Introduction to Probabilistic Graphical Models
Lecture 1

Course structure & Introduction

Télécom ParisTech,

Université Paris-Saclay, Paris, France

Instructor: Umut Şimşekli

Goals of this Course

- Provide a basic understanding of underlying principles of probabilistic modeling and inference

- Focus on fundamental concepts rather than technical details

. . . we avoid heavy use of algebra by a graphical notation
. . . but there will be some maths
  - Calculus
  - Linear Algebra
  - Probability Theory

- Model based approach

- Getting prepared for more advanced courses

The Topics to be Covered

- Probability background
- Conditional independence, Directed and undirected graphical models
- Inference/learning concepts, example applications
- Exponential family distributions
- Gaussian Mixture Models and the Expectation-Maximization algorithm
- Hidden Markov Models

Possible Applications

- Development of a probabilistic model in one application area (including but not limited to)
  - Computer Vision (Object tracking)
  - Robotics, Navigation, Self Localisation
  - Signal, Speech, Audio, Music Processing
  - Information Retrieval, Data mining, Text processing, Natural Language Processing
  - Scientific data analysis (DNA, Bioinformatics, Medicine, Seismology)
  - Sports, Finance, User Behaviour, Cognitive Science e.t.c.
- Reading a paper and writing a tutorial-like summary in own words and self designed examples
- Implementation and comparative study of inference algorithms on synthetic data

Reference Textbooks

- Handouts and Slides
- Pattern Recognition and Machine Learning,
  Christopher Bishop, Springer
  `http://research.microsoft.com/~cmbishop/PRML/index.htm`
- Information Theory, Inference, and Learning Algorithms
  David MacKay, Cambridge University Press – fourth printing (March 2005)
  `http://www.inference.phy.cam.ac.uk/mackay/itprnn/book.html`
- Machine Learning, A Probabilistic Approach,
  David Barber, Cambridge University Press
  `http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook`

Course Structure

- Weekly lectures on **Wednesdays** at **14:30**, at Télécom ParisTech
  - Lecture 1: (this one) 20/09/2017, 14:30, Télécom ParisTech, Room B310-B311
  - Lecture 2: 27/09/2017, 14:30, Télécom ParisTech, Room B214-B215-2 (Amphi meraude)
  - Lecture 3: 04/09/2017, 14:30, Télécom ParisTech, Room B310-B311
  - Lecture 4: 11/10/2017, 14:30, Télécom ParisTech, Room B310-B311
  - Lecture 5: 18/10/2017, 14:30, Télécom ParisTech, Room B310-B311
  - Lecture 6: 08/11/2017, 14:30, Télécom ParisTech, Room B310-B311
  - Lecture 6: 15/11/2017, 14:30, Télécom ParisTech, Room B312-313
- **Check the classroom before the lectures!**

Course Structure

Evaluation

- 2 Homeworks (Programming, Analytic Derivations): short and simple
- 1 Miniproject
- 1 Final Exam
  - Date, time, place will be announced

Course Structure

- Grading
  - % 20 Homeworks (there will be 2)
  - % 40 Miniproject – mostly programming
  - % 40 Final

Disclaimer

- All the material that will be used within this course is adapted from the "Bayesian Statistics and Machine Learning" course that has been given by A. Taylan Cemgil at Boğaziçi University, Istanbul

- For more info, please see `http://www.cmpe.boun.edu.tr/~cemgil/`

Introduction

Lecture Outline

- Introduction
  - Bayes' Theorem,
  - Trivial toy example to clarify notation
- Probability tables

Bayes' Theorem

Thomas Bayes (1702-1761)

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

$$
\begin{aligned}
p(\lambda|\mathcal{D}) &= \frac{p(\mathcal{D}|\lambda)p(\lambda)}{p(\mathcal{D})} \\
\text{Posterior} &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}
\end{aligned}
$$

An application of Bayes' Theorem: "Source Separation"

Given two fair dice with outcomes λ and y,

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

What is λ when $\mathcal{D} = 9$?

An application of Bayes' Theorem: "Source Separation"

$$\mathcal{D} = \lambda + y = 9$$

$\mathcal{D} = \lambda + y$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$\lambda = 1$	2	3	4	5	6	7
$\lambda = 2$	3	4	5	6	7	8
$\lambda = \mathbf{3}$	4	5	6	7	8	**9**
$\lambda = \mathbf{4}$	5	6	7	8	**9**	10
$\lambda = \mathbf{5}$	6	7	8	**9**	10	11
$\lambda = \mathbf{6}$	7	8	**9**	10	11	12

Bayes theorem "upgrades" $p(\lambda)$ into $p(\lambda|\mathcal{D})$.

But you have to provide an observation model: $p(\mathcal{D}|\lambda)$

“Bureaucratical” derivation

Formally we write

$$
\begin{aligned}
p(\lambda) &= \mathcal{C}(\lambda; [\ 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6\ ]) \\
p(y) &= \mathcal{C}(y; [\ 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6 \quad 1/6\ ]) \\
p(\mathcal{D}|\lambda, y) &= \delta(\mathcal{D} - (\lambda + y))
\end{aligned}
$$

$$
\begin{aligned}
p(\lambda, y|\mathcal{D}) &= \frac{1}{p(\mathcal{D})} \times p(\mathcal{D}|\lambda, y) \times p(y)p(\lambda) \\
\text{Posterior} &= \frac{1}{\text{Evidence}} \times \text{Likelihood} \times \text{Prior}
\end{aligned}
$$

Kronecker delta function denoting a degenerate (deterministic) distribution $\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}$

Prior

$$p(y)p(\lambda)$$

$p(y) \times p(\lambda)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$\lambda = 1$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$\lambda = 2$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$\lambda = 3$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$\lambda = 4$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$\lambda = 5$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$
$\lambda = 6$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$	$1/36$

- A table with indicies λ and y
- Each cell denotes the probability $p(\lambda, y)$

Likelihood

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

$p(\mathcal{D} = 9\|\lambda, y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$\lambda = 1$	0	0	0	0	0	0
$\lambda = 2$	0	0	0	0	0	0
$\lambda = 3$	0	0	0	0	0	**1**
$\lambda = 4$	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 indicies λ and y
- The likelihood is **not** a probability distribution, but a positive function.

Likelihood $\times$ Prior

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

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

Evidence (= Marginal Likelihood)

$$
\begin{aligned}
p(\mathcal{D}=9) &= \sum_{\lambda,y} p(\mathcal{D}=9|\lambda,y)p(\lambda)p(y) \\
&= 0+0+\cdots+1/36+1/36+1/36+1/36+0+\cdots+0 \\
&= 1/9
\end{aligned}
$$

$p(\mathcal{D}=9\vert\lambda,y)$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$	$y=6$
$\lambda=1$	0	0	0	0	0	0
$\lambda=2$	0	0	0	0	0	0
$\lambda=3$	0	0	0	0	0	$\mathbf{1/36}$
$\lambda=4$	0	0	0	0	$\mathbf{1/36}$	0
$\lambda=5$	0	0	0	$\mathbf{1/36}$	0	0
$\lambda=6$	0	0	$\mathbf{1/36}$	0	0	0

Posterior

$$p(\lambda, y|\mathcal{D} = 9) = \frac{1}{p(\mathcal{D})} p(\mathcal{D} = 9|\lambda, y) p(\lambda) p(y)$$

$p(\mathcal{D} = 9\|\lambda, y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$
$\lambda = 1$	0	0	0	0	0	0
$\lambda = 2$	0	0	0	0	0	0
$\lambda = 3$	0	0	0	0	0	$\mathbf{1/4}$
$\lambda = 4$	0	0	0	0	$\mathbf{1/4}$	0
$\lambda = 5$	0	0	0	$\mathbf{1/4}$	0	0
$\lambda = 6$	0	0	$\mathbf{1/4}$	0	0	0

$$1/4 = (1/36)/(1/9)$$

Marginal Posterior

$$p(\lambda|\mathcal{D}) = \sum_y \frac{1}{p(\mathcal{D})} p(\mathcal{D}|\lambda, y)p(\lambda)p(y)$$

	$p(\lambda\|\mathcal{D}=9)$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$	$y=6$
$\lambda=1$	0	0	0	0	0	0	0
$\lambda=2$	0	0	0	0	0	0	0
$\lambda=3$	$\mathbf{1/4}$	0	0	0	0	0	$1/4$
$\lambda=4$	$\mathbf{1/4}$	0	0	0	0	$1/4$	0
$\lambda=5$	$\mathbf{1/4}$	0	0	0	$1/4$	0	0
$\lambda=6$	$\mathbf{1/4}$	0	0	$1/4$	0	0	0

The "proportional to" notation

$$p(\lambda|\mathcal{D}=9) \quad \propto \quad p(\lambda, \mathcal{D}=9) = \sum_y p(\mathcal{D}=9|\lambda, y)p(\lambda)p(y)$$

	$p(\lambda, \mathcal{D}=9)$	$y=1$	$y=2$	$y=3$	$y=4$	$y=5$	$y=6$
$\lambda=1$	0	0	0	0	0	0	0
$\lambda=2$	0	0	0	0	0	0	0
$\lambda=3$	$1/36$	0	0	0	0	0	$\mathbf{1/36}$
$\lambda=4$	$1/36$	0	0	0	0	$\mathbf{1/36}$	0
$\lambda=5$	$1/36$	0	0	0	$\mathbf{1/36}$	0	0
$\lambda=6$	$1/36$	0	0	$\mathbf{1/36}$	0	0	0

Another application of Bayes' Theorem: "Model Selection"

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

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

How many dice are there when $\mathcal{D} = 9$?

Assume that any number n is equally likely *a-priori*

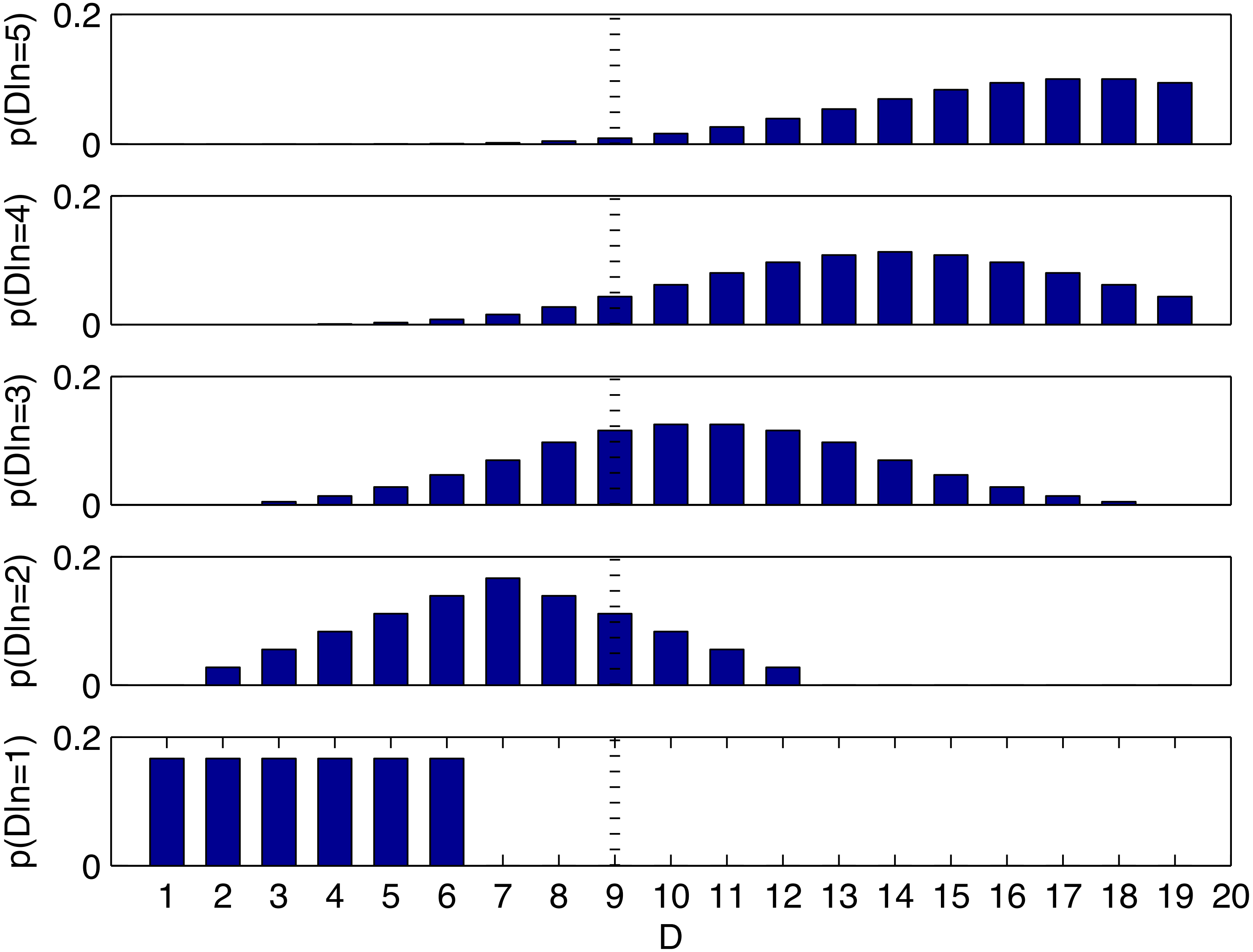
Another application of Bayes' Theorem: "Model Selection"

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

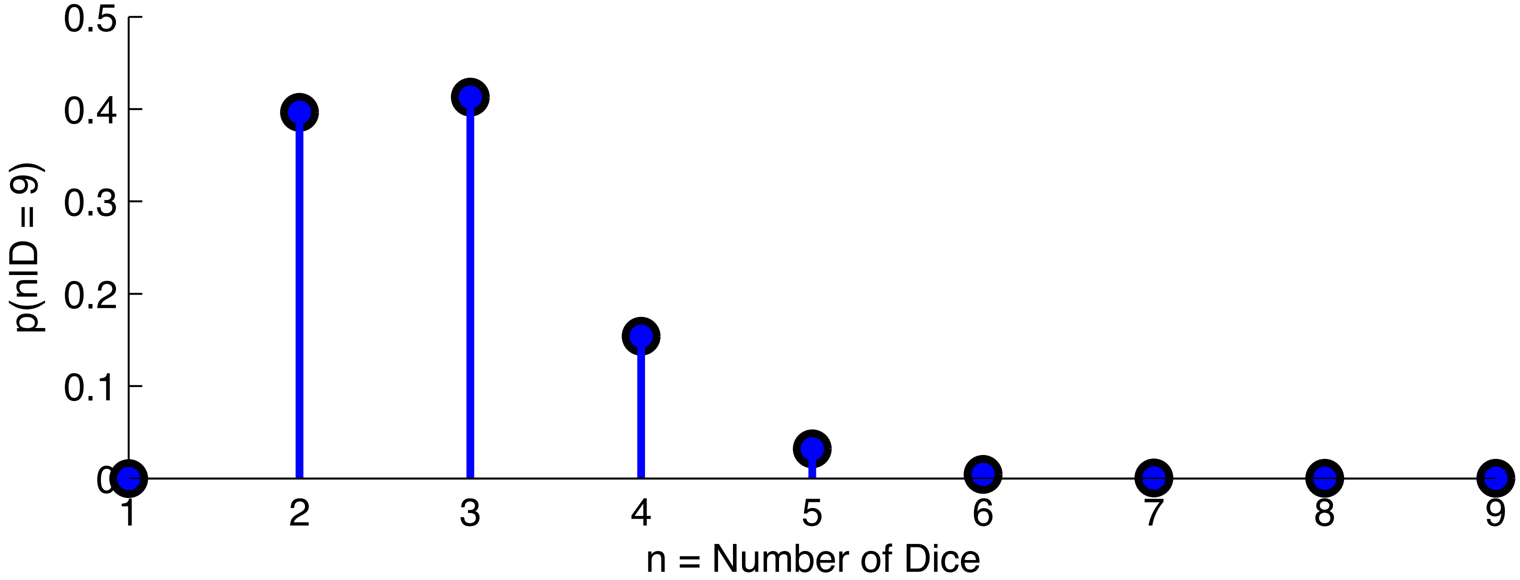
$$p(n|\mathcal{D}=9) = \frac{p(\mathcal{D}=9|n)p(n)}{p(\mathcal{D})} \propto p(\mathcal{D}=9|n)$$

$$\begin{aligned}
p(\mathcal{D}|n=1) &= \sum_{\lambda_1} p(\mathcal{D}|\lambda_1)p(\lambda_1) \\
p(\mathcal{D}|n=2) &= \sum_{\lambda_1}\sum_{\lambda_2} p(\mathcal{D}|\lambda_1,\lambda_2)p(\lambda_1)p(\lambda_2) \\
&\dots \\
p(\mathcal{D}|n=n') &= \sum_{\lambda_1,\dots,\lambda_{n'}} p(\mathcal{D}|\lambda_1,\dots,\lambda_{n'})\prod_{i=1}^{n'} p(\lambda_i)
\end{aligned}$$

$$p(\mathcal{D}|n) = \sum_{\boldsymbol{\lambda}} p(\mathcal{D}|\boldsymbol{\lambda}, n)p(\boldsymbol{\lambda}|n)$$



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 λ

Probabilistic Inference

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

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

$$\langle f(x) \rangle = \int_{\mathcal{X}} dx p(x) f(x) \qquad \langle f(x) \rangle = \sum_{x \in \mathcal{X}} p(x) f(x)$$

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

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} p(x) f(x)$$

- any "mix" of the above: e.g.,

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} p(x) = \operatorname*{argmax}_{x \in \mathcal{X}} \int_{\mathcal{Z}} dz p(z) p(x|z)$$

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

Bayes Theorem Repeated

$$
\begin{aligned}
p(B|A) &= \frac{p(A|B) \times p(B)}{\sum_B p(A|B)p(B)} \\
\text{Posterior} &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}
\end{aligned}
$$

- Think of A as an observation and B as its hidden cause.

- Bayes theorem says how to update our prior belief $p(B)$ given a new observation A. This gives a way of "reversing" the conditional probability $p(A|B)$.

Bayes Theorem Repeated

- This rather simple looking formula has surprisingly many applications
  - Medical Diagnosis (Symptoms/Diseases)
  - Speech Recognition (Signal/Phoneme)
  - Music Transcription (Audio/Score)
  - Computer Vision (Image/Object)
  - Robotics (Sensor/Position)
  - Finance (Past Price/Future Price)

- A natural way of combining prior knowledge with data $\Rightarrow$ Learning

Exercise

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.3	0.3
$x_1 = 2$	0.1	0.3

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

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.3	0.3
$x_1 = 2$	0.1	0.3

- Marginals:

$p(x_1)$	
$x_1 = 1$	0.6
$x_1 = 2$	0.4

$p(x_2)$	$x_2 = 1$	$x_2 = 2$
	0.4	0.6

- Conditionals:

$p(x_1 \vert x_2)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.75	0.5
$x_1 = 2$	0.25	0.5

$p(x_2 \vert x_1)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.5	0.5
$x_1 = 2$	0.25	0.75

Answers

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.3	0.3
$x_1 = 2$	0.1	0.3

- Posterior:

$p(x_1, x_2 = 2)$	$x_2 = 2$
$x_1 = 1$	0.3
$x_1 = 2$	0.3

$p(x_1 \vert x_2 = 2)$	$x_2 = 2$
$x_1 = 1$	0.5
$x_1 = 2$	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

$p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$
$x_1 = 1$	0.3	0.3
$x_1 = 2$	0.1	0.3

- Max: (get the value)

$$\max_{x_1} p(x_1|x_2 = 1) = 0.75$$

- Mode: (get the index)

$$\operatorname*{argmax}_{x_1} p(x_1|x_2 = 1) = 1$$

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

$\max_{x_1} p(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$
	0.3	0.3

Keywords Summary

Bayes Theorem

Likelihood

Prior

Posterior

Evidence, Marginal Likelihood