Probabilistic reasoning

Artificial intelligence (CK0031/CK0248)

Francesco Corona

Department of Computer Science
Federal University of Ceará, Fortaleza

Outline

Probability refresher

Probability refresher

A key concept in the field of artificial intelligence is that of **uncertainty**

- ⇝ Through noise on measurements
- ⇝ Through the finite size of data

Probability theory provides a consistent framework

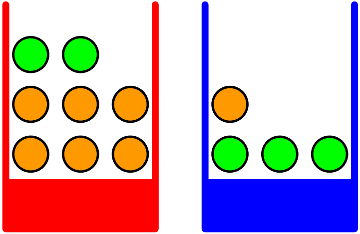
- Quantification and manipulation of uncertainty

Probability theory forms one of the central foundations of PRML

Probability refresher (cont.)

We have two boxes, one red and one blue

- In the red box, we have 2 apples and 6 oranges
- In the blue box, we have 3 apples and 1 orange



We randomly select one box

We randomly pick a fruit (from that box)

1. We check the fruit
2. We replace it in its box

We repeat the process *many* times

- 40% of the time, we pick the red box
- 60% of the time, we pick the blue box

We are equally likely to select any piece of fruit from the box

Probability refresher (cont.)

The **identity of the box** that will be chosen is a **random variable** B

- This random variable can take only two possible values
- Either `r`, for red box or `b`, for blue box

The **identity of the fruit** that will be chosen is a **random variable** F

- This random variable can take only two possible values
- Either `a`, for apple or `o`, for orange

Probability refresher (cont.)

Definition

*We define the **probability of an event** to be the fraction of times that some event occurs out of the total number of trials, in the limit that this number goes to infinity*

Example

- The probability of picking the red box is 4/10
- The probability of picking the blue box is 6/10

We write these probabilities

⇝ $p(B = \mathtt{r}) = 4/10$

⇝ $p(B = \mathtt{b}) = 6/10$

Probability refresher (cont.)

By definition, **probabilities must lie in the unit interval** $[0, 1]$

Consider the usual case in which events are **mutually exclusive**

Consider the case in which events **include all possible outcomes**

- The **probabilities** for such events **must sum to one**

Probability refresher (cont.)

Example

We have defined our experiment and we can start asking questions

- What is the probability that the selection procedure picks an apple?
- Given that we have picked an orange, what is the probability that the box we chose was the blue one?
- ...

Probability refresher (cont.)

We can answer questions such as these, and much more complex ones

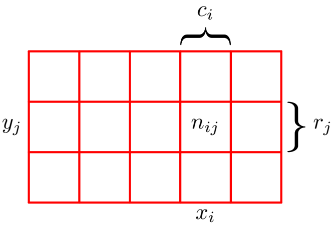
But, we need first the **two elementary rules of probability**

- ⇝ The **sum rule**
- ⇝ The **product rule**

To derive these rules, consider the slightly more general example

Probability refresher (cont.)

Suppose that we have **two random variables** X and Y



- X can take any of the values

  x_i, with $i = 1, \ldots, M$

- Y can take any of the values

  y_j, with $j = 1, \ldots, L$

Here, $M = 5$ and $L = 3$

Consider a **total of N trials** in which we sample both variable X and Y

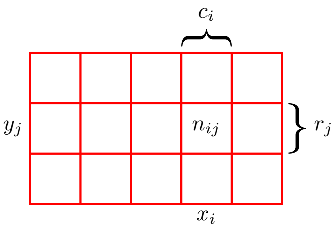
- Let n_{ij} be the number of such trials in which $X = x_i$ and $Y = y_j$
- Let c_i be the number of trials in which X takes the value x_i (irrespective of the value that Y takes)
- Let r_j be the number of trials in which Y takes the value y_j (irrespective of the value that X takes)

Probability refresher (cont.)

The probability that X will take the value x_i and Y will take the value y_j

$$p(X = x_i, Y = y_j)$$

This is the **joint probability** of $X = x_i$ and $Y = y_j$



The number of points falling in cell (i, j) as a fraction of the total number N of points

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \qquad (1)$$

Implicitly, in the limit $N \to \infty$

Probability refresher (cont.)

The probability that X takes the value x_i irrespective of the value of Y

$$p(X = x_i)$$

This is the fraction of the total number of points in column i

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{L} n_{ij}}{N} = \sum_{j=1}^{L} \underbrace{\frac{n_{ij}}{N}}_{p(X=x_i, Y=y_j)} = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \quad (2)$$

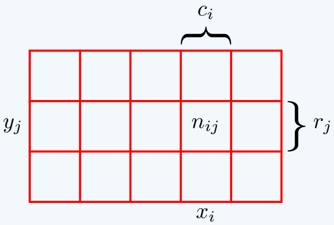
$p(X = x_i)$ is called the **marginal probability**

Obtained by marginalising, or summing out, the other variables (Y here)

Probability refresher (cont.)

The marginal probability sets us for the **sum rule** of probability

Definition



Sum rule

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{3}$$

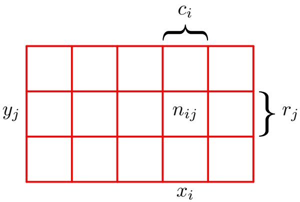
Probability refresher (cont.)

Consider only those instances for which $X = x_i$

Consider the fraction of such instances for which $Y = y_j$

$$p(Y = y_j | X = x_i)$$

This is the **conditional probability** of $Y = y_j$ given $X = x_i$



The fraction of points falling in column i that fall in cell (i, j)

$$p(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i} \qquad (4)$$

Probability refresher (cont.)

$$p(X = x_i, Y = y_j) = n_{ij}/N$$
$$p(Y = y_i | X = x_i) = n_{ij}/c_i$$
$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

From Equation (1), (2) and (4), we derive the **product rule** of probability

Definition

Product rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \underbrace{\frac{n_{ij}}{c_i}}_{p(Y=y_j|X=x_i)} \underbrace{\frac{c_i}{N}}_{p(X=x_i)} = p(Y = y_j | X = x_i) p(X = x_i) \quad (5)$$

Probability refresher (cont.)

Definition

The rules of probability

⇝ ***Sum rule***

$$p(X) = \sum_{Y} p(X, Y) \tag{6}$$

⇝ ***Product rule***

$$p(X, Y) = p(Y|X)p(X) \tag{7}$$

■

Probability refresher (cont.)

To compact notation, we use $p(\star)$ for a distribution[1] over some RV $\star$

- $p(X, Y)$ is a joint probability, the probability of X and Y
- $p(Y|X)$ is a conditional probability, the probability of Y given X
- $p(X)$ is a marginal probability, the probability of X

[1] $p(\star = \cdot)$ or simply $p(\cdot)$ denotes the distribution evaluated for the particular value $\cdot$

Probability refresher (cont.)

Definition

Consider the product rule and the symmetry property $p(X, Y) = p(Y, X)$

We obtain a relationship between conditional probabilities

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{8}$$

The relationship is called the ***Bayes' rule***

Probability refresher (cont.)

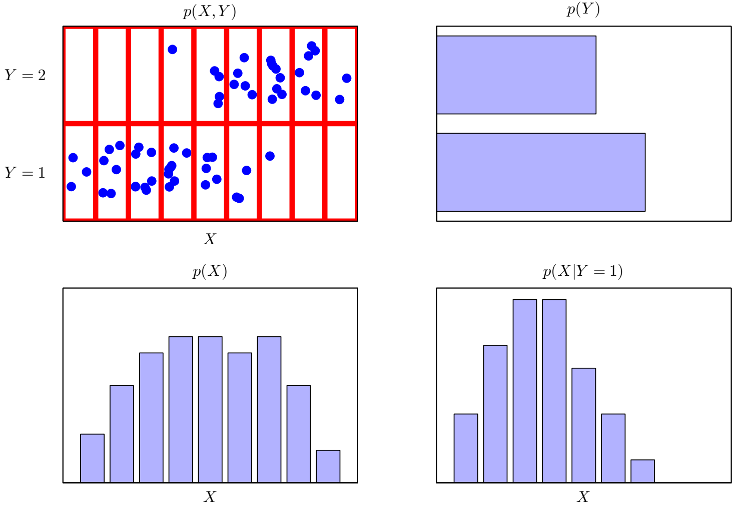
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Using the sum rule, the denominator in Bayes' theorem can be explicitated

$$p(X) = \sum_{Y} p(X|Y)p(Y) \tag{9}$$

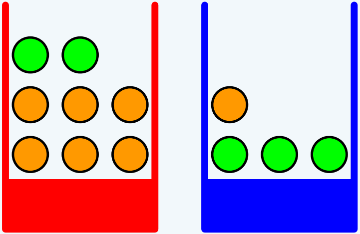
- Conditional probabilities $p(Y|X)$ over all values of Y must sum to one
- The denominator in terms of the quantities in the numerator
- The denominator is a normalisation constant

Probability theory (cont.)



Probability refresher (cont.)

Example



The probability of selecting either the red or the blue boxe

- $p(B = \texttt{r}) = 4/10$
- $p(B = \texttt{b}) = 6/10$

This satisfies $p(B = \texttt{r}) + p(B = \texttt{b}) = 4/10 + 6/10 = 1$

Probability refresher (cont.)

Suppose that we pick a box at random, say the blue box

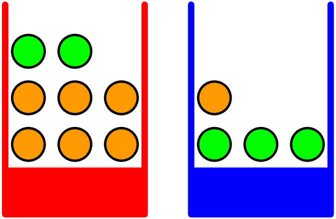
The probability of picking an apple is the fraction of apples in the blue box

$\rightsquigarrow$ 3/4

Thus, $p(F = \mathtt{a} | B = \mathtt{b}) = 3/4$

Probability refresher (cont.)

We write all conditional probabilities for the type of fruit, given the box



$$p(F=\mathtt{a}|B=\mathtt{r}) = 1/4 \quad (10)$$
$$p(F=\mathtt{o}|B=\mathtt{r}) = 3/4 \quad (11)$$
$$p(F=\mathtt{a}|B=\mathtt{b}) = 3/4 \quad (12)$$
$$p(F=\mathtt{o}|B=\mathtt{b}) = 1/4 \quad (13)$$

Note that these probabilities are normalised

$$p(F=\mathtt{a}|B=\mathtt{r}) + p(F=\mathtt{o}|B=\mathtt{r}) = 1 \quad (14)$$
$$p(F=\mathtt{a}|B=\mathtt{b}) + p(F=\mathtt{o}|B=\mathtt{b}) = 1 \quad (15)$$

Probability refresher (cont.)

Evaluate the overall probability of choosing an apple

We can use the sum and product rules of probability[2]

$$\begin{aligned} p(F = a) &= p(F = a | B = r) p(B = r) + p(F = a | B = b) p(B = b) \\ &= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \end{aligned} \tag{16}$$

It follows (sum rule) that $p(F = o) = 1 - 11/20 = 9/20$

[2] $P(X) = \sum_Y p(X, Y)$ with $p(X, Y) = p(Y | X) p(X) = p(Y, X) = p(X | Y) p(Y)$

Probability refresher (cont.)

Suppose that we are told that a piece of fruit has been selected

- Say, it is an orange

We would like to know which box it came from

The probability distribution over boxes conditioned on the fruit identity

$$P(B|F)$$

The probability distribution over fruits conditioned on the box identity

$$P(F|B)$$

- The probabilities in equations (10-13)

Probability refresher (cont.)

We need to reverse the conditional probability (Bayes' rule)

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3} \quad (17)$$

It follows (sum rule) that $p(B = b|F = o) = 1 - 2/3 = 1/3$

Probability refresher (cont.)

We can provide an important interpretation of Bayes' rule

$$p(B|F) = \frac{p(F|B)p(B)}{p(F)}$$

The most complete information about the box is initially probability $p(B)$

- It is the probability before we observe the identity of the fruit

⇝ We call this the **prior probability**

Once we are told that the fruit is an orange, it became probability $p(B|F)$

- It is the probability after we observe the identity of the fruit

⇝ We call this the **posterior probability**

Probability refresher (cont.)

$$\underbrace{p(B=r|F=o)}_{2/3} = \frac{p(F=o|B=r)}{p(F=o)}\underbrace{p(B=r)}_{4/10}$$

The prior probability of selecting the red box is 4/10

- (blue is more probable)

The posterior probability of the red box is 2/3

- (red is more probable)

Probability refresher (cont.)

Consider the joint distribution that factorises into the product of marginals

$$p(X, Y) = p(Y)p(X)$$

X and Y are said to be **independent**

$$p(X, Y) = p(Y|X)p(X)$$

The conditional distribution of Y given X is independent of the X value

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = P(Y) \qquad \Longleftarrow P(X|Y) = P(X)$$

Reasoning under uncertainty

Probabilistic modelling

Reasoning under uncertainty

Probabilistic modelling

Variables will be denoted using either upper case X or lower case x

Sets of variables will be typically denoted by calligraphic symbols

- For example, $\mathcal{V} = \{a, B, c\}$

The **domain of variable** x is $\mathrm{dom}(x)$, it denotes the **states** x can take

Probabilistic modelling (cont.)

Example

States will typically be represented using typewriter type fonts

- For a coin c, $\text{dom}(c) = \{\texttt{heads}, \texttt{tails}\}$
- $p(c = \texttt{heads})$ is the probability that variable c is in state $\texttt{heads}$

The meaning of $p(\texttt{state})$ is often clear, without reference to a variable

Suppose that we are discussing an experiment about a coin c

- ⇝ The meaning of $p(\texttt{heads})$ is clear from context
- ⇝ It is shorthand for $p(c = \texttt{heads})$

Probabilistic modelling (cont.)

When summing over a variable $\sum_x f(x)$ all states of x are included

- $\sum_x f(x) = \sum_{\mathsf{s} \in \mathrm{dom}(x)} f(x = \mathsf{s})$

Given variable x, its domain $\mathrm{dom}(x)$ and a full specification of probability values for each of the states, $p(x)$, we say we have a **distribution** for x

Probabilistic modelling (cont.)

For our purposes, **events** are expressions about random variables

⇝ *Two heads in 6 coin tosses*

Two events are **mutually exclusive** if they cannot both be true

⇝ Events *Coin is heads* and *Coin is tails* are mutually exclusive

Example

One can think of defining a new variable named by the event

⇝ $p(\textit{The coin is tails})$ can be interpreted as $p(\textit{The coin is tails} = \texttt{true})$

- $p(x = \texttt{tr})$, the probability of event/variable x being in state `true`
- $p(x = \texttt{fa})$, the probability of x being in state `false`

Probabilistic modelling (cont.)

Definition

Rules of probability for discrete variables (1)

Probability $p(x = \mathbf{x})$ *of variable* x *being in state* $\mathbf{x}$ *is represented by a value between* 0 *and* 1

- $p(x = \mathbf{x}) = 1$ *means that we are certain* x *is in state* $\mathbf{x}$
- $p(x = \mathbf{x}) = 0$ *means that we are certain* x *is NOT in state* $\mathbf{x}$

Values in $[0, 1]$ *represent the degree of certainty of state occupancy*

Probabilistic modelling (cont.)

Definition

Rules of probability for discrete variables (2)

The summation of the probability over all states is one

$$\sum_{x \in dom(x)} p(x = x) = 1 \tag{18}$$

Normalisation condition

$\rightsquigarrow$ $\sum_x p(x) = 1$

Probabilistic modelling (cont.)

Definition

Rules of probability for discrete variables (3)

x and y can interact

$$p(x = a \text{ or } y = b) = p(x = a) + p(y = b) - p(x = a \text{ and } y = b) \tag{19}$$

Or, more generally,

$$p(x \text{ or } y) = p(x) + p(y) - p(x \text{ and } y) \tag{20}$$

We use $p(x, y)$ for $p(x \text{ and } y)$

- $\rightsquigarrow$ $p(x, y) = p(y, x)$
- $\rightsquigarrow$ $p(x \text{ or } y) = p(y \text{ or } x)$

Probabilistic modelling (cont.)

Definition

Set notation

An alternative notation in terms of set theory

$$\begin{aligned} p(x \text{ or } y) &\equiv p(x \cup y) \\ p(x, y) &\equiv p(x \cap y) \end{aligned} \tag{21}$$

Probabilistic modelling (cont.)

Definition

Marginals

Given a ***joint distribution*** $p(x, y)$, *the distribution of a single variable*

$$p(x) = \sum_y p(x, y) \tag{22}$$

$p(x)$ *is termed a* ***marginal*** *of the joint probability distribution* $p(x, y)$

Marginalisation

Process of computing a marginal from a joint distribution

$$p(x_1, \dots, x_{i-1}, x_{x+1}, \dots, x_n) = \sum_{x_i} p(x_1, \dots, x_n) \tag{23}$$

Probabilistic modelling (cont.)

Definition

Conditional probability/Bayes' rule

The probability of some event x conditioned on knowing some event y

$\rightsquigarrow$ *The probability of x given y*

$$p(x|y) = \frac{p(x,y)}{p(y)} \tag{24}$$

IFF $p(y) = 0$, otherwise $p(x|y)$ is not defined

From this definition and $p(x,y) = p(y,x)$, we write the Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{25}$$

Probabilistic modelling (cont.)

Bayes' rule trivially follows from the definition of conditional probability

- Bayes' rule plays a central role in probabilistic reasoning
- It helps inverting probabilistic relations

$$p(y|x) \Leftrightarrow p(x|y)$$

Probabilistic modelling (cont.)

Remark

Subjective probability

Probability is a contentious topic and we do not debate it

- It is not the rules of probability that are contentious
- Rather what interpretation we should place on them

Suppose that potential repetitions of an experiment can be envisaged

- The frequentist definition of probability then makes sense
- (probabilities defined wrt a potentially infinite repetitions)

Probabilistic modelling (cont.)

In coin tossing, the interpretation of the probability of heads

- '*If I were to repeat the experiment of flipping a coin (at 'random'), the limit of the number of heads that occurred over the number of tosses is defined as the probability of a head occurring*'

Probabilistic modelling (cont.)

A typical problem and scenario in an AI situation

Example

A film enthusiast joins a new online film service

Based on a few films a user likes/dislikes, the company tries to estimate the probability that the user will like each of the $10K$ films in its offer

Suppose that we define probability as limiting case of infinite repetitions

- ⇝ This would not make much sense in this case
- ⇝ (we cannot repeat the experiment)

Suppose the user behaves in a manner that is consistent with other users

- We should be able to exploit the data from other users' ratings
- We can make a reasonable 'guess' as to what this consumer likes

Probabilistic modelling - Conditional probability

A **degree of belief** or **Bayesian** subjective interpretation of probability

- It sidesteps non-repeatability issues
- It is a framework for manipulating real values
- The framework is consistent with our intuition about probability

Probabilistic modelling - Conditional probability (cont.)

Conditional probability matches our intuition of uncertainty

Example

Imagine a circular dart board, split into 20 equal sections, labels $1 - 20$

- A dart thrower hits any one of the 20 sections uniformly at random

The probability that a dart occurs in any one of the 20 regions is simply

$$p(\texttt{region i}) = 1/20$$

Someone tells that the dart has not hit the 20-region

What is the probability that the dart thrower has hit the 5-region?

Probabilistic modelling - Conditional probability (cont.)

Conditioned on this information, only regions 1 to 19 remain possible

- There is no preference for the thrower to hit any of these regions
- The probability is 1/19

$$
\begin{aligned}
p(\texttt{region 5}|\texttt{not region 20}) &= \frac{p(\texttt{region 5}, \texttt{not region 20})}{p(\texttt{not region 20})} \\
&= \frac{p(\texttt{region 5})}{p(\texttt{not region 20})} \\
&= \frac{1/20}{19/20} = 1/19
\end{aligned}
$$

Probabilistic modelling - Conditional probability (cont.)

Remark

A not-fully-correct interpretation of $p(A = \mathtt{a}|B = \mathtt{b})$

- ‘*Given the event $B = \mathtt{b}$ has occurred, $p(A = \mathtt{a}|B = \mathtt{b})$ is the probability of the event $A = \mathtt{a}$ occurring*’

In most contexts, no explicit temporal causality can be implied

The correct interpretation

- ‘*$p(A = \mathtt{a}|B = \mathtt{b})$ is the probability of A being in state $\mathtt{a}$ under the constraint that B is in state $\mathtt{b}$*’

Probabilistic modelling - Conditional probability (cont.)

The relation between conditional distributionss and joint distributions

Between the conditional $p(A = \mathtt{a}|B = \mathtt{b})$ and the joint $p(A = \mathtt{a}, B = \mathtt{b})$

- There is a normalisation constant

$p(A = \mathtt{a}, B = \mathtt{b})$ is not a distribution in A

- $\sum_{\mathtt{a}} p(A = \mathtt{a}, B = \mathtt{b}) \neq 1$

To make it a distribution, we need to normalise

$$p(A = \mathtt{a}|B = \mathtt{b}) = \frac{p(A = \mathtt{a}, B = \mathtt{b})}{\sum_{\mathtt{a}} p(A = \mathtt{a}, B = \mathtt{b})}$$

$\rightsquigarrow$ This, summed over $\mathtt{a}$, does sum to 1

Probabilistic modelling - Conditional probability (cont.)

Definition

Independence

Variables x and y are independent if knowing the state (or value) of one variable gives no extra info about the other variable

$$p(x, y) = p(x)p(y) \tag{26}$$

For $p(x) \neq 0$ and $p(y) \neq 0$, the independence of x and y

$$p(x|y) = p(x) \Longleftrightarrow p(y|x) = p(y) \tag{27}$$

If $p(x|y) = p(x)$ for all states of x and y, then x and y are independent

■

Probabilistic modelling - Conditional probability (cont.)

Suppose that for some constant k and some positive functions $f(\cdot)$ and $g(\cdot)$,

$$p(x, y) = kf(x)g(y) \tag{28}$$

Then, we say that x and y are independent

- ⇝ We write $x \perp\!\!\!\perp y$

Probabilistic modelling - Conditional probability (cont.)

Example

Let x denote the day of the week in which females are born

Let y be the day in which males are born

$$\text{dom}(x) = \text{dom}(y) = \{\texttt{M}, \texttt{T}, \ldots, \texttt{S}\}$$

It seems reasonable to expect that x is independent of y

Suppose randomly select a woman from the phone book (Alice)

- We find out that she was born on a Tuesday

Probabilistic modelling - Conditional probability (cont.)

Knowing when Alice was born does not provide extra information

$$p(y|x) = p(y)$$

- The probabilities of Bob's birthday remain unchanged

This does not mean that the distribution of Bob's birthday is uniform

The distribution of birth days $p(y)$ and $p(x)$ are non-uniform

- (fewer babies are born on weekends, statistically)

Although nothing suggests that x and y are independent

Probabilistic modelling - Conditional probability (cont.)

Sometimes the concept of independence may perhaps appear a little strange

Example

Consider binary variables x and y (their domains consist of 2 states)

Define the distribution such that x and y are always both in a certain state

$$p(x = \mathtt{a}, y = \mathtt{1}) = 1$$
$$p(x = \mathtt{a}, y = \mathtt{2}) = 0$$
$$p(x = \mathtt{b}, y = \mathtt{2}) = 1$$
$$p(x = \mathtt{b}, y = \mathtt{1}) = 0$$

Are x and y dependent?

Probabilistic modelling - Conditional probability (cont.)

Since

- $p(x = \texttt{a}) = 1$, $p(x = \texttt{b}) = 0$
- $p(y = \texttt{1}) = 1$, $p(y = \texttt{2}) = 0$

We have that $p(x)p(y) = p(x, y)$ for ALL states of x and y

$\rightsquigarrow$ x and y are thus independent

This may seem strange, as we know of the relation between x and y

- They are always in the same joint state and yet independent

Probabilistic modelling - Conditional probability (cont.)

The distribution is concentrated in a single joint state

- Knowing the state of x tells nothing more about the state of y
- (and viceversa)

This potential confusion comes from using the term ‘independent’

- This may suggest that there is no relations between objects

Probabilistic modelling - Conditional probability (cont.)

Remark

To get the concept of statistical independence, ask whether or not knowing the state of y tells something more than we knew about the state of x

- '*knew before*' means reasoning with the joint distribution $p(x, y)$
- To figure out what we can know about x, or equivalently $p(x)$

■

Probabilistic modelling - Conditional probability (cont.)

Definition

Conditional independence

Sets of variables $\mathcal{X}$ *and* $\mathcal{Y}$ *are said to be independent of each other if, given all states of* $\mathcal{X}$, $\mathcal{Y}$ *and* $\mathcal{Z}$, *we have*

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}), \tag{29}$$

provided that we know the state of set $\mathcal{Z}$

We write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}|\mathcal{Z}$

Probabilistic modelling - Conditional probability (cont.)

If the conditioning set is empty, we may also write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ for $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \emptyset$

- $\mathcal{X}$ is (conditionally) independent of $\mathcal{Y}$

If $\mathcal{X}$ and $\mathcal{Y}$ are not conditionally independent $\Rightarrow$ conditionally dependent

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z} \tag{30}$$

Similarly, $\mathcal{X} \top\!\!\!\top \mathcal{Y} | \emptyset$ can be written as $\mathcal{X} \top\!\!\!\top \mathcal{Y}$

Probabilistic modelling - Conditional probability (cont.)

Intuitively, if x is conditionally independent of y given z

- ⇝ Then, given z, y contains no additional information about x
- ⇝ Then, given z, knowing x does not add information about y

Probabilistic modelling - Conditional probability (cont.)

Remark

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z} \Longrightarrow \mathcal{X}' \perp\!\!\!\perp \mathcal{Y}' | \mathcal{Z}, \quad \text{for } \mathcal{X}' \subseteq \mathcal{X} \text{ and } \mathcal{Y}' \subseteq \mathcal{Y}$$

Probability refresher - Conditional probability (cont.)

Remark

Independence implications

It is tempting to think that if 'a *is independent of* b' and 'b *is independent of* c', then 'a *must be independent of* c'

$$\{a \perp\!\!\!\perp b, b \perp\!\!\!\perp c\} \Longrightarrow a \perp\!\!\!\perp c \tag{31}$$

However, this does NOT necessarily hold true

Probabilistic modelling - Conditional probability (cont.)

Consider the distribution

$$p(a, b, c) = p(b)p(a, c) \tag{32}$$

From this,

$$p(a, b) = \sum_{\mathsf{c}} p(a, b, c) = p(b) \sum_{\mathsf{c}} p(a, b) \tag{33}$$

$p(a, b)$ is a function of b multiplied by a function of a

$\rightsquigarrow$ a and b are independent

- One can show that b and c are independent
- One can show that a is not necessarily independent of c

$\rightsquigarrow$ (distribution $p(a, c)$ can be set arbitrarily)

Probabilistic modelling - Conditional probability (cont.)

Remark

Similarly, it is tempting to think that if '*a and b are dependent*', and '*b and c are dependent*', then '*a and c must be dependent*'

$$\{a \top\!\!\top b, b \top\!\!\top c\} \Longrightarrow a \top\!\!\top c \tag{34}$$

However, this also does NOT follow ($\star$)

Remark

Conditional independence

$$x \perp\!\!\!\perp y | z$$

does not imply marginal independence

$$x \perp\!\!\!\perp y$$

Probabilistic modelling - Probability tables

Population of countries (CNT) England ($\texttt{E}$), Scotland ($\texttt{S}$) and Wales ($\texttt{W}$)

- England ($\texttt{E}$), 60776238
- Scotland ($\texttt{S}$), 5116900
- Wales ($\texttt{W}$), 2980700

A priori probability that a randomly selected person from the combined countries would live in England, Scotland or Wales is 0.88, 0.08 and 0.04

$$\begin{pmatrix} p(CNT = \texttt{E}) \\ p(CNT = \texttt{S}) \\ p(CNT = \texttt{W}) \end{pmatrix} = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix} \tag{35}$$

- These are based on population

Component values sum to 1 and the ordering is arbitrary

Probability modelling - Probability tables (cont.)

For simplicity, assume that only three mother tongues (MT) exist

- English (`Eng`)
- Scottish (`Scot`)
- Welsh (`Wel`)

Probability modelling - Probability tables (cont.)

The conditional probabilities $p(MT|CNT)$ by residence $\texttt{E}$, $\texttt{S}$ and $\texttt{W}$

$$p(MT = \texttt{Eng}|CNT = \texttt{E}) = 0.95$$
$$p(MT = \texttt{Scot}|CNT = \texttt{E}) = 0.04$$
$$p(MT = \texttt{Wel}|CNT = \texttt{E}) = 0.01$$

$$p(MT = \texttt{Eng}|CNT = \texttt{S}) = 0.70$$
$$p(MT = \texttt{Scot}|CNT = \texttt{S}) = 0.30$$
$$p(MT = \texttt{Wel}|CNT = \texttt{S}) = 0.00$$

$$p(MT = \texttt{Eng}|CNT = \texttt{W}) = 0.60$$
$$p(MT = \texttt{Scot}|CNT = \texttt{W}) = 0.00$$
$$p(MT = \texttt{Wel}|CNT = \texttt{W}) = 0.40$$

Probability modelling - Probability tables (cont.)

$$\begin{pmatrix} p(\texttt{Eng}|\texttt{E}) & p(\texttt{Eng}|\texttt{S}) & p(\texttt{Eng}|\texttt{W}) \\ p(\texttt{Scot}|\texttt{E}) & p(\texttt{Scot}|\texttt{S}) & p(\texttt{Scot}|\texttt{W}) \\ p(\texttt{Wel}|\texttt{E}) & p(\texttt{Wel}|\texttt{S}) & p(\texttt{Wel}|\texttt{W}) \end{pmatrix} = \begin{pmatrix} 0.95 & 0.70 & 0.60 \\ 0.04 & 0.30 & 0.00 \\ 0.01 & 0.00 & 0.40 \end{pmatrix}$$

Probabilistic modelling - Probability tables (cont.)

We can form a joint distribution $p(CNT, MT) = p(MT|CNT)p(CNT)$

$$\begin{pmatrix} p(\texttt{Eng}, \texttt{E}) & p(\texttt{Eng}, \texttt{S}) & p(\texttt{Eng}, \texttt{W}) \\ p(\texttt{Scot}, \texttt{E}) & p(\texttt{Scot}, \texttt{S}) & p(\texttt{Scot}, \texttt{W}) \\ p(\texttt{Wel}, \texttt{E}) & p(\texttt{Wel}, \texttt{S}) & p(\texttt{Wel}, \texttt{W}) \end{pmatrix}$$

We can write it in the form of a 3×3 matrix

- Columns indexed by country, rows indexed by mother tongue

$$\begin{pmatrix} 0.95 \times 0.88 & 0.70 \times 0.08 & 0.60 \times 0.04 \\ 0.04 \times 0.88 & 0.30 \times 0.08 & 0.00 \times 0.04 \\ 0.01 \times 0.88 & 0.00 \times 0.08 & 0.40 \times 0.04 \end{pmatrix} = \begin{pmatrix} 0.8360 & 0.056 & 0.024 \\ 0.0352 & 0.024 & 0.000 \\ 0.0088 & 0.000 & 0.016 \end{pmatrix}$$

Probabilistic modelling - Probability tables (cont.)

Remark

The joint distribution contains all the information about the model

Probabilistic modelling - Probability tables (cont.)

$$p(CNT, MT) = \begin{pmatrix} 0.8360 & 0.0560 & 0.0240 \\ 0.0352 & 0.0240 & 0.0000 \\ 0.0088 & 0.0000 & 0.0160 \end{pmatrix}$$

By summing the columns, we have the marginal $p(CNT)$

$$p(CNT) = \sum_{MT \in \text{dom}(MT)} p(CNT, MT)$$

That is,

$$\begin{pmatrix} p(CNT = \texttt{E}) \\ p(CNT = \texttt{S}) \\ p(CNT = \texttt{W}) \end{pmatrix} = \begin{pmatrix} 0.8352 + 0.0352 + 0.0088 = 0.88 \\ 0.0352 + 0.0240 + 0.0000 = 0.08 \\ 0.0088 + 0.0000 + 0.0160 = 0.04 \end{pmatrix} \tag{36}$$

Probabilistic modelling - Probability tables (cont.)

$$p(CNT, MT) = \begin{pmatrix} 0.8360 & 0.0560 & 0.0240 \\ 0.0352 & 0.0240 & 0.0000 \\ 0.0088 & 0.0000 & 0.0160 \end{pmatrix}$$

By summing the rows, we have the marginal $p(MT)$

$$p(MT) = \sum_{CNT \in \text{dom}(CNT)} p(CNT, MT)$$

That is,

$$\begin{pmatrix} p(MT = \texttt{Eng}) \\ p(MT = \texttt{Scot}) \\ p(MT = \texttt{Wel}) \end{pmatrix} = \begin{pmatrix} 0.8360 + 0.0560 + 0.0240 = 0.916 \\ 0.0352 + 0.0240 + 0.0000 = 0.059 \\ 0.0088 + 0.0000 + 0.0160 = 0.025 \end{pmatrix} \qquad (37)$$

Probabilistic modelling - Probability tables (cont.)

We infer the conditional distribution $p(CNT|MT) \propto p(MT|CNT)p(CNT)$

$$p(CNT|MT) = \begin{pmatrix} 0.913 & 0.061 & 0.026 \\ 0.590 & 0.410 & 0.000 \\ 0.360 & 0.000 & 0.640 \end{pmatrix}$$

The $p(CNT|MT)$ by dividing entries of $p(CNT, MT)$ by their rowsum

Probabilistic modelling - Probability tables (cont.)

Consider joint distributions over a larger set of variables $\{x_i\}_{i=1}^{D}$

- Suppose that each variable x_i taking K_i states

The table of the joint distribution is an array with $\prod_{i=1}^{D} K_i$ entries

Storing tables requires space exponential in the number of variables

- It rapidly becomes impractical for a large number D

Probabilistic modelling - Probability tables (cont.)

Remark

A distribution assigns a value to each of the joint states of variables

- $p(T, J, R, S)$ is equivalent to $p(J, S, R, T)$ or any reordering

The joint setting of variables is a different index to the same probability

More clear in the set theoretic notation $p(J \cap S \cap T \cap R)$

Probability reasoning

Reasoning under uncertainty

Probabilistic reasoning

The central paradigm of probabilistic reasoning

Identify all relevant variables $x_1, \dots, x_N$ in the environment

Make a **probabilistic model**

$$p(x_1, \dots, x_N)$$

Reasoning (or **inference**) is performed by introducing **evidence**

- Evidence sets variables in known states

Then, we compute probabilities of interest, conditioned on this evidence

Probabilistic reasoning (cont.)

Probability theory with Bayes' rule make for a complete reasoning system

- Deductive logic emerges as a special case

We discuss examples in which the number of variables is still very small

Then, we shall discuss reasoning in networks of many variables

- A graphical notation will play a central role

Probabilistic reasoning (cont.)

Example

Hamburgers and the KJ disease

People with Kreuzfel-Jacob (KJ) disease almost inevitably ate hamburgers

$$p(\textit{Hamburger eater} = \texttt{tr} | KJ = \texttt{tr}) = 0.9$$

The probability of a person having KJ is very low

$$p(KJ = \texttt{tr}) = 1/100K$$

Assume that eating hamburgers is spread

$$p(\textit{Hamburger eater} = \texttt{tr}) = 0.5$$

What is the probability that a hamburger eater will have KJ disease?

Probabilistic reasoning (cont.)

$$
\begin{aligned}
p(\texttt{KJ}|\texttt{Hamburger eater}) &= \frac{p(\texttt{Hamburger eater}, \texttt{KJ})}{p(\texttt{Hamburger eater})} \\
&= \frac{p(\texttt{Hamburger eater}|\texttt{KJ})p(\texttt{KJ})}{p(\texttt{Hamburger eater})} \\
&= \frac{9/10 \times 1/100K}{1/2} = 1.8 \times 10^{-5}
\end{aligned} \tag{38}
$$

Suppose that $p(\texttt{Hamburger eater}) = 0.001$

$\rightsquigarrow$ $p(\texttt{KJ}|\texttt{Hamburger eater}) \approx 1/100$

Probabilistic reasoning (cont.)

Example

Inspector Clouseau

Inspector Clouseau arrives at the scene of a crime

The victim lies dead near the possible murder weapon, a knife

$\rightsquigarrow$ K

- $\text{dom}(K) = \{\texttt{knife used}, \texttt{knife not used}\}$

The butler (B) and the maid (M) are the inspector's main suspects

$\rightsquigarrow$ B and M

- $\text{dom}(B) = \text{dom}(\text{M}) = \{\texttt{murderer}, \texttt{not murderer}\}$

Probabilistic reasoning (cont.)

Prior beliefs that they are the murderer

$$p(B = \texttt{murderer}) = 0.6$$
$$p(M = \texttt{murderer}) = 0.2$$

These beliefs are independent

$$p(B)p(M) = p(B, M)$$

Probabilistic reasoning (cont.)

Still possible that both the butler and the maid killed the victim or neither

$$
\begin{aligned}
p(K = \texttt{knife used} | B = \texttt{not murderer}, M = \texttt{not murderer}) &= 0.3 \\
p(K = \texttt{knife used} | B = \texttt{not murderer}, M = \texttt{murderer}) &= 0.2 \\
p(K = \texttt{knife used} | B = \texttt{murderer}, M = \texttt{not murderer}) &= 0.6 \\
p(K = \texttt{knife used} | B = \texttt{murderer}, M = \texttt{murderer}) &= 0.1
\end{aligned}
$$

In addition, $p(K, B, M) = p(K|B, M)p(B, M) = p(K|B, M)p(B)p(M)$

Probabilistic reasoning (cont.)

Assume that the knife is the murder weapon ($K = \texttt{tr}$)

What is the probability that the butler is the murderer

$$p(B = \texttt{murderer} | K = \texttt{tr})$$

Probabilistic reasoning (cont.)

- Let $b = \text{dom}(B)$ indicate the two states of B
- Let $m = \text{dom}(M)$ indicate the two states of M

$$
\begin{aligned}
p(B|K) &= \sum_{M \in m} p(B, M|K) = \sum_{M \in m} \frac{p(B, M, K)}{p(K)} = \frac{1}{p(K)} \sum_{M \in m} p(B, M, K) \\
&= \frac{1}{\sum_{\substack{B \in b \\ M \in m}} p(B, M, K)} \sum_{M \in m} p(K|B, M) p(B, M) \\
&= \frac{1}{\sum_{\substack{B \in b \\ M \in m}} p(K|B, M) p(B, M)} \sum_{M \in m} p(K|B, M) p(B, M) \\
&= \frac{1}{\sum_{\substack{B \in b \\ M \in m}} p(K|B, M) p(B) p(M)} \sum_{M \in m} p(K|B, M) p(B) p(M) \\
&= \frac{1}{\sum_{B \in b} \left[p(B) \sum_{M \in m} p(K|B, M) p(M) \right]} p(B) \sum_{M \in m} p(K|B, M) p(M)
\end{aligned} \tag{39}
$$

We used $p(B, M) = p(B)p(M)$

Probabilistic reasoning(cont.)

Plugging in the values, we have

$$p(B = \texttt{murderer} | K = \texttt{knife used})$$
$$= \frac{\frac{6}{10}\Big(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\Big)}{\frac{6}{10}\Big(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\Big) + \frac{4}{10}\Big(\frac{2}{10} \times \frac{2}{10} + \frac{8}{10} \times \frac{3}{10}\Big)}$$
$$= \frac{300}{412} \simeq 0.73 \quad (40)$$

Knowing it was the knife strengthens our belief that the butler did it

Probabilistic reasoning (cont.)

Remark

The role of $p(K = \texttt{knife used})$ in the example can cause confusion

$$p(K = \texttt{knife used}) = \sum_{B \in b} p(B) \sum_{M \in m} p(K = \texttt{knife used} | B, M) p(M) \\ = 0.412 \tag{41}$$

But surely also $p(K = \texttt{knife used}) = 1$, since this is given

Quantity $p(K = \texttt{knife used})$ relates to the **prior**

- The probability the model assigns to the knife being used
- (in the absence of any other info)

Probabilistic reasoning (cont.)

Clearly, if we know that the knife is used then the **posterior**

$$p(K = \texttt{knife used} | K = \texttt{knife used}) = \frac{p(K = \texttt{knife used}, K = \texttt{knife used})}{p(K = \texttt{knife used})} = \frac{p(K = \texttt{knife used})}{p(K = \texttt{knife used})} = 1 \quad (42)$$

Probabilistic reasoning (cont.)

Example

Who's in the bathroom?

A household of 3 persons

- Alice
- Bob
- Cecil

Cecil wants to go to the bathroom but finds it occupied

- So, he goes to Alice's room
- He sees she is there

Cecil knows that only either Bob or Alice can be in the bathroom

- He infers that Bob must be occupying it

Probabilistic reasoning (cont.)

We can arrive at the same conclusion mathematically

Define the events

$$\begin{cases} A: & \text{Alice is in her bedroom} \\ B: & \text{Bob is in his bedroom} \\ O: & \text{Bathroom is occupied} \end{cases} \tag{43}$$

We need to encode the available information

Probabilistic reasoning (cont.)

If either Alice or Bob are not in their rooms, they must be in the bathroom

- (both may be there)

$$\begin{aligned} p(O = \texttt{tr} | A = \texttt{fa}, B) = 1 \\ p(O = \texttt{tr} | B = \texttt{fa}, A) = 1 \end{aligned} \tag{44}$$

- The first term expresses that the bathroom is occupied ($O = \texttt{tr}$) if Alice is not in her bedroom ($A = \texttt{fa}$), wherever Bob is (B)
- The second term expresses that the bathroom is occupied ($O = \texttt{tr}$) if Bob is not in his bedroom ($B = \texttt{fa}$), wherever Alice is (A)

Probabilistic reasoning (cont.)

$$p(B = \texttt{fa} | O = \texttt{tr}, A = \texttt{tr}) = \frac{p(B = \texttt{fa}, O = \texttt{tr}, A = \texttt{tr})}{p(O = \texttt{tr}, A = \texttt{tr})}$$

$$= \frac{\underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa})}_{1} p(A = \texttt{tr}, B = \texttt{fa})}{p(O = \texttt{tr}, A = \texttt{tr})} \quad (45)$$

$$p(O = \texttt{tr}, A = \texttt{tr})$$
$$= \underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa})}_{1} p(A = \texttt{tr}, B = \texttt{fa})$$
$$+ \underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{tr})}_{0} p(A = \texttt{tr}, B = \texttt{tr}) \quad (46)$$

- If Alice is in her room and Bob is not, the bathroom must be occupied

$$p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa}) = 1$$

- If Alice and Bob are in their rooms, the bathroom cannot be occupied

$$p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{tr}) = 0$$

Probabilistic reasoning (cont.)

$$p(B = \texttt{fa} | O = \texttt{tr}, A = \texttt{tr}) = \frac{p(A = \texttt{tr}, A = \texttt{fa})}{p(A = \texttt{tr}, B = \texttt{fa})} = 1 \quad (47)$$

■

Probabilistic reasoning (cont.)

Remark

The example is interesting

We are not required to make a full probabilistic model

We do not need to specify $p(A, B)$

- The situation is common in limiting situations of probabilities
- Probabilities being either 0 or 1

Probabilistic reasoning corresponds to traditional logic systems

Probabilistic reasoning (cont.)

Example

Aristotles - Modus Ponens

According to logic, statements '*All apples are fruit*' and '*All fruits grow on trees*' lead to '*All apples grow on trees*'

This kind of reasoning is a form of transitivity

From the statements $A \Rightarrow F$ and $F \Rightarrow T$, we can infer $A \Rightarrow T$

This may be reduced to probabilistic reasoning

- '*All apples are fruits*' corresponds to $p(F = \texttt{tr}|A = \texttt{tr}) = 1$
- '*All fruits grow on trees*' corresponds to $p(T = \texttt{tr}|F = \texttt{tr}) = 1$

Probabilistic reasoning (cont.)

We want to show that this implies one of the two

- $p(T = \texttt{tr}|A = \texttt{tr}) = 1$, '*All apples grow on trees*'
- $p(T = \texttt{fa}|A = \texttt{tr}) = 0$, '*All apples do not grow on non-trees*'

$$p(T = \texttt{fa}|A = \texttt{tr}) = \frac{p(T = \texttt{fa}, A = \texttt{tr})}{p(A = \texttt{tr})}$$

Assuming that $p(A = \texttt{tr}) > 0$, these equal to $p(T = \texttt{fa}, A = \texttt{tr}) = 0$

$$p(T = \texttt{fa}, A = \texttt{tr}) = \\ p(T = \texttt{fa}, A = \texttt{tr}, F = \texttt{tr}) + p(T = \texttt{fa}, A = \texttt{tr}, F = \texttt{fa}) \quad (48)$$

We need to show that both terms on the right-hand side are zero

Probabilistic reasoning (cont.)

- Since $p(T = \texttt{fa} | F = \texttt{tr}) = 1 - p(T = \texttt{tr} | F = \texttt{tr}) = 1 - 1 = 0$,

$$p(T = \texttt{fa}, A = \texttt{tr}, F = \texttt{tr}) \\ \leq p(T = \texttt{fa}, F = \texttt{tr}) = p(T = \texttt{fa} | F = \texttt{tr}) p(F = \texttt{tr}) = 0, \quad (49)$$

- By assumption $p(F = \texttt{fa} | A = \texttt{tr}) = 0$,

$$p(T = \texttt{fa}, A = \texttt{tr}, F = \texttt{fa}) \\ \leq p(A = \texttt{tr}, F = \texttt{fa}) = p(F = \texttt{fa} | A = \texttt{tr}) p(A = \texttt{tr})) = 0, \quad (50)$$

Probabilistic reasoning (cont.)

Example

Aristotles - Inverse Modus Ponens

According to logic, statement '*If A is true then B is true*' leads to deduce that '*If B is false then A is false*'

We show how this can be represented by using probabilistic reasoning

'*If A is true then B is true*' corresponds to

$$p(B = \mathtt{tr} | A = \mathtt{tr}) = 1$$

We may infer

$$\begin{aligned} p(A = \mathtt{fa} | B = \mathtt{fa}) &= 1 - p(A = \mathtt{tr} | B = \mathtt{fa}) \\ &= 1 - \frac{p(B = \mathtt{fa} | A = \mathtt{tr}) p(A = \mathtt{tr})}{p(B = \mathtt{fa} | A = \mathtt{tr}) p(A = \mathtt{tr}) + p(B = \mathtt{fa} | A = \mathtt{fa}) p(A = \mathtt{fa})} \\ &= 1 \end{aligned} \quad (51)$$

It follows since $p(B = \mathtt{fa} | A = \mathtt{tr}) = 1 - p(B = \mathtt{br} | A = \mathtt{tr}) = 1 - 1 = 0$

Probabilistic reasoning (cont.)

Example

Soft XOR gate

What about inputs A and B, knowing the output is 0?

A	B	A xor B
0	0	0
0	1	1
1	0	1
1	1	0

The 'standard' XOR gate

1. A and B were both 0
2. A and B were both 1

We do not know which state A is in, it could equally likely be 0 or 1

Probabilistic reasoning (cont.)

A 'soft' XOR gate stochastically outputs $C = 1$ depending on its inputs

A	B	$p(C = 1 \vert A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

Additionally, let $A \perp\!\!\!\perp B$ and

- $p(A = 1) = 0.65$
- $p(B = 1) = 0.77$

Probabilistic reasoning (cont.)

What's up with $p(A = 1|C = 0)$?

$$
\begin{aligned}
p(A = 1, C = 0) = & \sum_B p(A = 1, B, C = 0) \\
= & \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\
= & p(A = 1)p(C = 0|A = 1, B = 0)p(B = 0)+ \\
& p(A = 1)p(C = 0|A = 1, B = 1)p(B = 1) \\
= & 0.65 \times (0.2 \times 0.23 + 0.75 \times 0.77) = 0.405275
\end{aligned} \tag{52}
$$

Probabilistic reasoning (cont.)

$$\begin{aligned} p(A = 0, C = 0) &= \sum_B p(A = 0, B, C = 0) \\ &= \sum_B p(C = 0 | A = 0, B) p(A = 0) p(B) \\ &= p(A = 0) p(C = 0 | A = 0, B = 0) p(B = 0) + \\ &\quad p(A = 1) p(C = 0 | A = 0, B = 1) p(B = 1) \\ &= 0.35 \times (0.9 \times 0.23 + 0.01 \times 0.77) = 0.075145 \end{aligned} \tag{53}$$

Probabilistic reasoning (cont.)

$$
\begin{aligned}
p(A = 1 | C = 0) &= \frac{p(A = 1, C = 0)}{p(A = 1, C = 0) + p(A = 0, C = 0)} \\
&= \frac{0.405275}{0.405275 + 0.075145} \\
&= 0.8436
\end{aligned}
\tag{54}
$$

Probabilistic reasoning (cont.)

Example

Larry

Larry is typically late for school

When his mum asks whether or not he was late, never admits to being late

- We denote Larry being late with $L = \texttt{late}$, otherwise $L = \texttt{not late}$

The response Larry gives is denoted by R_L

- $p(R_L = \texttt{not late} | L = \texttt{not late}) = 1$
- $p(R_L = \texttt{late} | L = \texttt{late}) = 0$

The remaining two values are determined by normalisation

- $p(R_L = \texttt{late} | L = \texttt{not late}) = 0$
- $p(R_L = \texttt{not late} | L = \texttt{late}) = 1$

Given that $R_L = \texttt{not late}$, what is the probability that Larry was late?

$$p(L = \texttt{late} | R_L = \texttt{not late})$$

Probabilistic reasoning (cont.)

Using Bayes' rule,

$$p(L = \texttt{late}|R_L = \texttt{not late}) = \frac{p(L = \texttt{late}, R_L = \texttt{not late})}{p(R_L = \texttt{not late})}$$
$$= \frac{p(L = \texttt{late}, R_L = \texttt{not late})}{p(L = \texttt{late}, R_L = \texttt{not late}) + p(L = \texttt{not late}, R_L = \texttt{not late})} \quad (55)$$

We recognise

$$p(L = \texttt{late}, R_L = \texttt{not late}) = \underbrace{p(R_L = \texttt{not late}|L = \texttt{late})}_{1} p(L = \texttt{late}) \quad (56)$$

$$p(L = \texttt{not late}, R_L = \texttt{not late}) = \underbrace{p(R_L = \texttt{not late}|L = \texttt{not late})}_{1} p(L = \texttt{not late}) \quad (57)$$

Probabilistic reasoning (cont.)

$$p(L = \texttt{late} | R_L = \texttt{not late}) = \frac{p(L = \texttt{late})}{p(L = \texttt{late}) + p(L = \texttt{not late})} = p(L = \texttt{late}) \qquad (58)$$

Larry's mother knows that he never admits to being late

- Her belief about whether or not he was late is unchanged
- (regardless of what Larry actually says)

In the last step we used normalisation, $p(L = \texttt{late}) + p(L = \texttt{not late}) = 1$

Probabilistic reasoning (cont.)

Example

Larry the lair and his sister Sue

Unlike Larry, his sister Sue always tells the truth to her mother

- (As to whether or not Larry is late for school)

$$p(R_S = \texttt{not late} | L = \texttt{not late}) = 1 \implies p(R_S = \texttt{late} | L = \texttt{not late}) = 0$$

$$p(R_S = \texttt{late} | L = \texttt{late}) = 1 \implies p(R_S = \texttt{not late} | L = \texttt{late}) = 0$$

We also assume that $p(R_S, R_L | L) = p(R_S | L) p(R_L | L)$

Then, we write

$$p(R_S, R_L, L) = p(R_L | L) p(R_S | L) p(L) \tag{59}$$

Given $R_S = \texttt{late}$ and $R_L = \texttt{not late}$, what the probability that he late?

Probabilistic reasoning (cont.)

Using Bayes' rule,

$$p(L = \texttt{late} | R_L = \texttt{nlate}, R_S = \texttt{late}) = \frac{1}{Z} p(R_S = \texttt{late} | L = \texttt{late}) p(R_L = \texttt{nlate} | L = \texttt{late}) p(L = \texttt{late}) \quad (60)$$

The normalisation term $1/Z$,

$$\frac{1}{Z} = p(R_S = \texttt{late} | L = \texttt{late}) p(R_L = \texttt{nlate} | L = \texttt{late}) p(L = \texttt{late}) + p(R_S = \texttt{late} | L = \texttt{nlate}) p(R_L = \texttt{nlate} | L = \texttt{nlate}) p(L = \texttt{nlate}) \quad (61)$$

Probabilistic reasoning (cont.)

Hence,

$$p(L = \texttt{late} | R_L = \texttt{not late}, R_S = \texttt{late}) = \frac{1 \times 1 \times p(L = \texttt{late})}{1 \times 1 \times p(L = \texttt{late}) + 0 \times 1 \times p(L = \texttt{not late})} = 1 \quad (62)$$

Larry's mother knows that Sue tells the truth, no matter what Larry says

Probabilistic reasoning (cont.)

Example

Luke

Luke has been told he is lucky and has won a prize in the lottery

5 prizes available

- 10 $\rightsquigarrow (p_1)$
- 100 $\rightsquigarrow (p_2)$
- $1K$ $\rightsquigarrow (p_3)$
- $10K$ $\rightsquigarrow (p_4)$
- $1M$ $\rightsquigarrow (p_5)$

p_0 is the prior probability of winning no prize

$$p_0 + p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

- Luke asks '*Did I win 1M*?!', '*I'm afraid not sir*' the lottery guy
- '*Did I win 10K*?!' asks Luke, '*Again, I'm afraid not sir*'

What is the probability that Luke has won 1K?

Probabilistic reasoning (cont.)

We denote

- $W = 1$ for the first prize (10)
- $W = 2, \dots, 5$ for the remaining prices (100, 1K, 10K, 1M)
- $W = 0$ for no prize (0)

$$\begin{aligned} p(W = 3 | W \neq 5, W \neq 4, W \neq 0) &= \frac{p(W = 3, W \neq 5, W \neq 4, W \neq 0)}{p(W \neq 5, W \neq 4, W \neq 0)} \\ &= \frac{p(W = 3)}{\underbrace{p(W = 1 \text{ or } W = 2 \text{ or } W = 3)}_{\text{events are mutually exclusive}}} \\ &= \frac{p_3}{p_1 + p_2 + p_3} \end{aligned} \tag{63}$$

Probabilistic reasoning (cont.)

The results makes intuitive sense

We remove the impossible states of W

The probability to win 1K is proportional to its prior probability (p_3)

- normalisation is the total set of possible probability left

Prior, likelihood and posterior

Reasoning under uncertainty

Prior, likelihood and posterior

Tell me something about variable Θ, given that

i) *I have observed data $\mathcal{D}$*

ii) *I have some knowledge of the data generating mechanism*

The quantity of interest

$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{\int_{\Theta} p(\mathcal{D}|\Theta)p(\Theta)} \quad (64)$$

A **generative model** $p(\mathcal{D}|\Theta)$ of the data

A **prior belief** $p(\Theta)$ about which variable values are appropriate

- We can infer the **posterior distribution** $p(\Theta|\mathcal{D})$ of the variables
- (In the light of the observed data)

Prior, likelihood and posterior (cont.)

The **most probable a posteriori** (**MAP**) setting maximises the posterior

$$\Theta_* = \arg\max_{\Theta} \big[p(\Theta|\mathcal{D}) \big]$$

Consider a flat prior, $p(\Theta)$ being a constant (with Θ)

The MAP solution is equivalent to the **maximum likelihood** solution

- The Θ that maximises the **likelihood** $p(\mathcal{D}|\Theta)$
- (of the model generating the data)

Prior, likelihood and posterior (cont.)

The use of the generative model suits well with physical modelling

- We typically postulate how to generate observed phenomena
- (Assuming we know the model)

Prior, likelihood and posterior (cont.)

One might postulate how to generate a time-series

Consider the displacements for a swinging pendulum

- Unknown mass, length and dumping constant

We infer the unknown physical properties of the pendulum

- Using the generative model
- Given the displacements

Prior, likelihood and posterior (cont.)

Example

Pendulum

Consider a pendulum

- Let x_t be the angular displacement at t

Assume that measurements are independent

- The likelihood of a sequence $x_1, \dots, x_T$

$$p(x_1, \dots, x_T|\Theta) = \prod_{t=1}^{T} p(x_t|\Theta) \tag{65}$$

Depends on the knowledge of the problem parameter Θ

Prior, likelihood and posterior (cont.)

Assume that the model is correct

Assume that our measurement of the displacement x is perfect

Then, the physical model

$$\rightsquigarrow \quad x_t = \sin{(\Theta t)}, \tag{66}$$

Θ is the unknown constants of the pendulum ($\sqrt{g/L}$)

- g is the gravitational attraction
- L the pendulum length

Prior, likelihood and posterior (cont.)

Assume that we have a poor instrument to measure the displacements

Measurements have a Gaussian distribution

- The variance σ^2 is known

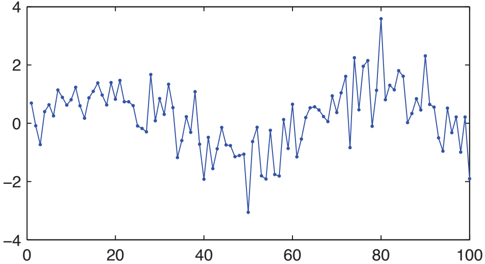
Then, the physical model

$$\rightsquigarrow \quad x_t = \sin(\Theta t) + \varepsilon_t \tag{67}$$

ε_t is zero mean Gaussian noise with variance σ^2

Prior, likelihood and posterior (cont.)

We have data: Noisy observations of displacements $x_1, \ldots, x_{100}$



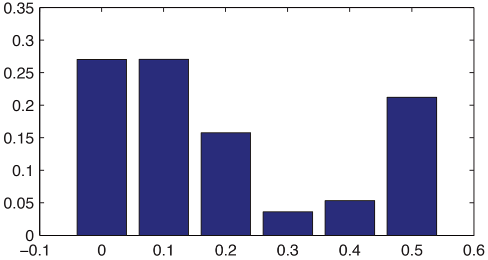
Prior, likelihood and posterior (cont.)

We can also consider a set of possible parameters Θ

We can place a prior $p(\Theta)$ over them

- Express our prior belief (before even seeing the measurements)
- Our trust in the appropriateness of different values of Θ

The prior belief on 5 possible values of Θ



Prior, likelihood and posterior (cont.)

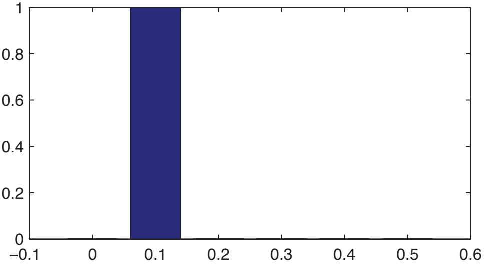
The posterior distribution

$$p(\Theta|x_1, \dots, x_N) \propto p(\Theta) \prod_{t=1}^{T} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{1}{2\sigma^2} \Big[x_t - \sin(\Theta t) \Big]^2 \Big\}}_{p(x_t|\theta)} \quad (68)$$

The posterior belief over the assumed values of Θ becomes strongly peaked

- For a large number of measurements, despite noisy measurements

The posterior belief on Θ



Two dice: Individual scores

Example

Two fair dice are rolled and someone tells that the sum of scores is 9

⇝ What is the posterior distribution of the dice scores?

⇝ The score of die a is denoted by s_a, with $\text{dom}(s_a) = \{1, 2, 3, 4, 5, 6\}$

⇝ The score of die b is denoted by s_b, with $\text{dom}(s_b) = \{1, 2, 3, 4, 5, 6\}$

The three variables involved are then s_a, s_b and $t = s_a + s_b$

We jointly model them

$$p(t, s_a, s_b) = \underbrace{p(t|s_a, s_b)}_{\text{likelihood}} \underbrace{p(s_a, s_b)}_{\text{prior}} \tag{69}$$

Two dice: Individual scores (cont.)

The prior $p(s_a, s_b)$ is the joint probability of scores s_a and s_b

- Without knowing anything else
- Assuming no dependency in the rolling

$$p(s_a, s_b) = p(s_a)p(s_b) \tag{70}$$

Since dice are fair, both $p(s_a)$ and $p(s_b)$ are uniform distributions

$$p(s_a) = p(s_b) = 1/6$$

Two dice: Individual scores (cont.)

The likelihood $p(t|s_a, s_b)$ states the total score $t = s_a + s_b$

$$p(t|s_a, s_b) = \mathbb{I}[t = s_a + s_b] \tag{71}$$

Function $\mathbb{I}[A]$ is such that $\mathbb{I}[A] = 1$ if statement A is true, 0 otherwise

Two dice: Individual scores (cont.)

$p(s_a)p(s_b)$	$s_a = 1$	$s_a = 2$	$s_a = 3$	$s_a = 4$	$s_a = 5$	$s_a = 6$
$s_b = 1$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b = 2$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b = 3$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b = 4$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b = 5$	1/36	1/36	1/36	1/36	1/36	1/36
$s_b = 6$	1/36	1/36	1/36	1/36	1/36	1/36

$p(t = 9 \mid s_a, s_b)$	$s_a = 1$	$s_a = 2$	$s_a = 3$	$s_a = 4$	$s_a = 5$	$s_a = 6$
$s_b = 1$	0	0	0	0	0	0
$s_b = 2$	0	0	0	0	0	0
$s_b = 3$	0	0	0	0	0	1
$s_b = 4$	0	0	0	0	1	0
$s_b = 5$	0	0	0	1	0	0
$s_b = 6$	0	0	1	0	0	0

Two dice: Individual scores (cont.)

The complete model is explicitly defined

$$p(t, s_a, s_b) = p(t = 9|s_a, s_b)p(s_a)p(s_b) \tag{72}$$

	$s_a = 1$	$s_a = 2$	$s_a = 3$	$s_a = 4$	$s_a = 5$	$s_a = 6$
$s_b = 1$	0	0	0	0	0	0
$s_b = 2$	0	0	0	0	0	0
$s_b = 3$	0	0	0	0	0	1/36
$s_b = 4$	0	0	0	0	1/36	0
$s_b = 5$	0	0	0	1/36	0	0
$s_b = 6$	0	0	1/36	0	0	0

Two dice: Individual scores (cont.)

The posterior

$$p(s_a, s_b | t = 9) = \frac{p(t = 9 | s_a, s_b) p(s_a) p(s_b)}{p(t = 9)} \qquad (73)$$

	$s_a = 1$	$s_a = 2$	$s_a = 3$	$s_a = 4$	$s_a = 5$	$s_a = 6$
$s_b = 1$	0	0	0	0	0	0
$s_b = 2$	0	0	0	0	0	0
$s_b = 3$	0	0	0	0	0	1/4
$s_b = 4$	0	0	0	0	1/4	0
$s_b = 5$	0	0	0	1/4	0	0
$s_b = 6$	0	0	1/4	0	0	0

$$p(t = 9) = \sum_{s_a, s_b} p(t = 9 | s_a, s_b) p(s_a) p(s_b) = 4 \times 1/36 = 1/9 \qquad (74)$$

The posterior is given by equal mass in only 4 non-zero elements