Probabilistic reasoning

Artificial intelligence (CK0031)

Francesco Corona

Outline

1. Probability theory
   - Expectations and covariances
   - Bayesian probabilities
2. Reasoning under uncertainty
   - Probabilistic modelling
   - Probabilistic reasoning
   - Prior, likelihood and posterior

Probability theory

Probability theory

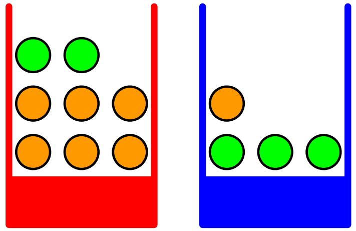
A key concept in the field of data science is that of **uncertainty**

- through noise on measurements
- through the finite size of data

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations of PRML

Probability theory (cont.)

We have two boxes, one red and one blue, and in the red box we have 2 apples and 6 oranges, and in the blue box we have 3 apples and 1 orange



We randomly select one box and from that box we randomly pick an item of fruit

- We check the fruit and we replace it in its box

We repeat this 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 theory (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 $\mathtt{r}$, for red box or $\mathtt{b}$, 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 $\mathtt{a}$, for apple or $\mathtt{o}$, for orange

Probability theory (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 selecting the red box is 4/10
- The probability of selecting the blue box is 6/10

We write these probabilities as $p(B = \mathtt{r}) = 4/10$ and $p(B = \mathtt{b}) = 6/10$

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

- If events are **mutually exclusive** and if they **include all possible outcomes**, then the **probabilities** for such events **must sum to one**

Probability theory (cont.)

Example

We have defined our experiment and we can start asking questions ...

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

Probability theory (cont.)

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

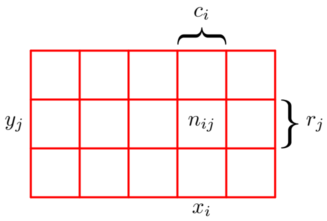
But, we need first the **two elementary rules of probability**

- the **sum rule** and the **product rule**

Probability theory (cont.)

To derive these rules, consider the slightly more general example

- **Two random variables** X and Y



We shall suppose that:

- X can take any of the values x_i, $i = 1, \dots, M$
- Y can take any of the values y_j, $j = 1, \dots, L$

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

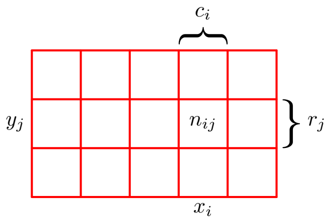
Consider a **total of N trials** in which we sample both variable X and Y

- 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 theory (cont.)

The probability that X will take the value x_i and Y will take the value y_j is written $p(X = x_i, Y = y_j)$

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



It is given by the number of points falling in the cell (i, j) as a fraction of the total number N of points

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

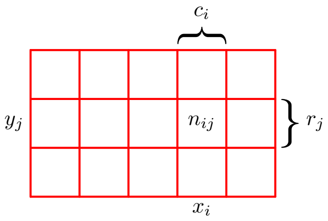
Implicitly, in the limit $N \to \infty$

Probability theory (cont.)

The probability that X takes the value x_i irrespective of the value of Y is $p(X = x_i)$ and is the fraction of the total number of points in column i

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

$p(X = x_i)$ is called the **marginal probability** because it is obtained by marginalising, or summing out, the other variables (i.e., Y here)



Definition

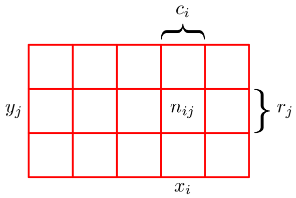
The marginal probability sets us for the **Sum rule** of probability

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

Probability theory (cont.)

If we consider only those instances for which $X = x_i$, then the fraction of such instances for which $Y = y_j$ is written $p(Y = y_j|X = x_i)$

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



It is given by the fraction of points in column i that fall in cell (i, j)

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

Definition

From Equation 1, 2 and 4, we derive the **Product rule** of probability

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

Probability theory (cont.)

Definition

The rules of probability

- **sum rule**

$$p(X) = \sum_Y p(X, Y) \qquad (6)$$

- **product rule**

$$p(X, Y) = p(Y|X)p(X) \qquad (7)$$

To compact notation, $p(\star)$ denotes a distribution over a RV $\star$ [1]

- $p(X, Y)$ 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 theory (cont.)

Definition

From the product rule and the symmetry property $p(X, Y) = p(Y, X)$, we obtain the following relationship between conditional probabilities

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

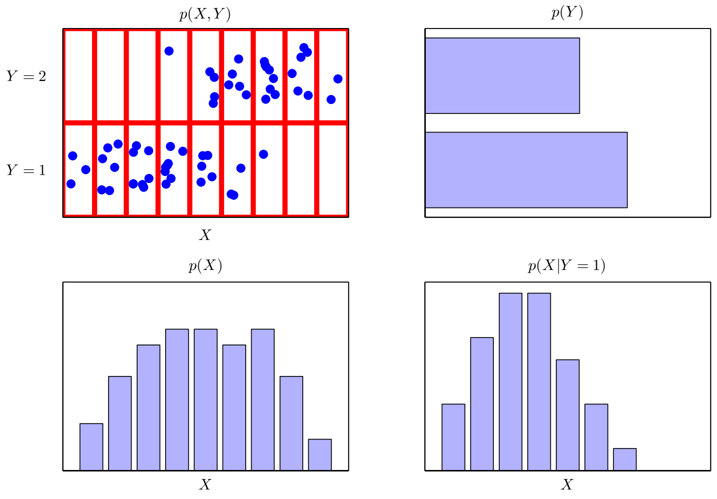
It is the **Bayes' rule**, plays a central role in reasoning under uncertainty

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator

$$p(X) = \sum_Y p(X|Y)p(Y) \qquad (9)$$

The denominator is a normalisation constant that ensures that the sum of the conditional probability $p(Y|X)$ over all values of Y is one

Probability theory (cont.)



Probability theory (cont.)

Example

Returning to the example involving the boxes of fruit

The probability of selecting either red or blue boxes are

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

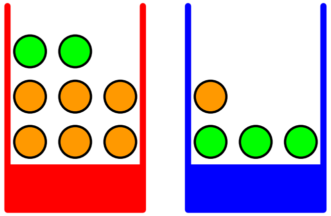
This satisfies $p(B = \mathtt{r}) + p(B = \mathtt{b}) = 4/10 + 6/10 = 1$

Now suppose that we pick a box at random, say the blue box

Then the probability of selecting an apple is just the fraction of apples in the blue box which is 3/4, so $p(F = \mathtt{a}|B = \mathtt{b}) = 3/4$

Probability theory (cont.)

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



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

Note that these probabilities are normalised

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

Probability theory (cont.)

We can now use the sum and product rules of probability to evaluate the overall probability of choosing an apple [2]

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

from which 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 theory (cont.)

Suppose instead we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from

We want the probability distribution over boxes conditioned on the identity of the fruit ($P(B|F)$)

The probabilities in Eq. 10-13 give the probability distribution over fruits conditioned on the identity of the box ($P(F|B)$)

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

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

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

Probabilistic reasoning

UFC/DC
AI (CK0031)
2016.2

Probability theory
Expectations and covariances
Bayesian probabilities

Reasoning under uncertainty
Probabilistic modelling
Probabilistic reasoning
Prior, likelihood and posterior

Probability theory (cont.)

We can provide an important interpretation of Bayes' theorem

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

- If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $p(B)$
- We call this the **prior probability** because it is the probability available before we observe the identity of the fruit

- Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $p(B|F)$
- We call this the **posterior probability** because it is the probability obtained after we have observed the identity of the fruit

Probabilistic reasoning

UFC/DC
AI (CK0031)
2016.2

Probability theory
Expectations and covariances
Bayesian probabilities

Reasoning under uncertainty
Probabilistic modelling
Probabilistic reasoning
Prior, likelihood and posterior

Probability theory (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), and once we observed that the selected fruit is an orange, the posterior probability of the red box is 2/3 (red is more probable)

Probabilistic reasoning

UFC/DC
AI (CK0031)
2016.2

Probability theory
Expectations and covariances
Bayesian probabilities

Reasoning under uncertainty
Probabilistic modelling
Probabilistic reasoning
Prior, likelihood and posterior

Probability theory (cont.)

If the joint distribution of two variables factorises into the product of the marginals, $p(X, Y) = p(X)p(Y)$, then X and Y are **independent**

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

From the product rule, $p(Y|X) = p(Y)$, and so the conditional distribution of Y given X is indeed independent of the X value

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

Probabilistic reasoning

UFC/DC
AI (CK0031)
2016.2

Probability theory
Expectations and covariances
Bayesian probabilities

Reasoning under uncertainty
Probabilistic modelling
Probabilistic reasoning
Prior, likelihood and posterior

Expectations and covariances

Probability theory

Expectations and covariances

One operation involving probabilities is **weighted averages of functions**

- The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the **expectation** of $f(x)$
- It will be denoted by $\mathbb{E}[f]$

Definition

For a discrete distribution, the expectation of $f(x)$ is given by the form

$$\mathbb{E}[f] = \sum_x p(x) f(x) \tag{18}$$

The average is weighted by the relative probabilities of the values of x

Expectations and covariances (cont.)

In either case, given a finite number N of observations drawn from the probability distribution or probability density $p(x)$, then the expectation of function $f(x)$ can be approximated as a finite sum over these points

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n) \tag{19}$$

The approximation becomes exact in the limit $N \rightarrow \infty$

Expectations and covariances

Sometimes we consider expectations of functions of several variables

- We use a subscript to indicate which variable is being averaged over

$\mathbb{E}_x[f(x,y)]$ is the average of function $f(x,y)$ wrt the distribution of x

- $\mathbb{E}_x[f(x,y)] = \sum_x p(x) f(x,y)$
- $\mathbb{E}_x[f(x,y)]$ is a function of y

Definition

The **conditional expectation** wrt a conditional distribution

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \tag{20}$$

Expectations and covariances (cont.)

Definition

The measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$ is called the **variance** of $f(x)$ and it is defined by

$$\text{var}[f] = \mathbb{E}\Big[\Big(f(x) - \mathbb{E}[f(x)]\Big)^2\Big] \tag{21}$$

Expanding the square, we can show ($\star$) that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{22}$$

- The variance of the variable x itself (i.e., $f(x) = x$) is

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{23}$$

Expectations and covariances (cont.)

Definition

For two random variables x and y, the extent to which x and y vary together is called **covariance** and it is defined by

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{xy}\Big[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\Big] \\ &= \mathbb{E}_{xy}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \qquad (24)$$

If x and y are independent, then their covariance vanishes ($\star$)

For two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$, the covariance is a matrix

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\Big[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T])\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned} \qquad (25)$$

Bayesian probabilities

Probability theory

Bayesian probabilities

We viewed probabilities as frequencies of repeatable random events

- It is the **frequentist** interpretation of probability

We can view probabilities also as quantification of uncertainty

- It is the **Bayesian** interpretation of probability

Example

In the boxes of fruit the observation of the identity of the fruit provided relevant information that altered the probability of the chosen box

- Bayes's rule converted a prior probability ($P(B = r) = 4/10$) into a posterior probability by incorporating evidence from observed data

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{2}{3}$$

Bayesian probabilities (cont.)

We can adopt this approach when making inference about quantities such as the parameters $\mathbf{w}$ in a regression or classification example

- We first capture our assumptions about $\mathbf{w}$, before observing the data in the form of a prior probability $p(\mathbf{w})$
- The effect of the observed data $\mathcal{D} = \{t_1, \dots, t_n\}$ is expressed through the conditional probability $p(\mathcal{D}|\mathbf{w})$
- Then, we evaluate the uncertainty in $\mathbf{w}$, after we observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \qquad (26)$$

The quantity $p(\mathcal{D}|\mathbf{w})$ is evaluated for the observed $\mathcal{D}$ and can be viewed as a function of the parameter(s) $\mathbf{w}$, as such it is a **likelihood function**

- How probable $\mathcal{D}$ is for different settings of the parameters $\mathbf{w}$

Bayesian probabilities (cont.)

The likelihood function $p(\mathcal{D}|\mathbf{w})$ plays a fundamental role

- In a frequentist setting, $\mathbf{w}$ is considered as a fixed parameter, whose value is determined by some form of *estimator*, and error bars on such an estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$
- In the Bayesian setting, there is only a single data set $\mathcal{D}$ (the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$ given that set

Remark

The likelihood $p(\mathcal{D}|\mathbf{w})$ is NOT a probability distribution

Bayesian probabilities (cont.)

A widely used frequentist estimator is **maximum likelihood**, in which $\mathbf{w}$ is set to the value that maximises the likelihood function $p(\mathcal{D}|\mathbf{w})$

- This corresponds to choosing the value of $\mathbf{w}$ for which the probability of the observed data set $\mathcal{D}$ is maximised

Definition

The negative log of the likelihood function is called an **error function**

- The negative logarithm is a monotonically decreasing function, maximising the likelihood is equivalent to minimising the error

Bayesian probabilities (cont.)

Definition

Given a definition of likelihood, we state Bayes' rule also in words

posterior $\propto$ **likelihood** $\times$ **prior**

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{27}$$

All quantities are intended as functions of $\mathbf{w}$ and the denominator is a normalisation constant ensuring that the posterior is a valid p(d/m)f

Integrating both sides of Bayes' rule with respect to $\mathbf{w}$, we can express the denominator in terms of prior distribution and likelihood function

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{28}$$

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 the calligraphic symbol

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

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

Example

States will typically be represented using typewriter type fonts

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

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

- If we are discussing an experiment about a coin c, the meaning of $p(\texttt{heads})$ is clear from context, being shorthand for $p(c = \texttt{heads})$

Probabilistic modelling (cont.)

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

- $\sum_x f(x) = \sum_{\texttt{s} \in \text{dom}(x)} f(x = \texttt{s})$

Given variable x, its domain $\text{dom}(x)$ and a full specification of the probability values for each of the states, $p(x)$

- We say that 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 $p(\textit{The coin is tails} = \texttt{true})$

We use $p(x = \texttt{tr})$ for the probability of event/variable x being in state $\texttt{true}$ and $p(x = \texttt{fa})$ for the probability of x being in state $\texttt{false}$

Probabilistic modelling (cont.)

Definition

Rules of probability for discrete variables (1): Probability $p(x = \mathtt{x})$ of variable x being in state $\mathtt{x}$ is represented by a value between 0 and 1

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

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

Definition

Rules of probability for discrete variables (2): The summation of the probability over all states is one:

$$\sum_{\mathtt{x} \in \text{dom}(x)} p(x = \mathtt{x}) = 1 \tag{29}$$

Normalisation condition: Often written as $\sum_x p(x) = 1$

Probabilistic modelling (cont.)

Definition

Rules of probability for discrete variables (3): x and y can interact

$$p(x = \mathtt{a} \text{ or } y = \mathtt{b}) = p(x = \mathtt{a}) + p(y = \mathtt{b}) - p(x = \mathtt{a} \text{ and } y = \mathtt{b}) \tag{30}$$

Or, more generally we write

$$p(x \text{ or } y) = p(x) + p(y) - p(x \text{ and } y) \tag{31}$$

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

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

Probabilistic modelling (cont.)

Definition

Set notation: An alternative notation in terms of set theory is

$$\begin{aligned} p(x \text{ or } y) &\equiv p(x \cup y) \\ p(x, y) &\equiv p(x \cap y) \end{aligned} \tag{32a}$$

Probabilistic modelling (cont.)

Definition

Marginals: Given a **joint distribution** $p(x, y)$, the distribution of a single variable is given by

$$p(x) = \sum_y p(x, y) \tag{33}$$

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

Marginalisation: Process of computing a marginal from a joint distro

More generally, one has

$$p(x_1, \dots, x_{i-1}, x_{x+1}, \dots, x_n) = \sum_{x_i} p(x_1, \dots, x_n) \tag{34}$$

Probabilistic modelling (cont.)

Definition

Conditional probability/**Bayes' rule**: The probability of some event x conditioned on knowing some event y, or the probability of x given y

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{35}$$

If $p(y) = 0$, then $p(x|y)$ is not defined

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

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{36}$$

Bayes' rule trivially follows from the definition of conditional probability, we can be loose in our language and use the terms as synonymous

- 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 here, apart from pointing out that it is not necessarily the rules of probability that are contentious, rather what interpretation we should place on them

If potential repetitions of an experiment can be envisaged, then the frequentist definition of probability in which probabilities are defined wrt a potentially infinite repetition of experiments makes sense

In coin tossing, the probability of heads might be interpreted as

- '*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

- If we define probability as a limiting case of infinite repetitions of the same experiment, this wouldn't make much sense in this case (we cannot repeat the experiment)
- If we assume that the user behaves in a manner that is consistent with other users, we should be able to exploit the large amount of data from other users' ratings to make a reasonable 'guess' as to what this consumer likes

Probabilistic modelling - Conditional probability

A **degree of belief** or **Bayesian** subjective interpretation of probability sidesteps non-repeatability issues

- It is just a framework for manipulating real values consistent with our intuition about probability

Probabilistic modelling - Conditional probability

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 and we want to know what is the probability that the dart thrower has hit the 5-region?

Probabilistic modelling - Conditional probability (cont.)

- Conditioned on this info, only regions 1 to 19 remain possible and 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 point to clarify is that $p(A = \mathrm{a}|B = \mathrm{b})$ should NOT be interpreted as

- '*Given the event $B = b$ has occurred, $p(A = a|B = b)$ is the probability of the event $A = a$ occurring*'

As, in most contexts, no such explicit temporal causality can be implied

The correct interpretation should be: '*$p(A = a|B = b)$ is the probability of A being in state a under the constraint that B is in state b*'

Probabilistic modelling - Conditional probability (cont.)

The relation between the conditional $p(A = \mathrm{a}|B = \mathrm{b})$ and the joint $p(A = \mathrm{a}, B = \mathrm{b})$ is just a normalisation constant

- $p(A = \mathrm{a}, B = \mathrm{b})$ is not a distribution in A
- In other words, $\sum_{\mathrm{a}} p(A = \mathrm{a}, B = \mathrm{b}) \neq 1$

To make it a distribution we need to divide:

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

which, when summed over a does sum to 1

This is just the definition of $p(A = \mathrm{a}|B = \mathrm{b})$

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

$$p(x, y) = p(x)p(y) \quad (37)$$

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

$$p(x|y) = p(x) \iff p(y|x) = p(y) \quad (38)$$

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

If for some constant k and some positive functions $f(\cdot)$ and $g(\cdot)$

$$p(x, y) = kf(x)g(y) \quad (39)$$

then we say that x and y are independent and 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 and let y be the day in which males are born, $\text{dom}(x) = \text{dom}(y) = \{\texttt{M}, \texttt{T}, \ldots, \texttt{S}\}$

- It is reasonable to expect that x is independent of y

We randomly select a woman from the phone book (Alice) and find out that she was born on a Tuesday and we randomly select a male (Bob)

- Before phoning Bob and asking him, what does knowing Alice's birth day add to which day we think Bob is born on?

Under the independence assumption, the answer is nothing

Probabilistic modelling - Conditional probability (cont.)

This doesn't mean that the distribution of Bob's birthday is uniform

- It means that knowing when Alice was born doesn't provide any extra information than we already knew about Bob's birthday

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

It is known that 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 is perhaps a little strange

Example

Consider binary variables (domains consist of 2 states) x and y and define the distribution st x and y are always both in a certain state

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

Are x and y dependent?

Probabilistic modelling - Conditional probability (cont.)

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

- $p(x)p(y) = p(x, y)$ for all states of x and y
- x and y are thus independent

This may seem strange, as we know that the relation between x and y is such that they are always in the same joint state and yet independent

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

- ‘knew before’ means working 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

$$p(\mathcal{X}, \mathcal{Y} | \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z}) p(\mathcal{Y} | \mathcal{Z}), \qquad (40)$$

provided that we know the state of set $\mathcal{Z}$, and we write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$

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 ⇒ conditionally dependent

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z} \qquad (41)$$

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 this means that

- given z, y contains no additional information about x, and similarly
- given z, knowing x does not tell anything more about y

Remark

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

Probability refresher - Conditional probability (cont.)

Remark

Independence implications: Tempting to think that if ‘a is independent of b’ and ‘b is independent of c’, then ‘a must be independent of c’

$$\{a \perp\!\!\!\perp b, b \perp\!\!\!\perp c\} \implies a \perp\!\!\!\perp c \qquad (42)$$

However, this does NOT follow

Consider a distribution of the form

$$p(a, b, c) = p(b)p(a, c) \qquad (43)$$

From this

$$p(a, b) = \sum_c p(a, b, c) = p(b) \sum_c p(a, b) \qquad (44)$$

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

- so that a and b are independent

Similarly, one can show that b and c are independent and that a is not necessarily independent of c (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’

$$\{b \top\!\!\top b, b \top\!\!\top c\} \implies a \top\!\!\top c \qquad (45)$$

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

Remark

Finally, note that conditional independence $x \perp\!\!\!\perp y | z$ does not imply marginal independence $x \perp\!\!\!\perp y$ ($\star$)

Probabilistic modelling - Probability tables

Based on populations 60776238, 5116900 and 2980700 of countries (CNT) England (E), Scotland (S) and Wales (W), *a priori* probability that a randomly selected person from the combined three countries would live in England, Scotland or Wales is 0.88, 0.08 and 0.04

$$\begin{pmatrix} p(CNT = \texttt{E}) \\ p(CNT = \texttt{S}) \\ p(CNT = \texttt{W}) \end{pmatrix} = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix} \qquad (46)$$

whose component values sum to 1 and the ordering is arbitrary

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

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

with conditional probabilities $p(MT|CNT)$ by residence E, S and W

Probability modelling - Probability tables (cont.)

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

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

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

Probability modelling - Probability tables (cont.)

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

Probabilistic modelling - Probability tables (cont.)

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

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

We can also 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 $p(CNT)$,

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

$$\begin{pmatrix} p(CNT = \mathtt{E}) \\ p(CNT = \mathtt{S}) \\ p(CNT = \mathtt{W}) \end{pmatrix} = \begin{pmatrix} 0.8352 + 0.0352 + 0.0088 = 0.88 \\ 0.0352 + 0.0240 + 0.0000 = 0.08 \\ 0.0088 + 0.0000 + 0.0160 = 0.04 \end{pmatrix} \tag{47}$$

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

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

$$\begin{pmatrix} p(MT = \mathtt{Eng}) \\ p(MT = \mathtt{Scot}) \\ p(MT = \mathtt{Wel}) \end{pmatrix} = \begin{pmatrix} 0.8360 + 0.0560 + 0.0240 = 0.916 \\ 0.0352 + 0.0240 + 0.0000 = 0.059 \\ 0.0088 + 0.0000 + 0.0160 = 0.025 \end{pmatrix} \tag{48}$$

Probabilistic modelling - Probability tables (cont.)

One can infer the conditional $p(CNT|MT) \propto p(MT|CNT)p(CNT)$

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

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

Probabilistic modelling - Probability tables (cont.)

For joint distributions over a larger set of variables $\{x_i\}_{i=1}^{D}$ with each variable x_i taking K_i states, the table describing the joint distro is an array with $\prod_{i=1}^{D} K_i$ entries

- Explicitly storing tables requires space exponential in the number of variables (rapidly becomes impractical for a large number D)

Remark

A probability 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

- in each case, the joint setting of variables is a different index to the same probability

Probability reasoning

Reasoning under uncertainty

Probabilistic reasoning

The central paradigm of probabilistic reasoning is to identify all relevant variables $x_1, \ldots, x_N$ in the environment, and make a **probabilistic model**

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

- **Reasoning** (or **inference**) is performed by introducing **evidence** that sets variables in known states, and subsequently computing probabilities of interest, conditioned on this evidence

The rules of probability, combined with Bayes' rule make for a complete reasoning system, one which includes deductive logic as a special case

We now discuss some examples in which the number of variables is still very small, and soon we discuss reasoning in networks of many variables

- There, a graphical notation will play a central role

Probabilistic reasoning (cont.)

Example

Hamburgers and the KJ disease: Consider this (factious) scenario

- Doctors found that people with Kreuzfel-Jacob disease (KJ) almost inevitably ate hamburgers $p(\mathit{Hamburger\ eater} = \texttt{tr} | \mathit{KJ} = \texttt{tr}) = 0.9$

The probability of a person having KJ is very low $p(\mathit{KJ} = \texttt{tr}) = \dfrac{1}{100K}$

Assuming eating hamburgers is spread $p(\mathit{Hamburger\ eater} = \texttt{tr}) = 0.5$, what is the probability that a hamburger eater will have KJ disease?

$$\begin{aligned} p(\texttt{KJ}|\texttt{Hamburger eater}) &= \frac{p(\texttt{Hamburger eater}, \texttt{KJ})}{p(\texttt{Hamburger eater})} \\ &= \frac{p(\texttt{Hamburger eater}|\texttt{KJ})p(\texttt{KJ})}{p(\texttt{Hamburger eater})} \\ &= \frac{9/10 \times 1/100K}{1/2} = 1.8 \times 10^{-5} \end{aligned} \tag{49}$$

If $p(\texttt{Hamburger eater}) = 0.001$, $p(\texttt{KJ}|\texttt{Hamburger eater}) \approx 1/100$

Example

Inspector Clouseau: Inspector Clouseau arrives at the scene of a crime

- The victim lies dead near the possible murder weapon, a knife (K, such that $\text{dom}(K) = \{\texttt{knife used}, \texttt{knife not used}\}$)

The butler (B) and the maid (M) are the inspector's main suspects (B and M, st $\text{dom}(B) = \text{dom}(M) = \{\texttt{murderer}, \texttt{not murderer}\}$)

Prior beliefs that they are the murderer quantifies as follows

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

These beliefs are independent ($p(B)p(M) = p(B, M)$) and it is 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, $p(K, B, M) = p(K|B, M)p(B)p(M)$

Probabilistic reasoning (cont.)

Assuming that the knife is the murder weapon ($K = \texttt{tr}$), what is the probability that the butler is the murderer, $p(B = \texttt{murderer}|K = \texttt{tr})$?

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

$$\begin{aligned} p(B|K) &= \sum_{M \in m} p(B, M|K) = \sum_{M \in m} \frac{p(B, M, K)}{p(K)} = \frac{1}{p(K)} \sum_{M \in m} p(B, M, K) \\ &= \frac{1}{\sum_{\substack{B \in b \\ M \in m}} p(K|B, M)p(B, M)} \sum_{M \in m} p(K|B, M)p(B, M) \\ &= \frac{1}{\sum_{B \in b} p(B) \sum_{M \in m} p(K|B, M)p(M)} p(B) \sum_{M \in m} p(K|B, M)p(M) \end{aligned} \tag{50}$$

where we used the fact that in our model $p(B, M) = p(B)p(M)$

Probabilistic reasoning(cont.)

Plugging in the values we have that $p(B = \texttt{murderer}|K = \texttt{knife used})$

$$\begin{aligned} &= \frac{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10} \times \frac{1}{10} + \frac{8}{10} \times \frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10} \times \frac{2}{10} + \frac{8}{10} \times \frac{3}{10}\right)} \\ &= \frac{300}{412} \simeq 0.73 \end{aligned} \tag{51}$$

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

Remark

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

$$p(K = \texttt{knife used}) = \sum_{B \in b} p(B) \sum_{M \in m} p(K = \texttt{knife used} | B, M) p(M) \\ = 0.412 \tag{52}$$

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

Quantity $p(K = \texttt{knife used})$ relates to the **prior** probability the model assigns to the knife being used (in the absence of any other info)

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

$$p(K = \texttt{knife used} | K = \texttt{knife used}) = \\ \frac{p(K = \texttt{knife used}, K = \texttt{knife used})}{p(K = \texttt{knife used})} = \\ \frac{p(K = \texttt{knife used})}{p(K = \texttt{knife used})} = 1 \tag{53}$$

Probabilistic reasoning (cont.)

Example

Who's in the bathroom? A household of 3 perosons: 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 and (knowing that only either Bob or Alice can be in the bathroom), he infers that Bob must be occupying it

To arrive at the same conclusion mathematically, define the events

$$\begin{cases} A: & \text{Alice is in her bedroom} \\ B: & \text{Bob is in his bedroom} \\ O: & \text{Bathroom is occupied} \end{cases} \tag{54}$$

Probabilistic reasoning (cont.)

We encode the information that if either Alice or Bob are not in their bedrooms, then they must be in the bathroom (both may be there) as

$$p(O = \texttt{tr} | A = \texttt{fa}, B) = 1 \\ p(O = \texttt{tr} | B = \texttt{fa}, A) = 1 \tag{55}$$

- 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.)

$$p(B = \texttt{fa} | O = \texttt{tr}, A = \texttt{tr}) = \\ \frac{p(B = \texttt{fa}, O = \texttt{tr}, A = \texttt{tr})}{p(O = \texttt{tr}, A = \texttt{tr})} = \\ \frac{\underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa})}_{1} p(A = \texttt{tr}, B = \texttt{fa})}{p(O = \texttt{tr}, A = \texttt{tr})} \tag{56}$$

$$p(O = \texttt{tr}, A = \texttt{tr}) = \\ \underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa})}_{1} p(A = \texttt{tr}, B = \texttt{fa}) + \\ \underbrace{p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{tr})}_{0} p(A = \texttt{tr}, B = \texttt{tr}) \tag{57}$$

- $p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{fa}) = 1$: If Alice is in her room and Bob is not, the bathroom must be occupied
- $p(O = \texttt{tr} | A = \texttt{tr}, B = \texttt{tr}) = 0$: If both Alice and Bob are in their rooms, the bathroom cannot be occupied

Probabilistic reasoning (cont.)

$$p(B = \mathtt{fa}|O = \mathtt{tr}, A = \mathtt{tr}) = \frac{p(A = \mathtt{tr}, A = \mathtt{fa})}{p(A = \mathtt{tr}, B = \mathtt{fa})} = 1 \qquad (58)$$

Remark

The example is interesting since we are not required to make a full probabilistic model, we don't need to specify $p(A, B)$)

- The situation is common in limiting situations of probabilities being either 0 or 1, corresponding to traditional logic systems

Probabilistic reasoning (cont.)

Example

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

We want to show that this implies one of the two

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

Assuming that $p(A = \mathtt{tr}) > 0$, these are equivalent to

- $p(T = \mathtt{fa}, A = \mathtt{tr}) = 0$

Probabilistic reasoning (cont.)

$$p(T = \mathtt{fa}, A = \mathtt{tr}) = \\ p(T = \mathtt{fa}, A = \mathtt{tr}, F = \mathtt{tr}) + p(T = \mathtt{fa}, A = \mathtt{tr}, F = \mathtt{fa}) \qquad (59)$$

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

- $$p(T = \mathtt{fa}, A = \mathtt{tr}, F = \mathtt{tr}) \leq \\ p(T = \mathtt{fa}, F = \mathtt{tr}) = p(T = \mathtt{fa}|F = \mathtt{tr})p(F = \mathtt{tr}) = 0, \qquad (60)$$

  since $p(T = \mathtt{fa}|F = \mathtt{tr}) = 1 - p(T = \mathtt{tr}|F = \mathtt{tr}) = 1 - 1 = 0$

- $$p(T = \mathtt{fa}, A = \mathtt{tr}, F = \mathtt{fa}) \leq \\ p(A = \mathtt{tr}, F = \mathtt{fa}) = p(F = \mathtt{fa}|A = \mathtt{tr})p(A = \mathtt{tr})) = 0, \qquad (61)$$

  where again, by assumption $p(F = \mathtt{fa}|A = \mathtt{tr}) = 0$

Probabilistic reasoning (cont.)

Example

Aristotle - 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

- $p(B = \mathtt{tr}|A = \mathtt{tr}) = 1$, corresponds to '*If A is true then B is true*'

We may infer

$$p(A = \mathtt{fa}|B = \mathtt{fa}) = 1 - p(A = \mathtt{tr}|B = \mathtt{fa}) = \\ 1 - \frac{p(B = \mathtt{fa}|A = \mathtt{tr})p(A = \mathtt{tr})}{p(B = \mathtt{fa}|A = \mathtt{tr})p(A = \mathtt{tr}) + p(B = \mathtt{fa}|A = \mathtt{fa})p(A = \mathtt{fa})} = 1 \qquad (62)$$

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	0	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

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	0	0.25

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

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

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

Probabilistic reasoning (cont.)

$$
\begin{aligned}
p(A = 1, C = 0) = & \sum_B p(A = 1, B, C = 0) \\
= & \sum_B p(C = 0|A = 1, B) p(A = 1) p(B) \\
= & p(A = 1) p(C = 0|A = 1, B = 0) p(B = 0) + \\
& p(A = 1) p(C = 0|A = 1, B = 1) p(B = 1) \\
= & 0.65 \times (0.2 \times 0.23 + 0.75 \times 0.77) = 0.405275
\end{aligned} \qquad (63)
$$

$$
\begin{aligned}
p(A = 0, C = 0) = & \sum_B p(A = 0, B, C = 0) \\
= & \sum_B p(C = 0|A = 0, B) p(A = 0) p(B) \\
= & p(A = 0) p(C = 0|A = 0, B = 0) p(B = 0) + \\
& p(A = 1) p(C = 0|A = 0, B = 1) p(B = 1) \\
= & 0.35 \times (0.9 \times 0.23 + 0.01 \times 0.77) = 0.075145
\end{aligned} \qquad (64)
$$

Probabilistic reasoning (cont.)

$$
\begin{aligned}
p(A = 1|C = 0) = & \frac{p(A = 1, C = 0)}{p(A = 1, C = 0) + p(A = 0, C = 0)} \\
= & \frac{0.405275}{0.405275 + 0.075145} \\
= & 0.8436
\end{aligned} \qquad (65)
$$

Probabilistic reasoning (cont.)

Example

Larry: Larry is typically late for school, but when his mother asks whether or not he was late for school, he never admits to being late

If Larry is late, we denote this with $L = \texttt{late}$, otherwise, $L = \texttt{not late}$

The response Larry gives is denoted by R_L and it is represented as

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

The remaining two values are determined by normalisation and are

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

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, we have

$$p(L=\texttt{late}|R_L=\texttt{not late}) = \frac{p(L=\texttt{late},R_L=\texttt{not late})}{p(R_L=\texttt{not late})}$$
$$= \frac{p(L=\texttt{late},R_L=\texttt{not late})}{p(L=\texttt{late},R_L=\texttt{not late})+p(L=\texttt{not late},R_L=\texttt{not late})} \quad (66)$$

In the above, we recognise

$$p(L=\texttt{late},R_L=\texttt{not late}) = \underbrace{p(R_L=\texttt{not late}|L=\texttt{late})}_{1}\,p(L=\texttt{late}) \quad (67)$$

$$p(L=\texttt{not late},R_L=\texttt{not late}) = \underbrace{p(R_L=\texttt{not late}|L=\texttt{not late})}_{1}\,p(L=\texttt{not late}) \quad (68)$$

Probabilistic reasoning (cont.)

$$p(L=\texttt{late}|R_L=\texttt{not late}) = \frac{p(L=\texttt{late})}{p(L=\texttt{late})+p(L=\texttt{not late})} \quad (69)$$
$$= p(L=\texttt{late})$$

The result is intuitive, 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 $p(R_S,R_L|L)=p(R_S|L)p(R_L|L)$ and then we write

$$p(R_S,R_L,L)=p(R_L|L)p(R_S|L)p(L) \quad (70)$$

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

Probabilistic reasoning (cont.)

Using Bayes' rule, we have

$$p(L=\texttt{late}|R_L=\texttt{nlate},R_S=\texttt{late}) =$$
$$\frac{1}{Z}p(R_S=\texttt{late}|L=\texttt{late})p(R_L=\texttt{nlate}|L=\texttt{late})p(L=\texttt{late}) \quad (71)$$

where the normalisation term $1/Z$ is given by

$$\frac{1}{Z}=p(R_S=\texttt{late}|L=\texttt{late})p(R_L=\texttt{nlate}|L=\texttt{late})p(L=\texttt{late})$$
$$+\,p(R_S=\texttt{late}|L=\texttt{nlate})p(R_L=\texttt{nlate}|L=\texttt{nlate})p(L=\texttt{nlate}) \quad (72)$$

Hence,

$$p(L=\texttt{late}|R_L=\texttt{not late},R_S=\texttt{late}) =$$
$$\frac{1\times 1\times p(L=\texttt{late})}{1\times 1\times p(L=\texttt{late})+0\times 1\times p(L=\texttt{not late})}=1 \quad (73)$$

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

Probabilistic reasoning (cont.)

Example

Luke: Luke has been told he's lucky and has won a prize in the lottery

- 5 prizes available: 10 (p_1); 100 (p_2); 1K (p_3); 10K (p_4); 1M (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*' answers 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) and $W = 0$ for no prize (0)

$$\begin{aligned} p(W=3|W\neq 5, W\neq 4, W\neq 0) &= \frac{p(W=3, W\neq 5, W\neq 4, W\neq 0)}{p(W\neq 5, W\neq 4, W\neq 0)} \\ &= \frac{p(W=3)}{\underbrace{p(W=1 \text{ or } W=2 \text{ or } W=3)}_{\text{events are mutually exclusive}}} \\ &= \frac{p_3}{p_1+p_2+p_3} \end{aligned} \qquad (74)$$

The results makes intuitive sense: Once removing the impossible states of W, the probability to win 1K is proportional to its prior probability (p_3), with normalisation being 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 $\mathcal{D}$ and ii) I have some knowledge of the data generating mechanism

Our interest is then the quantity

$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\Theta)p(\Theta)}{\int_\Theta p(\mathcal{D}|\Theta)p(\Theta)} \qquad (75)$$

From **generative model** $p(\mathcal{D}|\Theta)$ of the dataset, and coupled with 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

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

For 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 observed data)

Prior, likelihood and posterior (cont.)

The use of the generative model sits well with physical modelling

- We typically postulate how to generate observed phenomena, assuming we know the model

One might postulate how to generate a time-series of displacements for a swinging pendulum of unknown mass, length and dumping constant

- Using the generative model, and given only the displacements, we could infer the unknown physical properties of the pendulum

Prior, likelihood and posterior (cont.)

Example

Pendulum: Consider a pendulum, x_t is the angular displacement at t

Assuming that measurements are independent, given the knowledge of the problem parameter Θ, the likelihood of a sequence $x_1, \ldots, x_T$ is

$$p(x_1, \ldots, x_T|\Theta) = \prod_{t=1}^{T} p(x_t|\Theta) \qquad (76)$$

- If we assume that the model is correct and our measurement of the displacement x is perfect, then the physical model is

  $$x_t = \sin(\Theta t), \qquad (77)$$

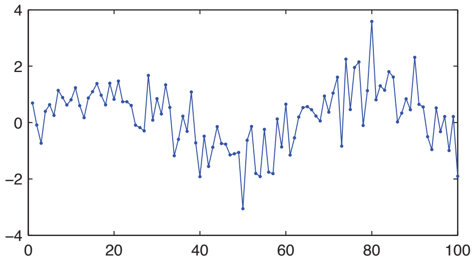
  where Θ is the unknown constants of the pendulum ($\sqrt{g/L}$, g is the gravitational attraction and L the pendulum length)
- If we assume that we have a poor instrument to measure the displacements, with some known variance σ^2, then

  $$x_t = \sin(\Theta t) + \varepsilon_t \qquad (78)$$

  where ε_t is zero mean Gaussian noise with variance σ^2

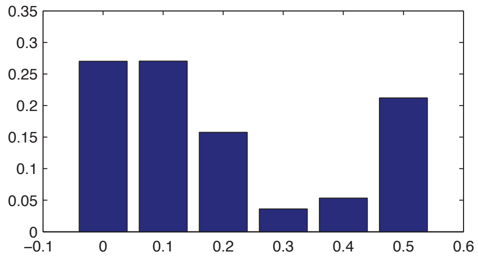
Prior, likelihood and posterior (cont.)

Noisy observations of displacements $x_1, \ldots, x_{100}$



Prior, likelihood and posterior (cont.)

We can also consider a set of possible parameters Θ and place a prior $p(\Theta)$ over them, expressing our prior belief (before even seeing the measurements) in the appropriateness of the different values of Θ



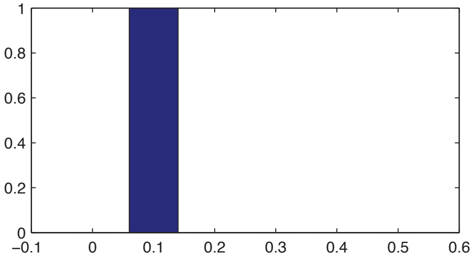
The prior belief on 5 possible values of Θ

Prior, likelihood and posterior (cont.)

The posterior distribution is then given by

$$p(\Theta|x_1,\dots,x_N) \propto p(\Theta)\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x_t-\sin(\Theta t))\right) \qquad (79)$$

Despite noisy measurements, the posterior over the assumed values of Θ becomes strongly peaked for a large number of 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 $\text{dom}(s_a) = \{1,2,3,4,5,6\}$ and similarly for die b we denote s_b and we let $\text{dom}(s_b) = \{1,2,3,4,5,6\}$

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

$$p(t, s_a, s_b) = \underbrace{p(t|s_a, s_b)}_{\text{likelihood}}\underbrace{p(s_a, s_b)}_{\text{prior}} \qquad (80)$$

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, and assuming no dependency in the rolling

  $$p(s_a, s_b) = p(s_a)p(s_b) \qquad (81)$$

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

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

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

  $$p(t|s_a, s_b) = \mathbb{I}[t = s_a + s_b] \qquad (82)$$

  Function $\mathbb{I}[A]$ is st $\mathbb{I}[A] = 1$ 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.)

Our complete model is explicitly defined using

$$p(t, s_a, s_b) = p(t = 9 \mid s_a, s_b)p(s_a)p(s_b) \tag{83}$$

	$s_a = 1$	$s_a = 2$	$s_a = 3$	$s_a = 4$	$s_a = 5$	$s_a = 6$
$s_b = 1$	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 is given by

$$p(s_a, s_b \mid t = 9) = \frac{p(t = 9 \mid s_a, s_b)p(s_a)p(s_b)}{p(t = 9)} \tag{84}$$

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

$$p(t = 9) = \sum_{s_a, s_b} p(t = 9 \mid s_a, s_b)p(s_a)p(s_b) = 4 \times 1/36 = 1/9 \tag{85}$$

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