IMPERIAL

Probability and Information Theory

Payam Barnaghi
Department of Brain Sciences &
School of Convergence Science (Human and Artificial Intelligence)
Imperial College London
January 2026

ML4NS Machine Learning for Neuroscience

1

IMPERIAL

Probability

– The probability of an event is the fraction of times that the event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity.
– By definition, probabilities must be between $[0,1]$

– The probability that X will take the value x_i and Y will take the value y_j is written as:

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

Where n_{ij} is the number of points that x_i, y_j occur together.

2

2

Mathematical notations

IMPERIAL

- In machine learning papers and books, you will come across several mathematical notations describing different concepts.
- Throughout this series, we will present (some of) the concepts in mathematical form to help you become more familiar with notations and how to read and interpret them.

3

3

Probability

IMPERIAL

- The probability that X takes the value x_j irrespective of the value of Y is written as p(x=X)

$$p(X = x_i) = \frac{c_i}{N}$$

4

4

Example

IMPERIAL

ATG GCT AAC TGG CCA
ATG TTT GGA TAC CAG
ATG ACC GTC CTT AGG
ATG GAA GCT TGC TAA
ATG CCA TGG AAC CTC
ATG AGT GGC TTG TGA
ATG TCG AGG CCA ACC
ATG GGT TTT CAT TAG
ATG ACG CGA TCC GGT
ATG TAC GGC AGT TAA

5

5

Example- How do you represent these as probability?

IMPERIAL

- Alzheimer's disease is most common in people over the age of 65.
- The risk of Alzheimer's disease and other types of dementia increases with age, affecting an estimated 1 in 14 people over the age of 65 and 1 in every 6 people over the age of 80.
- But around 1 in every 20 people with Alzheimer's disease is under the age of 65. This is called early- or young-onset Alzheimer's disease.

Source of the statistics: NHS, https://www.nhs.uk/conditions/alzheimers-disease/

6

6

Random variable

IMPERIAL

- A random variable is a variable that can take on different values randomly.
- Random variables may be discrete or continuous. A discrete random variable has a finite or countably infinite number of states.
- Note that these states are not necessarily integers; they can also just be named states that are not considered to have any numerical value.

7

7

Probability distribution

IMPERIAL

- A probability distribution describes how likely a random variable or set of random variables is to take on each of its possible states.
- How we describe probability distributions depends on whether the variables are discrete or continuous.

8

8

Probability

IMPERIAL

- The expression $p(A)$ denotes the probability that the event A is true.
- For example, A might be the logical expression "A patient has Alzheimer's disease".
- We require that $0 \leq p(A) \leq 1$,

Where $p(A) = 0$ means the person definitely does not have the condition, and $p(A) = 1$ means the patient definitely has the condition.

9

9

p($\bar{A}$)

IMPERIAL

- We write $p(\bar{A})$ to denote the probability of the event not A; this is defined as: $p(\bar{A}) = 1 - p(A)$.
- We will often write $A = 1$ to mean the event A is true, and $A = 0$ to mean the event A is false.

10

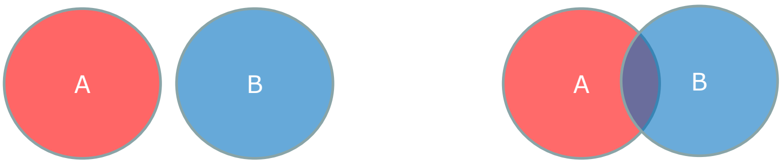
10

Probability of a union of two events

IMPERIAL

– Given two events, A and B, we define the probability of A or B as follows:

$$\begin{aligned} p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\ &= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned}$$



11

11

Joint probabilities

IMPERIAL

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

– This is sometimes called the product rule. Given a joint distribution on two events $p(A, B)$, we define the marginal distribution as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

12

12

Marginalisation

IMPERIAL

- Let's assume that you want to compute $P(X = x)$, but we are not given the direct probability distribution over X.
- We are instead given a joint probability distribution over X and some other random variable(s) Y.
- In this case, we can just say:

$$p(X = x) = \sum_Y p(X = x, Y)$$

Adapted from Rohan Saxena, Quora, https://qr.ae/pvjMYC

13

13

Marginalisation

IMPERIAL

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

This means to find $P(X{=}x)$, we sum all the probability values where $X{=}x$ occurs with all possible values of Y.

This makes sense, intuitively. To see how, let's say Y can take on n values: $y_1, y_2, \ldots, y_n$.

We can find how often $X{=}x$ occurs if we consider how often $X{=}x$ occurs with each individual value of Y, and sum up all such values to get the total value of the "often-ness" of X.

Adapted from Rohan Saxena, Quora, https://qr.ae/pvjMYC

14

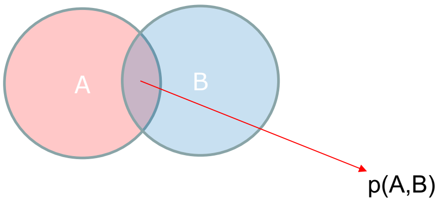
14

Conditional probability

IMPERIAL

- We define the conditional probability of event A, given that event B is true, as follows:

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$



15

15

Bayes Rule

IMPERIAL

Likelihood

Prior

Posterior

$$p(A|B) = \frac{p(B|A).p(A)}{p(B)}$$

Marginalisation

16

16

Example: medical diagnosis

IMPERIAL

- As an example of how to use the Bayes rule, consider the following medical diagnosis problem.
- Suppose you have a patient who is in their 50s, and you decide to have a medical test for diagnosing a neurological condition. If the test is positive, what is the probability of the patient has the disease?
- That obviously depends on how reliable the test is.

17

17

Example: medical diagnosis

IMPERIAL

- Suppose you are told the test has a **sensitivity** of 80%, which means that if a patient has the disease, the test will be positive with a probability of 0.8.
- In other words,

$$p(x = 1|y = 1) = 0.8$$

- where x = 1 is the test is positive, and y = 1 is the event the patient has the disease.
- Many people conclude they are, therefore, 80% likely to have the neurological condition. But this is not true! It ignores the prior probability of having the disease in the given age group, which is quite low (e.g., let's assume for this example it is 0.004 – note: this is not a clinically verified number and only an example to explain the concept of conditional probability):

$$p(y = 1) = 0.004$$

18

18

Example: medical diagnosis

IMPERIAL

- Ignoring this prior is called the base rate fallacy. We also need to take into account the fact that the test may be a false positive or false alarm. Unfortunately, such false positives are quite likely (for example, due to the screening technology):

$$p(x=1|y=0)=0.1$$

- Combining these three terms using the Bayes rule, we can compute the correct answer as follows:

$$\begin{aligned} p(y=1|x=1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} \\ &= \frac{0.8\times 0.004}{0.8\times 0.004+0.1\times 0.996}=0.031 \end{aligned}$$

- where p(y = 0) = 1 − p(y = 1) = 0.996. In other words, if the test is positive, the patient has about a 3% chance of having the condition.

19

19

Example: medical diagnosis

IMPERIAL

$$p(x=1|y=1)=0.8$$

$$p(y=1)=0.004$$

$$p(x=1|y=0)=0.1$$

$$\begin{aligned} p(y=1|x=1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} \\ &= \frac{0.8\times 0.004}{0.8\times 0.004+0.1\times 0.996}=0.031 \end{aligned}$$

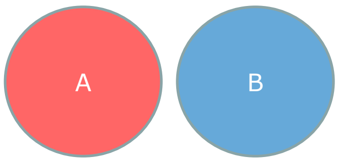
20

20

Independence

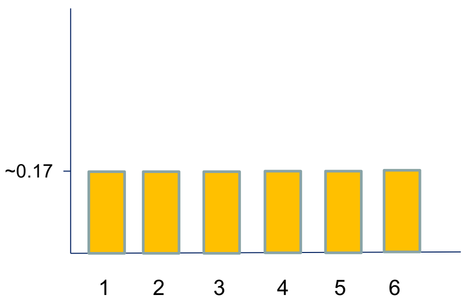
- We say X and Y are unconditionally independent or marginally independent, denoted X ⊥ Y, if we can represent the joint as:

$$X \perp Y \iff p(X, Y) = p(X)p(Y)$$



Probability density function

- Consider a dice



Continuous random variables

IMPERIAL

- So far, we have only considered reasoning about uncertain discrete quantities.
- Suppose X is some uncertain continuous quantity. The probability that X lies in any interval $a \leq X \leq b$ can be computed as follows.
- Let's define the events A = (X ≤ a), B = (X ≤ b) and W = (a < X ≤ b).

$$p(W) = p(B) - p(A)$$
$$p(B) = p(A) + p(W)$$

23

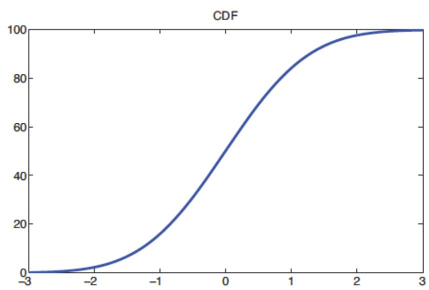
23

Cumulative Distribution Function (cdf)

IMPERIAL

- Define the function $\mathrm{F(q)} \triangleq p(x \leq q)$. This is called the cumulative distribution function or cdf of X.

$$p(a < X \leq b) = F(b) - F(a)$$



24

24

Probability Density Function*

IMPERIAL

- Now define f(x)= $\frac{d}{dx}F(x)$ (we assume this derivative exists); this is called the probability density function or pdf.
- Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

$$P(a < X \leq b) = \int_a^b f(x)dx$$

- As the size of the interval gets smaller, we can write

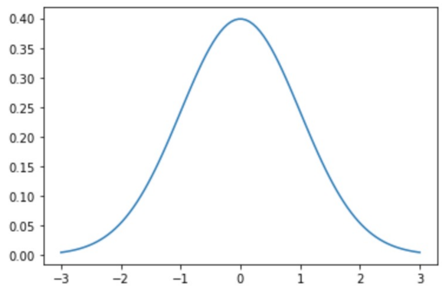
$$P(x \leq X \leq x + dx) \approx p(x)dx$$

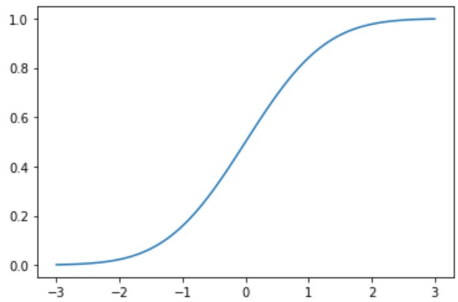
25

25

Probability Density Function vs Cumulative Distribution Function

IMPERIAL





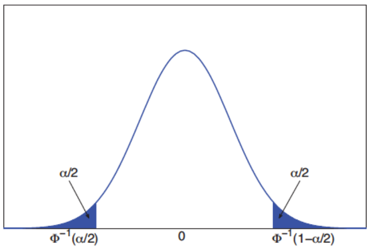
Code: PDF_CDF.ipynb

26

26

pdf for the standard normal

IMPERIAL



- Corresponding pdf for a Gaussian distribution. The shaded regions each contain $\alpha/2$ of the probability mass. Therefore, the non-shaded region contains $1 - \alpha$ of the probability mass.

27

27

Mean

IMPERIAL

- The most familiar property of a distribution is its mean, or expected value, denoted by μ.
- For discrete random variable, it is defined as:

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x \; p(x)$$

- and for continuous random variables, it is defined as

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x \; p(x) dx.$$

28

28

Example: Expected value

IMPERIAL

29

29

Bernoulli

IMPERIAL

- Suppose we toss a coin only once. Let $X \in \{0,1\}$ be a binary random variable with a probability of “success” or “heads” of θ.
- We say that X has a Bernoulli distribution. This is written as $X \sim Ber(\theta)$, where the Probability Mass Function (pmf) is defined as

$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

$$\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

30

30

Binomial distribution

IMPERIAL

- Suppose we toss a coin n times. Let X ∈ {0, . . . , n} be the number of heads. If the probability of heads is θ, then we say X has a binomial distribution, written as X ∼ Bin(n, θ).

31

31

Binomial distribution

IMPERIAL

- If we toss a coin n times and want to determine the probability of k heads (the probability of a head is θ).

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

$and\ n! = n \ast (n-1) \ast (n-2) \ast \ \ldots \ast 1$

32

32

The multinomial and multinoulli distributions*

IMPERIAL

- The binomial distribution can be used to model the outcomes of coin tosses. To model the outcomes of tossing a K-sided die, we can use the multinomial distribution.
- This is defined as follows: let x = ($x_1, \ldots, x_K$) be a random vector, where x_j is the number of times side j of the die occurs.

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$

Example*

IMPERIAL

- Imagine a medical study categorising the type of patient outcomes after a specific intervention. The possible outcomes could be:

P(Full Recovery)=0.5

P(Partial Recovery)=0.3

P(No Improvement)=0.15

P(Adverse Reaction)=0.05

$$P(X=x) = \begin{cases} 0.5, & \text{if } X = \text{Full Recovery}, \\ 0.3, & \text{if } X = \text{Partial Recovery}, \\ 0.15, & \text{if } X = \text{No Improvement}, \\ 0.05, & \text{if } X = \text{Adverse Reaction}. \end{cases}$$

$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$

Gaussian (normal) distribution

IMPERIAL

- The Gaussian or normal distribution is the most widely used in statistics and machine learning. Its pdf is given by:

$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- We write $X \sim \mathcal{N}(\mu,\sigma^2)$ to denote that $p(X = x) = \mathcal{N}(x|\mu,\sigma^2)$. If X ~ $\mathcal{N}$(0, 1), we say X follows a standard normal distribution. This is sometimes called the bell curve.

35

35

The standard normal distribution

IMPERIAL

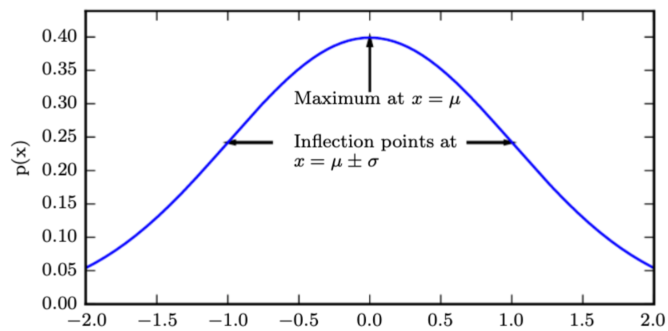


Figure 3.1: The normal distribution. The normal distribution $\mathcal{N}(x;\mu,\sigma^2)$ exhibits a classic "bell curve" shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ. In this example, we depict the **standard normal distribution**, with $\mu = 0$ and $\sigma = 1$.

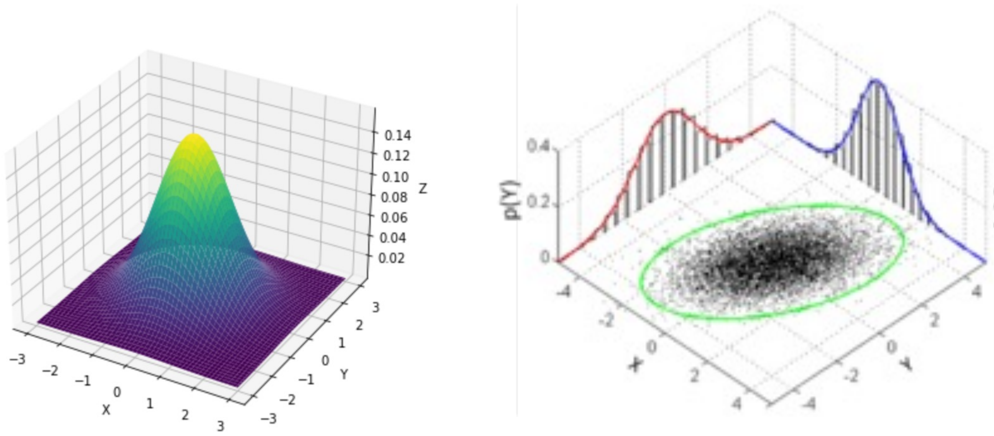
Source: Goodfellow *et al.*, Deep Learning, MIT Press, https://www.deeplearningbook.org/contents/prob.html

36

36

Multivariate Gaussian

IMPERIAL



Source: https://commons.wikimedia.org/wiki/File:MultivariateNormal.png

37

37

Information Theory

IMPERIAL

- The basic intuition behind information theory is that learning about an unlikely event is more informative than learning that a likely event has occurred.

- Likely events should have low information content, and in extreme cases, events that are guaranteed to happen should have no information content whatsoever.
- Less likely events should have higher information content.

38

38

Information Theory

IMPERIAL

- Independent events should have additive information.
- For example, finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once.

- To satisfy all three of these properties, the self-information

of an event x = x is defined as:

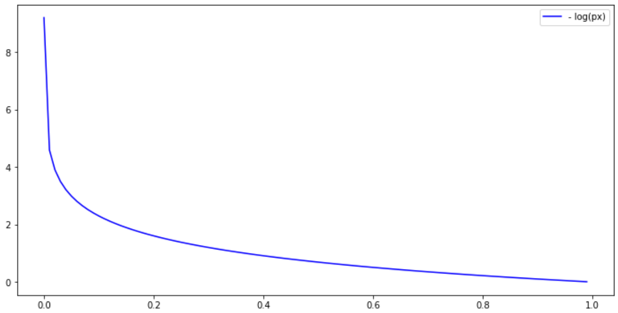
$$I(x) = -\log p(x)$$

39

39

$I(x) = -\log p(x)$

IMPERIAL




```
import numpy as np
import matplotlib.pyplot as plt

fig, ax = plt.subplots(figsize=(12, 6))

epsilon = 0.0001
#an epsilon to avoid numercial calculaton error
px = np.arange(0+epsilon,1, 0.01)
y =  - np.log(px)

ax.plot(px, y, color='blue', label='- log(px)')
plt.legend()
plt.show()
```

40

40

Shannon's entropy

IMPERIAL

- Self-information deals only with a single outcome.
- We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy:

$$H = -\sum_{x} p(x)\,.\log p(x)$$

41

41

Example 1

IMPERIAL

- Calculate the entropy of tossing a fair coin.
- Note: $\log_2(1/2) = -1$

42

42

Example 2

IMPERIAL

- Calculate the entropy of tossing an unfair coin for which

$$p(H) = 0.3 \qquad p(T) = 0.7$$

- Note: $Log_2(0.7) = -0.515$; $Log_2(0.3) = -1.737$

43

43

Example 3

IMPERIAL

- Calculate the entropy of each codon in the given position (with the given data).

ATG GCT AAC TGG CCA
ATG TTT GGA TAC CAG
ATG ACC GTC CTT AGG
ATG GAA GCT TGC TAA
ATG CCA TGG AAC CTC
ATG AGT GGC TTG TGA
ATG TCG AGG CCA ACC
ATG GGT TTT CAT TAG
ATG ACG CGA TCC GGT
ATG TAC GGC AGT TAA

44

44

Conditional independence

IMPERIAL

- Suppose we assume that x_{t+1} is independent from $x_{1:t-1}$
- In words, “the future is independent of the past given the present”.
- This is called the (first-order) Markov assumption.

45

45

Markov chains

IMPERIAL

46

46

Behaviour modelling using MC

IMPERIAL

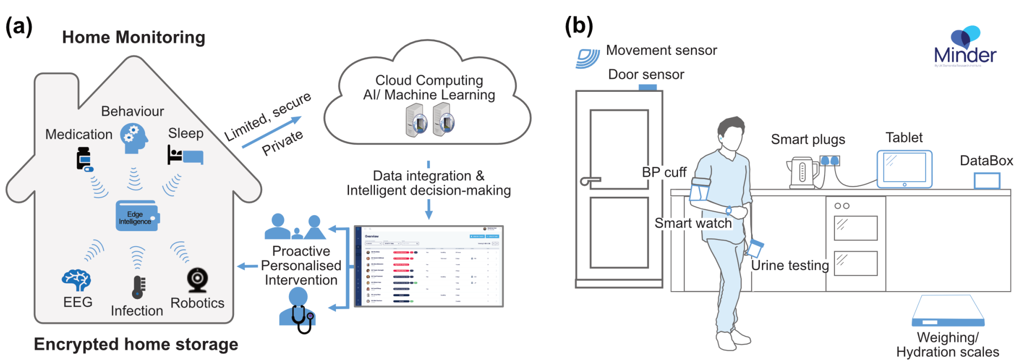
- Imagine each activity/behaviour state is modelled as a Markov chain state.
- Identify the transitions.
- Determine the probability of being in each state and the probability of transitioning between the states.
- You can use this information to build a Markov Chain model.
- You can use the chain probability rule to calculate the probability of being at a current state given an earlier observation.

47

47

Example – daily activity monitoring

IMPERIAL

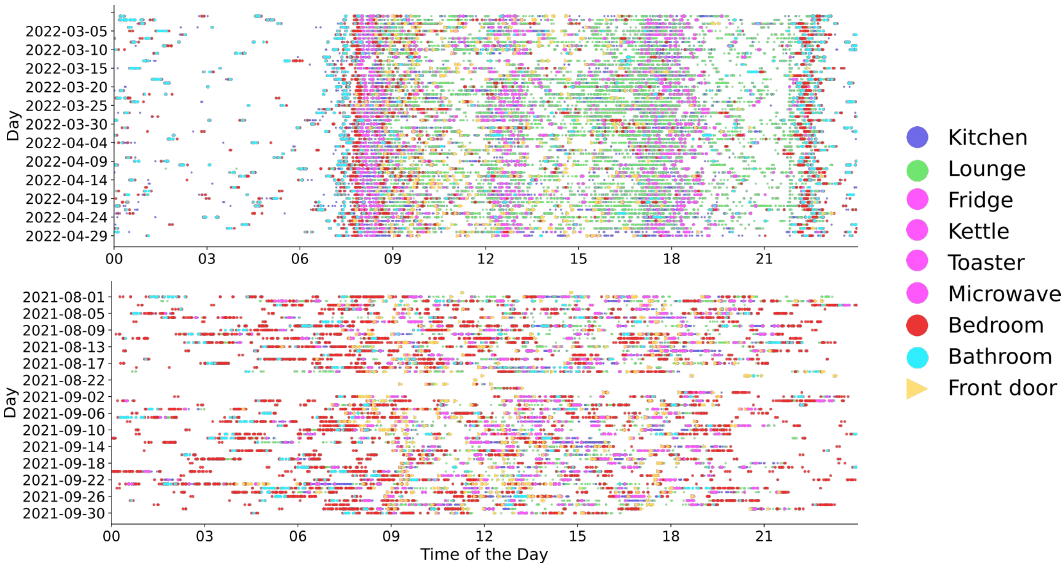


48

48

Continuous remote monitoring – activity data

IMPERIAL

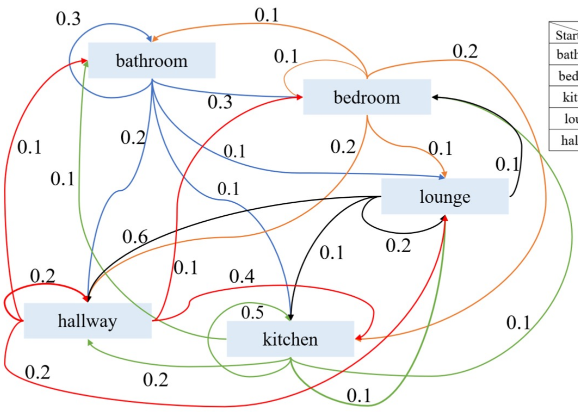


49

49

Changes in activity patterns

IMPERIAL



Start \ End	bathroom	bedroom	kitchen	lounge	hallway
bathroom	0.3	0.3	0.1	0.1	0.2
bedroom	0.1	0.4	0.2	0.1	0.2
kitchen	0.1	0.1	0.5	0.1	0.2
lounge	0	0.1	0.1	0.2	0.6
hallway	0.1	0.1	0.4	0.2	0.2

$$\begin{bmatrix} 0.3 & 0.3 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0.2 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.5 & 0.1 & 0.2 \\ 0.0 & 0.1 & 0.1 & 0.2 & 0.6 \\ 0.1 & 0.1 & 0.4 & 0.2 & 0.2 \end{bmatrix}$$

Transition Matrix T

For more information, refer to: Y Huang et al., https://doi.org/10.1016/j.artmed.2024.102821

50

50

Open Discussion

IMPERIAL

- You are analysing spike train data from a neuron responding to a visual stimulus.
- Over 100 trials, the neuron fires within 10 ms of stimulus onset in 40 trials.

- **Question**: What is the probability that the neuron fires within 10 ms on the next trial?
- How would this probability change if you continued your experiment and observed 10 consecutive trials with no spikes?

51

51

IMPERIAL

Review questions

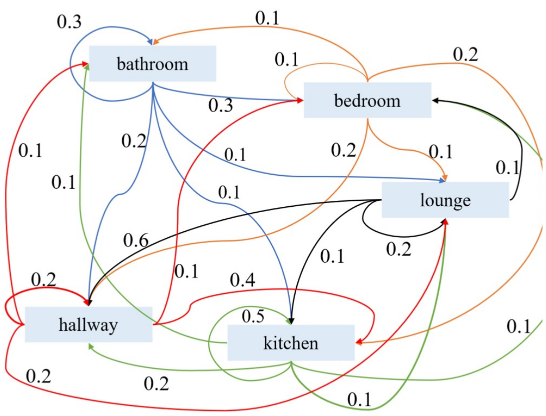
52

52

Q1. Entropy

IMPERIAL

- Let's assume we build a Markov chain and then measure the entropy in the activity states of a patient at home over time.
- If the patient's activity suddenly reduces over time, will the entropy decrease or increase?



53

Q2. Overfitting

IMPERIAL

- If we have more features than the number of training samples, there is a risk of overfitting the model.
- True or False?

54

Q3. Probability

IMPERIAL

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

If we are not given the direct distribution of x to find $P(X=x)$, we sum all the probability values where $X=x$ occurs with all possible values of Y.

What is this called in probability theory?

55

55

Acknowledgement

IMPERIAL

– Some of the content for the slides in this lecture is adapted from Kevin Murphy's book:
  – Machine Learning: A Probabilistic Perspective Kevin P. Murphy, MIT Press.

56

56

If you have any questions

IMPERIAL

- Please feel free to arrange a meeting or email (p.barnaghi@imperial.ac.uk).
- To arrange a meeting, please email my colleague, Ms Rhiannon Kirby.
- My office: 928, Sir Michael Uren Research Hub, White City Campus.

57

57