An Introduction to Bayesian Machine Learning for Multimedia Information Processing
Part I - Introduction

A. Taylan Cemgil
Signal Processing and Communications Lab.

UNIVERSITY OF CAMBRIDGE
Department of Engineering

2008 IEEE ICME Tutorial, 23 June, Hannover
Goals of this Tutorial

• Provide a basic understanding of underlying principles of probabilistic modeling and Bayesian inference

• Orientation in the broad literature of Bayesian machine learning and statistical signal processing

• Focus on fundamental concepts rather than technical details,
  … we avoid heavy use of algebra by a graphical notation
  … but there will be some maths
Goals of this Tutorial

• Model based approach
  . . . rather than description of algorithms for solving specific problems

• Illustrate with examples how certain problems in multimedia signal analysis can be approached using generic tools

• Motivate participants to investigate further
  . . . provide alternative perspective to existing solutions
  . . . and hopefully provide new inspiration
Part I, Introduction

- Introduction
  - Bayes’ Theorem,
  - Trivial toy example to clarify notation

- Graphical Models
  - Bayesian Networks
  - Undirected Graphical models, Markov Random Fields
  - Factor graphs

- Maximum Likelihood, Penalised Likelihood, Bayesian Learning
Part II, Basic Modelling and Inference Strategies

- Basic Building Blocks in model construction
  - Probability distributions, Exponential family

- Approximate Inference
  - Stochastic
    * Markov Chain Monte Carlo (MCMC), Gibbs sampler
    * Simulated Annealing, Iterative Improvement (SA - II)
  - Deterministic Inference
    * Variational Bayes (VB)
    * Expectation-Maximisation (EM)
    * Iterative conditional modes (ICM)
Part III, Models and Applications

• Hidden Markov Models,
  – Tempo tracking, Score-performance matching
  – Inference in Hidden Markov Models
    • Forward Backward Algorithm
    • Viterbi
    • Exact inference by message passing: Belief Propagation

• Linear Dynamical systems, Kalman Filter Models
  – Tracking
  – Computer Accompaniment
  – Kalman Filtering and Smoothing
  – Audio Restoration and Interpolation∗
• Nonlinear Dynamical Systems
  – Object tracking in video
  – Importance sampling, Particle Filtering, Sequential Monte Carlo
  – Switching State Space models, Changepoint Models
  – Pitch tracking

• Markov Random Fields
  – Denoising, Source Separation

• Topic-Term-Document Models
  – Latent Semantic indexing
  – Generative aspect model
  – Non Negative Matrix Factorisation
• Factorial Models, Model selection
  – Audio Source Separation
  – Polyphonic Pitch Tracking

• Final Remarks and Bibliography
Bayes’ Theorem \[1, 3\]

Thomas Bayes (1702-1761)

What you know about a parameter $\lambda$ after the data $\mathcal{D}$ arrive is what you knew before about $\lambda$ and what the data $\mathcal{D}$ told you.

$$p(\lambda | \mathcal{D}) = \frac{p(\mathcal{D} | \lambda) p(\lambda)}{p(\mathcal{D})}$$

Posterior = \frac{Likelihood \times Prior}{Evidence}
An application of Bayes’ Theorem: “Source Separation”

Given two fair dice with outcomes $\lambda$ and $y$,

$$\mathcal{D} = \lambda + y$$

What is $\lambda$ when $\mathcal{D} = 9$?
An application of Bayes’ Theorem: “Source Separation”

\[ D = \lambda + y = 9 \]

<table>
<thead>
<tr>
<th>( D = \lambda + y )</th>
<th>( y = 1 )</th>
<th>( y = 2 )</th>
<th>( y = 3 )</th>
<th>( y = 4 )</th>
<th>( y = 5 )</th>
<th>( y = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda = 1 )</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>( \lambda = 2 )</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>( \lambda = 3 )</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>( \lambda = 4 )</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>( \lambda = 5 )</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>( \lambda = 6 )</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Bayes theorem “upgrades” \( p(\lambda) \) into \( p(\lambda|D) \).

But you have to provide an observation model: \( p(D|\lambda) \)
“Bureaucratical” derivation

Formally we write

\[
p(\lambda) = C(\lambda; \left[ \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \right])
\]

\[
p(y) = C(y; \left[ \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \frac{1}{6} \right])
\]

\[
p(\mathcal{D}|\lambda, y) = \delta(\mathcal{D} - (\lambda + y))
\]

\[
p(\lambda, y|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \times p(\mathcal{D}|\lambda, y) \times p(y)p(\lambda)
\]

Posterior = \frac{1}{\text{Evidence}} \times \text{Likelihood} \times \text{Prior}

Kronecker delta function denoting a degenerate (deterministic) distribution

\[
\delta(x) = \begin{cases} 
1 & x = 0 \\
0 & x \neq 0 
\end{cases}
\]
Prior

\[ p(y)p(\lambda) \]

<table>
<thead>
<tr>
<th>( p(y) \times p(\lambda) )</th>
<th>( y = 1 )</th>
<th>( y = 2 )</th>
<th>( y = 3 )</th>
<th>( y = 4 )</th>
<th>( y = 5 )</th>
<th>( y = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda = 1 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 2 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 3 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 4 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 5 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 6 )</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
</tbody>
</table>

- A table with indices \( \lambda \) and \( y \)
- Each cell denotes the probability \( p(\lambda, y) \)
**Likelihood**

\[ p(D = 9|\lambda, y) \]

| \( p(D = 9|\lambda, y) \) | \( y = 1 \) | \( y = 2 \) | \( y = 3 \) | \( y = 4 \) | \( y = 5 \) | \( y = 6 \) |
|-------------------------|---------|---------|---------|---------|---------|---------|
| \( \lambda = 1 \)      | 0       | 0       | 0       | 0       | 0       | 0       |
| \( \lambda = 2 \)      | 0       | 0       | 0       | 0       | 0       | 0       |
| \( \lambda = 3 \)      | 0       | 0       | 0       | 0       | 0       | 1       |
| \( \lambda = 4 \)      | 0       | 0       | 0       | 0       | 1       | 0       |
| \( \lambda = 5 \)      | 0       | 0       | 0       | 1       | 0       | 0       |
| \( \lambda = 6 \)      | 0       | 0       | 1       | 0       | 0       | 0       |

- A table with indices \( \lambda \) and \( y \)
- The likelihood is **not** a probability distribution, but a positive function.
Likelihood $\times$ Prior

$$\phi_D(\lambda, y) = p(\mathcal{D} = 9|\lambda, y)p(\lambda)p(y)$$

| $p(\mathcal{D} = 9|\lambda, y)$ | $y = 1$ | $y = 2$ | $y = 3$ | $y = 4$ | $y = 5$ | $y = 6$ |
|-------------------------------|-------|-------|-------|-------|-------|-------|
| $\lambda = 1$                 | 0     | 0     | 0     | 0     | 0     | 0     |
| $\lambda = 2$                 | 0     | 0     | 0     | 0     | 0     | 0     |
| $\lambda = 3$                 | 0     | 0     | 0     | 0     | 0     | $1/36$|
| $\lambda = 4$                 | 0     | 0     | 0     | $1/36$| 0     | 0     |
| $\lambda = 5$                 | 0     | 0     | $1/36$| 0     | 0     | 0     |
| $\lambda = 6$                 | $1/36$| 0     | 0     | 0     | 0     | 0     |
Evidence (= Marginal Likelihood)

\[ p(\mathcal{D} = 9) = \sum_{\lambda, y} p(\mathcal{D} = 9|\lambda, y)p(\lambda)p(y) \]

\[ = 0 + 0 + \cdots + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + 0 + \cdots + 0 \]

\[ = \frac{1}{9} \]

| \( p(\mathcal{D} = 9|\lambda, y) \) | \( y = 1 \) | \( y = 2 \) | \( y = 3 \) | \( y = 4 \) | \( y = 5 \) | \( y = 6 \) |
|---|---|---|---|---|---|---|
| \( \lambda = 1 \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( \lambda = 2 \) | 0 | 0 | 0 | 0 | 0 | 0 |
| \( \lambda = 3 \) | 0 | 0 | 0 | 0 | 0 | \( \frac{1}{36} \) |
| \( \lambda = 4 \) | 0 | 0 | 0 | 0 | \( \frac{1}{36} \) | 0 |
| \( \lambda = 5 \) | 0 | 0 | 0 | \( \frac{1}{36} \) | 0 | 0 |
| \( \lambda = 6 \) | 0 | 0 | \( \frac{1}{36} \) | 0 | 0 | 0 |
Posterior

\[ p(\lambda, y|D = 9) = \frac{1}{p(D)} p(D = 9|\lambda, y)p(\lambda)p(y) \]

| \( p(D = 9|\lambda, y) \) | \( y = 1 \) | \( y = 2 \) | \( y = 3 \) | \( y = 4 \) | \( y = 5 \) | \( y = 6 \) |
|-----------------|-------|-------|-------|-------|-------|-------|
| \( \lambda = 1 \) | 0     | 0     | 0     | 0     | 0     | 0     |
| \( \lambda = 2 \) | 0     | 0     | 0     | 0     | 0     | 0     |
| \( \lambda = 3 \) | 0     | 0     | 0     | 0     | 0     | 1/4   |
| \( \lambda = 4 \) | 0     | 0     | 0     | 0     | 1/4   | 0     |
| \( \lambda = 5 \) | 0     | 0     | 1/4   | 0     | 0     | 0     |
| \( \lambda = 6 \) | 0     | 0     | 1/4   | 0     | 0     | 0     |

\[ 1/4 = (1/36)/(1/9) \]
Marginal Posterior

\[ p(\lambda|\mathcal{D}) = \sum_{y} \frac{1}{p(\mathcal{D})} p(\mathcal{D}|\lambda, y) p(\lambda) p(y) \]

| $\lambda$  | $p(\lambda|\mathcal{D} = 9)$ | $y = 1$ | $y = 2$ | $y = 3$ | $y = 4$ | $y = 5$ | $y = 6$ |
|-----------|-------------------------------|---------|---------|---------|---------|---------|---------|
| $\lambda = 1$ | 0                             | 0       | 0       | 0       | 0       | 0       | 0       |
| $\lambda = 2$ | 0                             | 0       | 0       | 0       | 0       | 0       | 0       |
| $\lambda = 3$ | 1/4                           | 0       | 0       | 0       | 0       | 0       | 1/4     |
| $\lambda = 4$ | 1/4                           | 0       | 0       | 0       | 0       | 1/4     | 0       |
| $\lambda = 5$ | 1/4                           | 0       | 0       | 0       | 1/4     | 0       | 0       |
| $\lambda = 6$ | 1/4                           | 0       | 0       | 1/4     | 0       | 0       | 0       |
The “proportional to” notation

\[ p(\lambda|\mathcal{D} = 9) \propto p(\lambda, \mathcal{D} = 9) = \sum_{y} p(\mathcal{D} = 9|\lambda, y)p(\lambda)p(y) \]

<table>
<thead>
<tr>
<th>( \lambda = 1 )</th>
<th>( p(\lambda, \mathcal{D} = 9) )</th>
<th>( y = 1 )</th>
<th>( y = 2 )</th>
<th>( y = 3 )</th>
<th>( y = 4 )</th>
<th>( y = 5 )</th>
<th>( y = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda = 1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \lambda = 2 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \lambda = 3 )</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
</tr>
<tr>
<td>( \lambda = 4 )</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>0</td>
</tr>
<tr>
<td>( \lambda = 5 )</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \lambda = 6 )</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Another application of Bayes’ Theorem: “Model Selection”

Given an unknown number of fair dice with outcomes $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$D = \sum_{i=1}^{n} \lambda_i$$

How many dice are there when $D = 9$?

Assume that any number $n$ is equally likely
Another application of Bayes’ Theorem: “Model Selection”

Given all \( n \) are equally likely (i.e., \( p(n) \) is flat), we calculate (formally)

\[
p(n | \mathcal{D} = 9) = \frac{p(\mathcal{D} = 9 | n)p(n)}{p(\mathcal{D})} \propto p(\mathcal{D} = 9 | n)
\]

\[
p(\mathcal{D} | n = 1) = \sum_{\lambda_1} p(\mathcal{D} | \lambda_1)p(\lambda_1)
\]

\[
p(\mathcal{D} | n = 2) = \sum_{\lambda_1} \sum_{\lambda_2} p(\mathcal{D} | \lambda_1, \lambda_2)p(\lambda_1)p(\lambda_2)
\]

\[
\ldots
\]

\[
p(\mathcal{D} | n = n') = \sum_{\lambda_1, \ldots, \lambda_{n'}} p(\mathcal{D} | \lambda_1, \ldots, \lambda_{n'}) \prod_{i=1}^{n'} p(\lambda_i)
\]
\[ p(D|n) = \sum_\lambda p(D|\lambda, n)p(\lambda|n) \]
Another application of Bayes’ Theorem: “Model Selection”

- Complex models are more flexible but they spread their probability mass
- Bayesian inference inherently prefers “simpler models” – Occam’s razor
- Computational burden: We need to sum over all parameters $\lambda$
Probabilistic Inference

A huge spectrum of applications – all boil down to computation of

- **expectations** of functions under probability distributions: **Integration**

\[
\langle f(x) \rangle = \int_{\mathcal{X}} dx p(x) f(x) \quad \langle f(x) \rangle = \sum_{x \in \mathcal{X}} p(x) f(x)
\]

- **modes** of functions under probability distributions: **Optimization**

\[
x^* = \arg\max_{x \in \mathcal{X}} p(x) f(x)
\]

- any “mix” of the above: e.g.,

\[
x^* = \arg\max_{x \in \mathcal{X}} p(x) = \arg\max_{x \in \mathcal{X}} \int_{\mathcal{Z}} dz p(z) p(x|z)
\]
Divide and Conquer

Probabilistic modelling provides a methodology that puts a clear division between

• What to solve : Model Construction
  – Both an Art and Science
  – Highly domain specific

• How to solve : Inference Algorithm
  – Mechanical (In theory! not in practice)
  – Generic
Exercise

\[
p(x_1, x_2) & | \quad x_2 = 1 & \quad x_2 = 2 \\
\hline
x_1 = 1 & 0.3 & 0.3 \\
x_1 = 2 & 0.1 & 0.3 \\
\hline
\]

1. Find the following quantities

- **Marginals:** \( p(x_1), p(x_2) \)
- **Conditionals:** \( p(x_1|x_2), p(x_2|x_1) \)
- **Posterior:** \( p(x_1, x_2 = 2), p(x_1|x_2 = 2) \)
- **Evidence:** \( p(x_2 = 2) \)
- **\( p(\{\}) \)**
- **Max:** \( p(x_1^*) = \max_{x_1} p(x_1|x_2 = 1) \)
- **Mode:** \( x_1^* = \arg \max_{x_1} p(x_1|x_2 = 1) \)
- **Max-marginal:** \( \max_{x_1} p(x_1, x_2) \)

2. Are \( x_1 \) and \( x_2 \) independent? (i.e., Is \( p(x_1, x_2) = p(x_1)p(x_2) \)?)
## Answers

<table>
<thead>
<tr>
<th>$p(x_1, x_2)$</th>
<th>$x_2 = 1$</th>
<th>$x_2 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 = 1$</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>$x_1 = 2$</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

- **Marginals:**

<table>
<thead>
<tr>
<th>$p(x_1)$</th>
<th>$x_1 = 1$</th>
<th>$x_1 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x_2)$</td>
<td>$x_2 = 1$</td>
<td>$x_2 = 2$</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

- **Conditionals:**

| $p(x_1 | x_2)$ | $x_2 = 1$ | $x_2 = 2$ |
|--------------|-----------|-----------|
| $p(x_2 | x_1)$  | $x_2 = 1$ | $x_2 = 2$ |
| $x_1 = 1$    | 0.75      | 0.5       |
| $x_1 = 2$    | 0.25      | 0.5       |
| $x_1 = 1$    | 0.5       | 0.5       |
| $x_1 = 2$    | 0.25      | 0.75      |
Answers

\[
p(x_1, x_2) \quad x_2 = 1 \quad x_2 = 2 \\
x_1 = 1 \quad 0.3 \quad 0.3 \\
x_1 = 2 \quad 0.1 \quad 0.3
\]

• Posterior:

\[
p(x_1, x_2 = 2) \quad x_2 = 2 \\
x_1 = 1 \quad 0.3 \\
x_1 = 2 \quad 0.3
\]

\[
p(x_1|x_2 = 2) \quad x_2 = 2 \\
x_1 = 1 \quad 0.5 \\
x_1 = 2 \quad 0.5
\]

• Evidence:

\[
p(x_2 = 2) = \sum_{x_1} p(x_1, x_2 = 2) = 0.6
\]

• Normalisation constant:

\[
p(\{\}) = \sum_{x_1} \sum_{x_2} p(x_1, x_2) = 1
\]
Answers

<table>
<thead>
<tr>
<th>$p(x_1, x_2)$</th>
<th>$x_2 = 1$</th>
<th>$x_2 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1 = 1$</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>$x_1 = 2$</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

- Max: (get the value)
  \[ \max_{x_1} p(x_1|x_2 = 1) = 0.75 \]

- Mode: (get the index)
  \[ \arg\max_{x_1} p(x_1|x_2 = 1) = 1 \]

- Max-marginal: (get the “skyline”) \( \max_{x_1} p(x_1, x_2) \)

<table>
<thead>
<tr>
<th>$\max_{x_1} p(x_1, x_2)$</th>
<th>$x_2 = 1$</th>
<th>$x_2 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Exercise: Continuous Random variables

- Evaluate $p(c), p(x = 0)$ and $p(c|x = 10)$
Exercise: Continuous Random variables

- $x$, a conditionally Gaussian Random variable and $c$ is discrete

\[
p(x|c) = \mathcal{N}(x; \mu(c), \nu(c)) \equiv \frac{1}{\sqrt{2\pi\nu(c)}} \exp\left(-\frac{1}{2} \frac{(x - \mu(c))^2}{\nu(c)}\right)
\]

- In this example we take

\[
p(x, c = 1) = 0.6\mathcal{N}(x; 0, 0.01) \quad p(x, c = 2) = 0.4\mathcal{N}(x; 0.2, 0.03)
\]
Solutions

• \(p(x = 0)\) is a number, that we calculate here with Matlab (or octave)

\[
\begin{align*}
\text{>> } & x = 0; \text{ mu} = [0, 0.2]; \text{ v} = [0.01 0.03]; \text{ pc} = [0.6 0.4]; \\
\text{>> } & \text{pc.*(2*pi*v).}^{(-1/2)}.*\text{exp}(-0.5*(x-mu).^{2}./v) \\
\text{ans} & = 2.3937 \quad 0.4730 \\
\text{>> } & \text{sum(ans)} \\
\text{ans} & = 2.8667
\end{align*}
\]

• Note: This works here adding exp’s can be numerically instable

• How come that the “probability” is larger than one ?
  – \(p(x)\) is a density. The probability is \(p(x)dx\)
Solution

\[ p(x) \]

\[ p(x, c=1) \]

\[ p(x, c=2) \]
Solutions

- $p(c|x = 10)$ is a distribution,

$$p(c|x = 10) = \frac{p(c, x = 10)}{p(x = 10)}$$

- We calculate here with Matlab (or octave)

```matlab
>> x = 10; mu = [0, 0.2]; v = [0.01 0.03]; pc = [0.6 0.4];
>> pcx = pc.*(2*pi*v).^(-1/2).*exp(-0.5*(x-mu).^2./v)
pcx =
    0     0
>> pcx/sum(pcx)
Warning: Divide by zero.
ans =
    NaN   NaN
```

- **Problem: Underflow.** We ALWAYS work with log-densities in practice
Solutions
Solutions

• The log-density is

\[ \log p(x|c) = \log \mathcal{N}(x; \mu(c), \nu(c)) \equiv -\frac{1}{2} \log(2\pi \nu(c)) - \frac{1}{2} \frac{(x - \mu(c))^2}{\nu(c)} \]

\[
\begin{align*}
> & \text{lpc} = \log([0.6 \ 0.4]); \\
> & \text{lpcx} = \text{lpc} - 1/2*\log(2*\text{pi}^*\nu) -0.5*(x-\text{mu}).^2./\nu \\
& \quad 1.0e+003 * \\
& -4.9991 \quad -1.6007 \\
> & \text{lpcx} - \log(\sum(\exp(\text{lpcx}))) \\
\text{Warning: Log of zero.} \\
> & \text{ans} = \\
& \quad \text{Inf} \quad \text{Inf}
\end{align*}
\]

• Problem still persists. (we exp very small numbers)
Numerically Stable computation of $\log(\sum_i \exp(l_i))$

- Derivation

\[
L = \log(\sum_i \exp(l_i))
\]
\[
= \log(\sum_i \exp(l_i) \frac{\exp(l^*)}{\exp(l^*)})
\]
\[
= \log(\exp(l^*) \sum_i \exp(l_i - l^*))
\]
\[
= l^* + \log(\sum_i \exp(l_i - l^*))
\]

- We take $l^*$ as the maximum $l^* = \max_i l_i$

- Exercise: Implement above as a function $\log\text{sumexp}(l)$
Solutions

- \[
\begin{align*}
\text{lpc} &= \log([0.6 \ 0.4]); \\
\text{lpcx} &= \text{lpc} - \frac{1}{2} \log(2\pi\nu) - 0.5\cdot(x-\mu)^2/\nu \\
\text{ans} &= \exp(\text{ans})
\end{align*}
\]
Probability Models

+  

Inference Algorithms

=  

Bayesian Numerical Methods

Applications of Probability Models

• Classification

• Optimal Decision, given a loss function

• Finding interesting (hidden) structure
  – Clustering, Segmentation
  – Dimensionality Reduction
  – Outlier Detection

• Finding a compact representation = Data Compression

• Prediction
Graphical Models

• formal languages for specification of probability models and associated inference algorithms

• historically, introduced in probabilistic expert systems (Pearl 1988) as a visual guide for representing expert knowledge

• today, a standard tool in machine learning, statistics and signal processing
Graphical Models

• provide graph based algorithms for derivations and computation

• pedagogical insight/motivation for model/algorithm construction
  – Statistics:
    “Kalman filter models and hidden Markov models (HMM) are equivalent up to parametrisation”
  – Signal processing:
    “Fast Fourier transform is an instance of sum-product algorithm on a factor graph”
  – Computer Science:
    “Backtracking in Prolog is equivalent to inference in Bayesian networks with deterministic tables”

• Automated tools for code generation start to emerge, making the design/implement/test cycle shorter
Important types of Graphical Models

• Useful for Model Construction
  – Directed Acyclic Graphs (DAG), Bayesian Networks
  – Undirected Graphs, Markov Networks, Random Fields
  – Influence diagrams
  – ...

• Useful for Inference
  – Factor Graphs
  – Junction/Clique graphs
  – Region graphs
  – ...
Directed Graphical models (DAG)
DAG Example: Two dice

\[ p(\lambda) \quad p(y) \]

\[ \begin{array}{c}
\lambda \\
\hline
\end{array} \quad \begin{array}{c}
y \\
\hline
\end{array} \quad \begin{array}{c}
\mathcal{D} \\
\hline
\end{array} \]

\[ p(\mathcal{D}|\lambda, y) \]

\[ p(\mathcal{D}, \lambda, y) = p(\mathcal{D}|\lambda, y)p(\lambda)p(y) \]
DAG with observations

\[ p(\lambda) \quad p(y) \]

\[ \lambda \quad y \]

\[ D \]

\[ p(D = 9|\lambda, y) \]

\[ \phi_D(\lambda, y) = p(D = 9|\lambda, y)p(\lambda)p(y) \]
Directed Graphical models

• Each random variable is associated with a node in the graph,

• We draw an arrow from $A \rightarrow B$ if $p(B|\ldots,A,\ldots) (A \in \text{parent}(B))$,

• The edges tell us qualitatively about the factorization of the joint probability

• For $N$ random variables $x_1, \ldots, x_N$, the distribution admits

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | \text{parent}(x_i))$$

• Describes in a compact way an algorithm to “generate” the data – “Generative models”
### Examples

<table>
<thead>
<tr>
<th>Model</th>
<th>Structure</th>
<th>factorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td><img src="image" alt="Full Graph" /></td>
<td>( p(x_1)p(x_2</td>
</tr>
<tr>
<td>Markov(2)</td>
<td><img src="image" alt="Markov(2) Graph" /></td>
<td>( p(x_1)p(x_2</td>
</tr>
<tr>
<td>Markov(1)</td>
<td><img src="image" alt="Markov(1) Graph" /></td>
<td>( p(x_1)p(x_2</td>
</tr>
<tr>
<td>Factorized</td>
<td><img src="image" alt="Factorized Graph" /></td>
<td>( p(x_1)p(x_2</td>
</tr>
</tbody>
</table>

Removing edges eliminates a term from the conditional probability factors.
Undirected Graphical Models
Undirected Graphical Models

- Define a distribution by non-negative local compatibility functions $\phi(x_\alpha)$

$$p(x) = \frac{1}{Z} \prod_{\alpha} \phi(x_\alpha)$$

where $\alpha$ runs over cliques: fully connected subsets

- Examples

$$p(x) = \frac{1}{Z} \phi(x_1, x_2) \phi(x_1, x_3) \phi(x_2, x_4) \phi(x_3, x_4)$$

$$p(x) = \frac{1}{Z} \phi(x_1, x_2, x_3) \phi(x_2, x_3, x_4)$$
Possible Model Topologies

Diagram showing various network topologies.
Factor graphs
Factor graphs \([2]\)

- A bipartite graph. A powerful graphical representation of the inference problem
  - **Factor nodes**: Black squares. Factor potentials (local functions) defining the posterior.
  - **Variable nodes**: White Nodes. Define collections of random variables
  - **Edges**: denote membership. A variable node is connected to a factor node if a member variable is an argument of the local function.

\[
p(D = 9 | \lambda, y) = \phi_D(\lambda, y) = p(D = 9 | \lambda, y)p(\lambda)p(y) = \phi_1(\lambda, y)\phi_2(\lambda)\phi_3(y)
\]
Exercise

- For the following Graphical models, write down the factors of the joint distribution and plot an equivalent factor graph and an undirected graph.
Answer (Markov(1))

\[ p(x_1) \quad p(x_2|x_1) \quad p(x_3|x_2) \quad p(x_4|x_3) \]

\[ \phi(x_1, x_2) \quad \phi(x_2, x_3) \quad \phi(x_3, x_4) \]
Answer (IFA – Factorial)

\[
p(h_1)p(h_2) \prod_{i=1}^{4} p(x_i|h_1, h_2)
\]
Answer (IFA – Factorial)

- We can also cluster nodes together
Inference and Learning

- Data set
  \[ D = \{ x_1, \ldots, x_N \} \]

- Model with parameter \( \lambda \)
  \[ p(D|\lambda) \]

- Maximum Likelihood (ML)
  \[ \lambda^{\text{ML}} = \arg \max_\lambda \log p(D|\lambda) \]

- Predictive distribution
  \[ p(x_{N+1}|D) \approx p(x_{N+1}|\lambda^{\text{ML}}) \]
Regularisation

- Prior
  
  \[ p(\lambda) \]

- Maximum a-posteriori (MAP) : Regularised Maximum Likelihood

  \[ \lambda^{\text{MAP}} = \arg \max_{\lambda} \log p(D|\lambda)p(\lambda) \]

- Predictive distribution

  \[ p(x_{N+1}|D) \approx p(x_{N+1}|\lambda^{\text{MAP}}) \]
Bayesian Learning

- We treat parameters on the same footing as all other variables
- We integrate over unknown parameters rather than using point estimates (remember the many-dice example)
  - Self-regularisation, avoids overfitting
  - Natural setup for online adaptation
  - Model selection
Bayesian Learning

• Predictive distribution

\[ p(x_{N+1} | \mathcal{D}) = \int d\lambda \ p(x_{N+1} | \lambda) p(\lambda | \mathcal{D}) \]

\[ \begin{array}{c}
  x_1 \\
  x_2 \\
  \ldots \\
  x_N \\
  x_{N+1}
\end{array} \]

• Bayesian learning is just inference ...
Example Applications and Models
Audio Restoration

- During download or transmission, some samples of audio are lost
- Estimate missing samples given clean ones
Examples: Audio Restoration

\[ p(x_{\neg \kappa} | x_{\kappa}) \propto \int d\mathcal{H} p(x_{\neg \kappa} | \mathcal{H}) p(x_{\kappa} | \mathcal{H}) p(\mathcal{H}) \]

\[ \mathcal{H} \equiv \text{(parameters, hidden states)} \]

\[ H \equiv \text{parameters, hidden states} \]

\[ x_{\neg \kappa} \quad \text{Missing} \quad x_{\kappa} \quad \text{Observed} \]
Restoration

(Cemgil and Godsill 2005 [?])

- Piano
  - $S \rightarrow R \rightarrow \hat{S}$
  - $R \rightarrow \hat{S}$

- Trumpet
  - $S \rightarrow R \rightarrow \hat{S}$
  - $R \rightarrow \hat{S}$
Interpolation of Images
Interpolation of Images

Data (25% Missing)  Variational Bayes+ICM NMF  ML NMF2
Medical Expert Systems
Medical Expert Systems

Visit to Asia?  Smoking?  Tuberculosis?  Lung Cancer?  Bronchitis?
Either T or L?
Positive X Ray?  Dyspnoea?
Medical Expert Systems

Visit to Asia?
0 99% 1 1%

Smoking?
0 50% 1 50%

Tuberculosis?
0 99% 1 1%

Lung Cancer?
0 5% 1 94.5%

Bronchitis?
0 55% 1 45%

Either T or L?
0 35% 1 6.5%

Positive X Ray?
0 89% 1 11%

Dyspnoea?
0 56.4% 1 43.6%
Medical Expert Systems

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0%98.7</td>
<td>0%31.2</td>
<td>0%90.8</td>
<td>0%51.1</td>
<td>0%49.4</td>
<td>0%42.4</td>
<td>0%0</td>
<td>0%35.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1%1.3</td>
<td>1%68.8</td>
<td>1%9.2</td>
<td>1%48.9</td>
<td>1%50.6</td>
<td>1%57.6</td>
<td>1%100</td>
<td>1%64.1</td>
</tr>
</tbody>
</table>
Medical Expert Systems

- **Visit to Asia?**
  - 0: 98.5%
  - 1: 1.5%

- **Smoking?**
  - 0: 100%
  - 1: 0%

- **Tuberculosis?**
  - 0: 14.8%
  - 1: 85.2%

- **Lung Cancer?**
  - 0: 14.2%
  - 1: 85.8%

- **Bronchitis?**
  - 0: 70%
  - 1: 30%

- **Either T or L?**
  - 0: 28.9%
  - 1: 71.1%

- **Positive X Ray?**
  - 0: 100%
  - 1: 0%

- **Dyspnoea?**
  - 0: 56%
  - 1: 44%
Model Selection: Variable selection in Polynomial Regression

- Given $\mathcal{D} = \{t_j, x(t_j)\}_{j=1...J}$, what is the order $N$ of the polynomial?

$$x(t) = s_1 + s_2 t + s_3 t^2 + s_4 t^3 + \cdots + \epsilon(t)$$
Bayesian Variable Selection

- Generalized Linear Model – Column’s of $C$ are the basis vectors
- The exact posterior is a mixture of $2^W$ Gaussians
- When $W$ is large, computation of posterior features becomes intractable.
Regression

$p(x, r_1; W)$

All on Configurations All off
Clustering
Clustering

\[
(\mu_a^*, \mu_b^*, \pi^*) = \underset{\mu_a, \mu_b, \pi}{\text{argmax}} \sum_{c_1:N} \prod_{i=1}^{N} p(x_i | \mu_a, \mu_b, c_i) p(c_i | \pi)
\]
Computer vision / Cognitive Science

How many rectangles are there in this image?
Computer Vision

How many people are there in these images?
Visual Tracking

- Norm of Transverse Plane
- Norm of Coronal Plane
- Norm of Sagittal Plane
Navigation, Robotics : Sensor Fusion
Navigation, Robotics : Sensor Fusion

GPS status

GPS reading

Other sensors (magnetic, pressure, e.t.c.)

Linear accelerometer sensor

Gyroscope

Attitude Variables

Linear Kinematic Variables

Set of feature points (Camera Frame)

Set of feature points (World Coordinates)

Global Static Map (Intensity function)
Computer Accompaniment

(Music Plus One, Raphael 2000 [?], Dannenberg and Raphael 2006)
References

