Introduction to Probability

Lecture 1: Conditional probabilities and Bayes' theorem

Mateja Jamnik, Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: {mateja.jamnik,thomas.sauerwald}@cl.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Outline

Lecturers

Mateja Jamnik

Thomas Sauerwald

Course logistics

Rough syllabus:

- Introduction to probability: 1 lecture
- Discrete and continuous random variables: 6 lectures
- Moments and limit theorems: 3 lectures
- Applications/statistics: 2 lectures

Recommended reading:

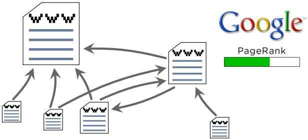
- **Ross, S.M. (2014). A First course in probability. Pearson (9th ed.).**
- **Dekking, F.M., et. al. (2005) A modern introduction to probability and statistics. Springer.**
- Bertsekas, D.P. & Tsitsiklis, J.N. (2008). Introduction to probability. Athena Scientific.
- Grimmett, G. & Welsh, D. (2014). Probability: an Introduction. Oxford University Press (2nd ed.).

Why probability?

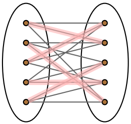
- Gives us mathematical tools to deal with uncertain events.
- It is used everywhere, especially in applications of machine learning.
- Machine learning: use probability to compute predictions about and from data.
- Probability is not statistics:
  - Both about random processes.
  - Probability: logically self-contained, few rules for computing, one correct answer.
  - Statistics: messier, more art, get experimental data and try to draw probabilistic conclusions, no single correct answer.

Applications of probability

Ranking Websites

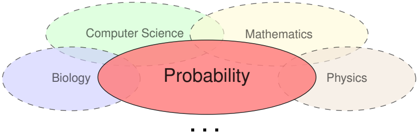


Matching

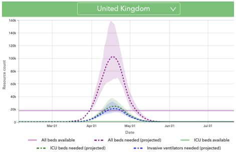


$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

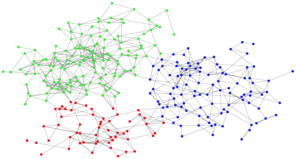
Finance



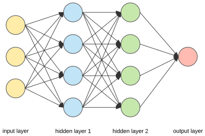
Medicine



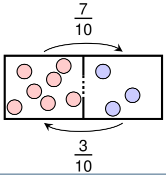
Data Mining



Deep Learning



Particle Processes



Prerequisite background

- Set theory
- Counting: product rule, sum rule, inclusion-exclusion
- Combinatorics: permutations
- Probability space: sample space, event space
- Axioms
- Union bound

- Look for revision material of above on the course website:

  `https://www.cl.cam.ac.uk/teaching/2526/IntroProb/`

Outline

Logistics, motivation, background

Conditional probability

Bayes' Theorem

Independence

Definition

Conditional probability

Consider an experiment with sample space S, and two events E and F. Then, the (conditional) probability of event E given F has occurred (denoted $\mathbf{P}[E|F]$) with $\mathbf{P}[F] > 0$ is defined by

$$\mathbf{P}[E|F] = \frac{\mathbf{P}[E \cap F]}{\mathbf{P}[F]} = \frac{\mathbf{P}[EF]}{\mathbf{P}[F]}$$

Sample space: all possible outcomes consistent with F (i.e., $S \cap F = F$)
Event space: all outcomes in E consistent with F (i.e., $E \cap F$)
Note: we assume that all outcomes are equally likely

$$\mathbf{P}[E|F] = \frac{\#\text{ outcomes in } E \cap F}{\#\text{ outcomes in } F} = \frac{\frac{\#\text{ outcomes in } E \cap F}{\#\text{ outcomes in } S}}{\frac{\#\text{ outcomes in } F}{\#\text{ outcomes in } S}} = \frac{\mathbf{P}[E \cap F]}{\mathbf{P}[F]}$$

Example

Example

Two dice are rolled yielding value D_1 and D_2. Let E be event that $D_1 + D_2 = 4$.

1. What is $\mathbf{P}[\,E\,]$?
2. Let event F be $D_1 = 2$. What is $\mathbf{P}[\,E|F\,]$?

Answer

1. $|S| = 36$, $E = \{(1,3),(2,2),(3,1)\}$, thus $\mathbf{P}[\,E\,] = \frac{3}{36} = \frac{1}{12}$.
2. $S = \{(2,1),(2,2),2,3),(2,4),(2,5),(2,6)\}$, $E = \{(2,2)\}$, thus $\mathbf{P}[\,E|F\,] = \frac{1}{6}$

Rules revisited

Chain rule

Rearranging the definition of conditional probability gives us:

$$\mathbf{P}[\,EF\,] = \mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,]$$

Generalisation of the Chain rule:

Multiplication rule

$$\mathbf{P}[\,E_1 E_2 \cdots E_n\,] = \mathbf{P}[\,E_1\,]\,\mathbf{P}[\,E_2|E_1\,]\,\mathbf{P}[\,E_3|E_2 E_1\,] \cdots \mathbf{P}[\,E_n|E_1 \cdots E_{n-1}\,]$$

Example

Example

An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. What is the probability that each pile has exactly 1 ace?

Answer

Define:

E_1= *ace*♥ is in any one pile

E_2= *ace*♥ and *ace*♠ are in different piles

E_3= *ace*♥, *ace*♠ and *ace*♣ are in different piles

E_4= all aces are in different piles

$$\mathbf{P}[\,E_1E_2E_3E_4\,] = \mathbf{P}[\,E_1\,]\,\mathbf{P}[\,E_2|E_1\,]\,\mathbf{P}[\,E_3|E_1E_2\,]\,\mathbf{P}[\,E_4|E_1E_2E_3\,]$$

We have $\mathbf{P}[\,E_1\,] = 1$. For rest we consider complement of next ace being in the same pile and thus have:

$\mathbf{P}[\,E_2|E_1\,] = 1 - \frac{12}{51}$

$\mathbf{P}[\,E_3|E_1E_2\,] = 1 - \frac{24}{50}$

$\mathbf{P}[\,E_4|E_1E_2E_3\,] = 1 - \frac{36}{49}$

Thus:

$\mathbf{P}[\,E_1E_2E_3E_4\,] = \frac{39 \cdot 26 \cdot 13}{51 \cdot 50 \cdot 49} \approx 0.105$

Outline

Law of total probability

The law of total probability (a.k.a. Partition theorem)

For events E and F where $\mathbf{P}[F] > 0$, then for any event E

$$\mathbf{P}[E] = \mathbf{P}[EF] + \mathbf{P}[EF^c] = \mathbf{P}[E|F]\,\mathbf{P}[F] + \mathbf{P}[E|F^c]\,\mathbf{P}[F^c]$$

In general, for disjoint events $F_1, F_2, \ldots F_n$ s.t. $F_1 \cup F_2 \cup \cdots \cup F_n = S$,

$$\mathbf{P}[E] = \sum_{i=1}^{n} \mathbf{P}[E|F_i]\,\mathbf{P}[F_i]$$

Intuition:

Want to know probability of E. There are two scenarios, F and F^c. If we know these and the probability of E conditioned on each scenario, we can compute the probability of E.

Lightbulb example

Example

There are 3 boxes each containing a different number of light bulbs. The first box has 10 bulbs of which 4 are dead, the second has 6 bulbs of which 1 is dead, and the third box has 8 bulbs of which 3 are dead. What is the probability of a dead bulb being selected when a bulb is chosen at random from one of the 3 boxes (each box has equal chance of being picked)?

Answer

Let event E = "dead bulb is picked", and F_1 = "bulb is picked from first box", F_2 = "bulb is picked from second box" and F_3 = "bulb is picked from third box". We know:

$$\mathbf{P}[\,E|F_1\,] = \frac{4}{10}, \mathbf{P}[\,E|F_2\,] = \frac{1}{6}, \mathbf{P}[\,E|F_3\,] = \frac{3}{8}$$

We need to compute $\mathbf{P}[\,E\,]$, and we know that $\mathbf{P}[\,F_i\,] = \frac{1}{3}$:

$$\mathbf{P}[\,E\,] = \sum_{i=1}^{n} \mathbf{P}[\,E|F_i\,]\,\mathbf{P}[\,F_i\,] = \frac{4}{10}\frac{1}{3} + \frac{1}{6}\frac{1}{3} + \frac{3}{8}\frac{1}{3} = \frac{113}{360} \approx 0.31$$

Bayes' theorem

How many spam emails contain the word "Dear"?

$$\mathbf{P}[\,E|F\,] = \mathbf{P}[\,\text{"Dear"}|\text{spam}\,]$$

But how about what is the probability that an email containing "Dear" is spam?

$$\mathbf{P}[\,F|E\,] = \mathbf{P}[\,\text{spam}|\text{"Dear"}\,]$$

Bayes' theorem

For any events E and F where $\mathbf{P}[\,E\,] > 0$ and $\mathbf{P}[\,F\,] > 0$,

$$\mathbf{P}[\,F|E\,] = \frac{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,]}{\mathbf{P}[\,E\,]}$$

and in expanded form,

$$\mathbf{P}[\,F|E\,] = \frac{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,]}{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,] + \mathbf{P}[\,E|F^c\,]\,\mathbf{P}[\,F^c\,]} = \frac{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,]}{\sum_{i=1}^{n} \mathbf{P}[\,E|F_i\,]\,\mathbf{P}[\,F_i\,]}$$

using the Law of Total Probability. Note that all events F_i must be mutually exclusive (non-overlapping) and exhaustive (their union is the complete sample space) .

Example – Do it at home

Example

60% of all email in 2022 is spam. 20% of spam contains the word "Dear". 1% of non-spam contains the word "Dear". What is the probability that an email is spam given it contains the word "Dear"?

Answer

- Let event E ="Dear", event F = spam.
- $\mathbf{P}[\,F\,] = 0.6$ thus $\mathbf{P}[\,F^c\,] = 0.4$.
- $\mathbf{P}[\,E|F\,] = 0.2$.
- $\mathbf{P}[\,E|F^c\,] = 0.01$.
- Compute $\mathbf{P}[\,F|E\,]$.

$$\mathbf{P}[\,F|E\,] = \frac{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,]}{\mathbf{P}[\,E|F\,]\,\mathbf{P}[\,F\,] + \mathbf{P}[\,E|F^c\,]\,\mathbf{P}[\,F^c\,]} =$$

$$= \frac{(0.2)(0.6)}{(0.2)(0.6) + (0.01)(0.4)} \approx 0.968$$

Bayes' terminology

posterior — likelihood — prior

$$\mathbf{P}[\,F|E\,] = \frac{\mathbf{P}[\,E|F\,] \cdot \mathbf{P}[\,F\,]}{\mathbf{P}[\,E\,]}$$

normalisation constant

F: hypothesis, E: evidence

$\mathbf{P}[\,F\,]$: "prior probability" of hypothesis

$\mathbf{P}[\,E|F\,]$: probability of evidence given hypothesis (likelihood)

$\mathbf{P}[\,E\,]$: calculated by making sure that probabilities of all outcomes sum to 1 (they are "normalised")

Confusion matrix (error matrix)

Used in classification tasks for predicting output error.

		True condition	
	Total population	Condition positive F	Condition negative F^c
Predicted condition	Predicted condition positive E	True positive $\mathbf{P}[E \mid F]$	False positive $\mathbf{P}[E \mid F^c]$
	Predicted condition negative E^c	False negative $\mathbf{P}[E^c \mid F]$	True negative $\mathbf{P}[E^c \mid F^c]$

Medical testing example

Example

- A test is 98% effective at detecting the disease COVID-19 ("true positive").
- The test has a "false positive" rate of 1%.
- 0.5% of the population has COVID-19.
- What is the likelihood you have COVID-19 if you test positive?

Answer

- Let E: test positive, F: actually have COVID-19.
- Need to find $\mathbf{P}[F|E]$.
- We know:
  - $\mathbf{P}[E|F] = 0.98$
  - $\mathbf{P}\left[E|F^c\right] = 0.01$
  - $\mathbf{P}[F] = 0.005$ thus $\mathbf{P}\left[F^c\right] = 0.995$
- Thus

$$\mathbf{P}[F|E] = \frac{\mathbf{P}[E|F]\mathbf{P}[F]}{\mathbf{P}[E|F]\mathbf{P}[F] + \mathbf{P}[E|F^c]\mathbf{P}[F^c]} =$$

$$= \frac{(0.98)(0.005)}{(0.98)(0.005) + (0.01)(0.995)} \approx 0.33$$

Bayesian intuition

- 33% chance of having COVID-19 after testing positive may seem surprising.
- But the space of facts is now conditioned on a positive test result (people who test positive and have COVID-19 **and** people who test positive and don't have COVID-19).

	F yes disease	F^c no disease
E test+	True positive $\mathbf{P}\left[\, E \mid F \,\right] = 0.98$	False positive $\mathbf{P}\left[\, E \mid F^c \,\right] = 0.01$
E^c test-	False negative $\mathbf{P}\left[\, E^c \mid F \,\right] = 0.02$	True negative $\mathbf{P}\left[\, E^c \mid F^c \,\right] = 0.99$

- But what is a chance of having COVID-19 if you test and it comes back negative?

$$\mathbf{P}\left[\, F \mid E^c \,\right] = \frac{\mathbf{P}\left[\, E^c \mid F \,\right] \mathbf{P}\left[\, F \,\right]}{\mathbf{P}\left[\, E^c \mid F \,\right] \mathbf{P}\left[\, F \,\right] + \mathbf{P}\left[\, E^c \mid F^c \,\right] \mathbf{P}\left[\, F^c \,\right]} \approx 0.0001$$

- We update our beliefs with Bayes' theorem:
  I have 0.5% chance of having COVID-19. I take the test:
  - Test is positive: I now have 33% chance of having COVID-19.
  - Test is negative: I now have 0.01% chance of having COVID-19.
- So it makes sense to take the test.

Outline

Independent events

Independence

Two events E and F are independent if and only if

$$\mathbf{P}[EF] = \mathbf{P}[E]\,\mathbf{P}[F]$$

Otherwise, they are called dependent events.

In general, n events $E_1, E_2, \ldots, E_n$ are mutually independent if for every subset of these events with r elements (where $r \leq n$) it holds that

$$\mathbf{P}[E_a E_b \cdots E_r] = \mathbf{P}[E_a]\,\mathbf{P}[E_b] \cdots \mathbf{P}[E_r]$$

Therefore for 3 events E, F, G to be independent, we must have

$$\begin{aligned}
\mathbf{P}[EFG] &= \mathbf{P}[E]\,\mathbf{P}[F]\,\mathbf{P}[G] \\
\mathbf{P}[EF] &= \mathbf{P}[E]\,\mathbf{P}[F] \\
\mathbf{P}[EG] &= \mathbf{P}[E]\,\mathbf{P}[G] \\
\mathbf{P}[FG] &= \mathbf{P}[F]\,\mathbf{P}[G]
\end{aligned}$$

Independence of complement

Notice an equivalent definition for independent events E and F ($\mathbf{P}[F] > 0$)

$$\mathbf{P}[E|F] = \mathbf{P}[E]$$

Proof:

$$\mathbf{P}[E|F] = \frac{\mathbf{P}[EF]}{\mathbf{P}[F]} = \frac{\mathbf{P}[E]\mathbf{P}[F]}{\mathbf{P}[F]} = \mathbf{P}[E]$$

Independence of complement

If events E and F are independent, then E and F^c are independent:

$$\mathbf{P}[EF^c] = \mathbf{P}[E]\mathbf{P}[F^c]$$

Proof:

$$\mathbf{P}[EF^c] = \mathbf{P}[E] - \mathbf{P}[EF] = \mathbf{P}[E] - \mathbf{P}[E]\mathbf{P}[F] =$$
$$= \mathbf{P}[E](1 - \mathbf{P}[F]) = \mathbf{P}[E]\mathbf{P}[F^c]$$

Example

Example

Each roll of a die is an independent trial. We have two rolls of D_1 and D_2. Let event $E : D_1 = 1$, $F : D_2 = 6$ and event $G : D_1 + D_2 = 7$ (thus $G = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$).

1. Are E and F independent?
2. Are E and G independent?
3. Are E, F, G independent?

Answer

1. Yes, since $\mathbf{P}[\,E\,] = \frac{1}{6}, \mathbf{P}[\,F\,] = \frac{1}{6}$ and $\mathbf{P}[\,EF\,] = \frac{1}{36}$.
2. Yes, since $\mathbf{P}[\,E\,] = \frac{1}{6}, \mathbf{P}[\,G\,] = \frac{6}{36} = \frac{1}{6}$ and $\mathbf{P}[\,EG\,] = \frac{1}{36}$.
3. No, since $\mathbf{P}[\,EFG\,] = \frac{1}{36} \neq \frac{1}{6}\frac{1}{6}\frac{1}{6}$.

Conditional independence – Convince yourself at home

Conditional independence

Two events E and F are called conditionally independent given a third event G if

$$\mathbf{P}[\,EF|G\,] = \mathbf{P}[\,E|G\,]\,\mathbf{P}[\,F|G\,]$$

Or equivalently,

$$\mathbf{P}[\,E|FG\,] = \mathbf{P}[\,E|G\,]$$

Notice that:

- Dependent events can become conditionally independent.
- Independent events can become conditionally dependent.
- Knowing when conditioning breaks or creates independence is a big part of building complex probabilistic models.

Example revisited

Example

Each roll of a die is an independent trial. We have two rolls of D_1 and D_2. Let event $E : D_1 = 1$, $F : D_2 = 6$ and event $G : D_1 + D_2 = 7$ (thus $G = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$).

1. Are E and F independent?
2. Are E and F independent given G?

Answer

1. Yes, since $\mathbf{P}[\,E\,] = \frac{1}{6}$, $\mathbf{P}[\,F\,] = \frac{1}{6}$ and $\mathbf{P}[\,EF\,] = \frac{1}{36}$.
2. No, since $\mathbf{P}[\,E|G\,] = \frac{1}{6}$ and $\mathbf{P}[\,F|G\,] = \frac{1}{6}$, but $\mathbf{P}[\,EF|G\,] = \frac{1}{6} \neq \mathbf{P}[\,E|G\,]\mathbf{P}[\,F|G\,]$.

Summary of conditional probability

Conditioning on event G:

Name of rule	Original rule	Conditional rule
1st axiom of probability	$0 \leq \mathbf{P}[E] \leq 1$	$0 \leq \mathbf{P}[E\|G] \leq 1$
Complement	$\mathbf{P}[E] = 1 - \mathbf{P}[E^c]$	$\mathbf{P}[E\|G] = 1 - \mathbf{P}[E^c\|G]$
Chain rule	$\mathbf{P}[EF] = \mathbf{P}[E\|F]\mathbf{P}[F]$	$\mathbf{P}[EF\|G] = \mathbf{P}[E\|FG]\mathbf{P}[F\|G]$
Bayes' theorem	$\mathbf{P}[F\|E] = \frac{\mathbf{P}[E\|F]\mathbf{P}[F]}{\mathbf{P}[E]}$	$\mathbf{P}[F\|EG] = \frac{\mathbf{P}[E\|FG]\mathbf{P}[F\|G]}{\mathbf{P}[E\|G]}$