

$$\ln P(y) \geq \int dx \, d\theta \, Q_x(x)Q_\theta(\theta) \ln \frac{P(y, x, \theta)}{Q_x(x)Q_\theta(\theta)}$$

$$= \mathcal{F}(Q_x(x), Q_\theta(\theta), y).$$
Variational Bayesian Learning . . .

Maximizing this lower bound, $\mathcal{F}$, leads to EM-like updates:

\[
Q^*_x(x) \propto \exp \langle \ln P(x,y|\theta) \rangle_{Q_\theta(\theta)} \quad \text{E-like step}
\]

\[
Q^*_\theta(\theta) \propto P(\theta) \exp \langle \ln P(x,y|\theta) \rangle_{Q_x(x)} \quad \text{M-like step}
\]

Maximizing $\mathcal{F}$ is equivalent to minimizing KL-divergence between the approximate posterior, $Q(\theta)Q(x)$ and the true posterior, $P(\theta, x|y)$. 
Conjugate-Exponential models

Let’s focus on conjugate-exponential (CE) models, which satisfy (1) and (2):

**Condition (1).** The joint probability over variables is in the exponential family:

\[
P(x, y | \theta) = f(x, y) g(\theta) \exp \{ \phi(\theta)^\top u(x, y) \}
\]

where \( \phi(\theta) \) is the vector of natural parameters, \( u \) are sufficient statistics

**Condition (2).** The prior over parameters is conjugate to this joint probability:

\[
P(\theta | \eta, \nu) = h(\eta, \nu) g(\theta)^\eta \exp \{ \phi(\theta)^\top \nu \}
\]

where \( \eta \) and \( \nu \) are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- \( \eta \): number of pseudo-observations
- \( \nu \): values of pseudo-observations
Conjugate-Exponential examples

In the CE family:

• Gaussian mixtures
• factor analysis, probabilistic PCA
• hidden Markov models and factorial HMMs
• linear dynamical systems and switching models
• discrete-variable belief networks

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Not in the CE family:

• Boltzmann machines, MRFs (no simple conjugacy)
• logistic regression (no simple conjugacy)
• sigmoid belief networks (not exponential)
• independent components analysis (not exponential)

Note: one can often approximate these models with models in the CE family.
A Useful Result

**Theorem** Given an iid data set \( y = (y_1, \ldots y_n) \), if the model is CE then:

(a) \( Q_\theta(\theta) \) is also conjugate, i.e.

\[
Q_\theta(\theta) = h(\tilde{\eta}, \tilde{\nu}) g(\theta) \tilde{\eta} \exp \{ \phi(\theta)^\top \tilde{\nu} \}
\]

where \( \tilde{\eta} = \eta + n \) and \( \tilde{\nu} = \nu + \sum_i u(x_i, y_i) \).

(b) \( Q_x(x) = \prod_{i=1}^n Q_{x_i}(x_i) \) is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over \( Q_\theta(\theta) \)

\[
Q_{x_i}(x_i) \propto f(x_i, y_i) \exp \{ \phi(\theta)^\top u(x_i, y_i) \} = P(x_i|y_i, \phi(\theta))
\]

**KEY points:**

(a) the approximate parameter posterior is of the same form as the prior, so it is easily summarized in terms of two sets of hyperparameters, \( \tilde{\eta} \) and \( \tilde{\nu} \);

(b) the approximate hidden variable posterior, averaging over all parameters, is of the same form as the hidden variable posterior for a single setting of the parameters, so again, it is easily computed using the usual methods.
## The Variational Bayesian EM algorithm

<table>
<thead>
<tr>
<th><strong>EM for MAP estimation</strong></th>
<th><strong>Variational Bayesian EM</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal:</strong> maximize $p(\theta</td>
<td>y, m)$ w.r.t. $\theta$</td>
</tr>
<tr>
<td><strong>E Step:</strong> compute</td>
<td><strong>VB-E Step:</strong> compute</td>
</tr>
<tr>
<td>$q_{\theta}^{(t+1)}(x) = p(x</td>
<td>y, \theta^{(t)})$</td>
</tr>
<tr>
<td><strong>M Step:</strong></td>
<td><strong>VB-M Step:</strong></td>
</tr>
<tr>
<td>$\theta^{(t+1)} = \arg\max \int q_{\theta}^{(t+1)}(x) \ln p(x, y, \theta) , dx$</td>
<td>$q_{\theta}^{(t+1)}(\theta) = \exp \left[ \int q_{\theta}^{(t+1)}(x) \ln p(x, y, \theta) , dx \right]$</td>
</tr>
</tbody>
</table>

### Properties:
- Reduces to the EM algorithm if $q_\theta(\theta) = \delta(\theta - \theta^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\tilde{\phi}$. 
Variational Bayes: History of Models Treated

- multilayer perceptrons (Hinton & van Camp, 1993)
- mixture of experts (Waterhouse, MacKay & Robinson, 1996)
- hidden Markov models (MacKay, 1995)
- other work by Jaakkola, Jordan, Barber, Bishop, Tipping, etc

Examples of Variational Learning of Model Structure

- mixtures of factor analysers (Ghahramani & Beal, 1999)
- mixtures of Gaussians (Attias, 1999)
- independent components analysis (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- principal components analysis (Bishop, 1999)
- linear dynamical systems (Ghahramani & Beal, 2000)
- mixture of experts (Ueda & Ghahramani, 2000)
- discrete graphical models (Beal & Ghahramani, 2002)
- VIBES software for conjugate-exponential graphs (Winn, 2003)
Mixture of Factor Analysers

Goal:

- Infer number of clusters
- Infer intrinsic dimensionality of each cluster

Under the assumption that each cluster is Gaussian

embed_demo
Mixture of Factor Analysers

**True data:** 6 Gaussian clusters with dimensions: (1 7 4 3 2 2) embedded in 10-D

**Inferred structure:**

<table>
<thead>
<tr>
<th>number of points per cluster</th>
<th>intrinsic dimensionalities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td></td>
</tr>
</tbody>
</table>

- Finds the clusters and dimensionalities efficiently.
- The model complexity reduces in line with the lack of data support.

demos: run_simple and ueda_demo
Hidden Markov Models

Discrete hidden states, $s_t$.  
Observations $y_t$.

How many hidden states?  
What structure state-transition matrix?

demo: vbhmm_demo
Hidden Markov Models:
Discriminating Forward from Reverse English

First 8 sentences from Alice in Wonderland.
Compare VB-HMM with ML-HMM.

Discriminating Forward and Reverse English
Model Selection using F

Classification Accuracy
Hidden States
Linear Dynamical Systems

- Assumes $y_t$ generated from a sequence of Markov hidden state variables $x_t$
- If transition and output functions are linear, time-invariant, and noise distributions are Gaussian, this is a linear-Gaussian state-space model:

$$x_t = Ax_{t-1} + w_t, \quad y_t = Cx_t + v_t$$

- Three levels of inference:
  1. Given data, structure and parameters, **Kalman smoothing** $\rightarrow$ hidden state;
  2. Given data and structure, **EM** $\rightarrow$ hidden state and parameter point estimates;
  3. Given data only, **VEM** $\rightarrow$ model structure and distributions over parameters and hidden state.
Linear Dynamical System Results

Inferring model structure:

a) SSM(0,3) i.e. FA  b) SSM(3,3)  c) SSM(3,4)

Inferred model complexity reduces with less data:

True model:  SSM(6,6)  • 10-dim observation vector.

demo: bayeslds
Independent Components Analysis

Blind Source Separation: $5 \times 100$ msec speech and music sources linearly mixed to produce 11 signals (microphones)

Figure 2. Application of VB to blind source separation algorithm (see text).

from Attias (2000)
Summary & Conclusions

- Bayesian learning avoids overfitting and can be used to do model selection.

- But we need approximations:

  - Laplace
  - BIC
  - Sampling
  - Variational...