Lecture 4: Introduction to Probability Theory

COMP90049
Knowledge Technologies

Lea Frermann and Karin Verspoor, CIS

Semester 2, 2019

THE UNIVERSITY OF MELBOURNE

Roadmap

Last time... Similarity

- retrieve similar documents
- recommend books to similar users
- quantification of similarity
  ...define features and measure overlap
  ...map items into (eucledian) space and measure spatial distances

Roadmap

Last time... Similarity

- retrieve similar documents
- recommend books to similar users
- quantification of similarity
  ...define features and measure overlap
  ...map items into (eucledian) space and measure spatial distances

Today... Probability

- probability of this email being spam?
- probability that User A will like this book?
- probability of getting home dry tonight?

Estimating confidence in different possible outcomes

Probability Theory

“The calculus of probability theory provides us with a **formal framework** for considering multiple possible **outcomes** and their **likelihood**. It defines a set of **mutually exclusive** and **exhaustive** possibilities, and associates each of them with a probability — **a number between 0 and 1**, so that the **total probability of all possibilities is 1**. This framework allows us to consider options that are **unlikely, yet not impossible**, without reducing our conclusions to content-free lists of every possibility.”

From Probabilistic Graphical Models: Principles and Techniques (2009; Koller and Friedman) `http://pgm.stanford.edu/intro.pdf`

(Very) Basics of Probability Theory

P(A): the probability of A — the fraction of times the event is true in independent trials

$$0 <= P(A) <= 1$$

$$P(True) = 1$$

$$P(False) = 0$$

(Very) Basics of Probability Theory

P(A): the probability of A — the fraction of times the event is true in independent trials

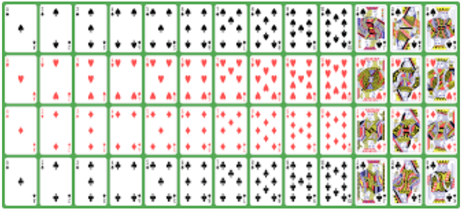
$$0 <= P(A) <= 1$$

$$P(True) = 1$$

$$P(False) = 0$$

Given a deck of 52 cards

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)



https://commons.wikimedia.org/wiki/File:English_pattern_playing_cards_deck.svg

(Very) Basics of Probability Theory

P(A): the probability of A — the fraction of times the event is true in independent trials

$$0 <= P(A) <= 1$$

$$P(\mathit{True}) = 1$$
$$P(\mathit{False}) = 0$$

Given a deck of 52 cards

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)

$$P(\texttt{queen}) = ?$$

$$P(\texttt{red}) = ?$$

$$P(\texttt{heart}) = ?$$

(Very) Basics of Probability Theory

P(A): **the probability of A** the fraction of times the event is true in independent trials

$$0 <= P(A) <= 1$$

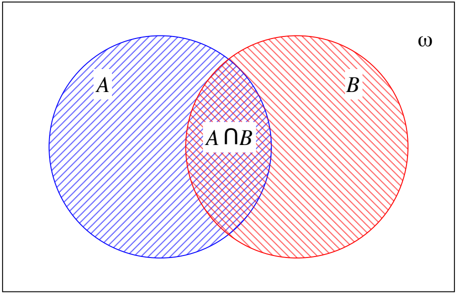
$$P(True) = 1$$
$$P(False) = 0$$

Given a deck of 52 cards

- 13 ranks (ace, 2-10, jack, queen, king)
- of each of four suits (clubs, spades = black; hearts, diamonds = red)

$$P(\texttt{queen}) = \frac{1}{13}$$
$$P(\texttt{heart}) = \frac{1}{4}$$
$$P(\texttt{red}) = \frac{1}{2}$$

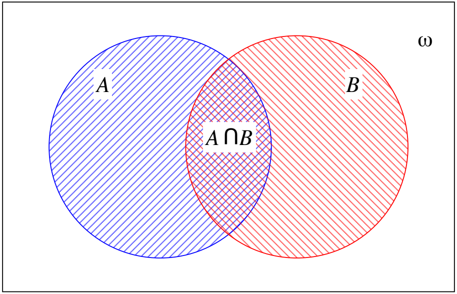
P(A, B):	**joint probability of A and B**	the probability of both A and B occurring = $P(A \cap B)$



$$P(\texttt{ace}, \texttt{heart}) = ?$$

$$P(\texttt{heart}, \texttt{red}) = ?$$

P(A, B):	**joint probability of A and B**	the probability of both A and B occurring = $P(A \cap B)$

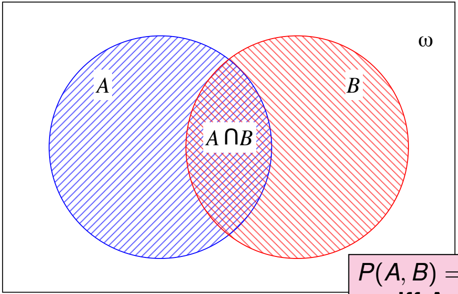


$$P(\texttt{ace}, \texttt{heart}) = \frac{1}{52}$$

$$P(\texttt{heart}, \texttt{red}) = \frac{1}{4}$$

Basics of Probability Theory

P(A, B):	**joint probability of A and B**	the probability of both A and B occurring = $P(A \cap B)$



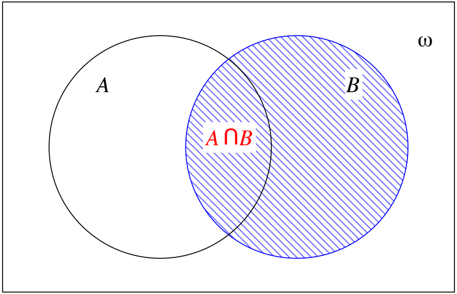
$P(A, B) = P(A) * P(B)$
iff A and B are independent

$$P(\texttt{ace}, \texttt{heart}) = \frac{1}{52}$$

$$P(\texttt{heart}, \texttt{red}) = \frac{1}{4}$$

Conditional Probability

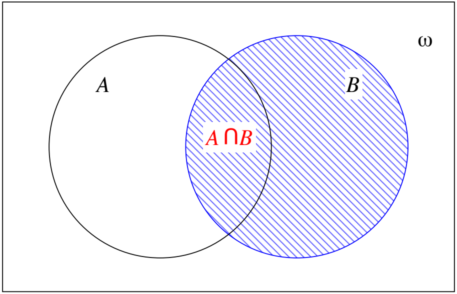
P(A|B): **conditional probability** the probability of A given the occurrence of B = $\frac{P(A \cap B)}{p(B)}$



$$P(\texttt{ace}|\texttt{heart}) = ?$$
$$P(\texttt{heart}|\texttt{red}) = ?$$

Conditional Probability

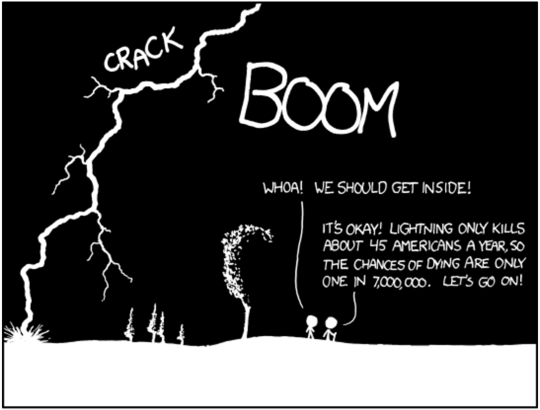
P(A|B): **conditional probability** the probability of A given the occurrence of B = $\frac{P(A \cap B)}{p(B)}$



$$P(\texttt{ace}|\texttt{heart}) = \frac{1}{52} / \frac{1}{4} = \frac{1}{13}$$

$$P(\texttt{heart}|\texttt{red}) = \frac{1}{4} / \frac{1}{2} = \frac{1}{2}$$

What type of probability?



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

https://imgs.xkcd.com/comics/conditional_risk.png

Rules of Probability I

- *Independence:* A and B are independent iff $P(A \cap B) = P(A)P(B)$
- *Sum rule*: $P(A) = \sum_B P(A \cap B)$
- *Multiplication rule*: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- *Chain rule*:
  $P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n| \cap_{i=1}^{n-1} A_i)$

Rules of Probability I

- *Independence:* A and B are independent iff $P(A \cap B) = P(A)P(B)$
- *Sum rule*: $P(A) = \sum_B P(A \cap B)$

$$\boxed{\begin{gathered} P(king) = P(king, diamond) + \\ P(king, heart) + \\ P(king, spades) + P(King, clubs) \end{gathered}}$$

- *Multiplication rule*: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- *Chain rule*:
  $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \ldots P(A_n| \cap_{i=1}^{n-1} A_i)$

Rules of Probability I

- *Independence:* A and B are independent iff $P(A \cap B) = P(A)P(B)$
- *Sum rule*: $P(A) = \sum_B P(A \cap B)$

$$P(king) = P(king, diamond) + P(king, heart) + P(king, spades) + P(King, clubs)$$

- *Multiplication rule*: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- *Chain rule*:
  $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \dots P(A_n| \cap_{i=1}^{n-1} A_i)$

again, we can choose the factorization, e.g., :

$$p(July, 5^\circ C, sick) = p(July) \times p(5^\circ C|July) \times p(sick|5^\circ C, July)$$

makes sense

$$= p(5^\circ C) \times p(sick|5^\circ C) \times p(July|5^\circ C, sick)$$

???

Rules of Probability II

Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

Posterior Probability $P(A|B)$

- the degree of belief having accounted for B.

Prior Probability $P(A)$

- the initial degree of belief in A.
- the probability of A occurring, given no additional knowledge about A

Likelihood $P(B|A)$

- the support B provides for A

Normalizing constant ('Evidence') $p(B) = \sum_A P(B|A)P(A)$

Rules of Probability II

Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

So what?

1. Bayes' Rule allows us to compute $P(A|B)$ given knowledge of the 'inverse' probability $P(B|A)$.
2. Bayes' Rule allows us to update prior belief with empirical evidence

Rules of Probability II

Bayes Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad \left(\text{cf., } P(A|B) = \frac{P(A \cap B)}{P(B)}\right)$$

So what?

1. Bayes' Rule allows us to compute $P(A|B)$ given knowledge of the 'inverse' probability $P(B|A)$.
2. Bayes' Rule allows us to update prior belief with empirical evidence

Example

Estimate the probability of a student **being smart** given that (s)he **achieved H1** score, $P(smart|H1)$ from the following information:

$P(Smart) = 0.3$	prior rate of smart students
$P(H1\|Smart) = 0.6$	empirically measured $H1\|smart$
$P(H1\|!Smart) = 0.1$	empirically measured $H1\|!smart$
$P(H1) = 0.25$	derived from the above

(What if $P(Smart) = 0.1$?)

Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes (aka *Bernoulli trials*)

  e.g. multiple coin tosses ($\langle H, T, H, H, ..., T \rangle$)

Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes (aka *Bernoulli trials*)
  e.g. multiple coin tosses ($\langle H, T, H, H, ..., T \rangle$)

- The probability of an event with probability p occurring exactly m out of n times is given by

$$P(m, n, p) = \binom{n}{m} p^m (1-p)^{n-m}$$

$$P(m, n, p) = \underbrace{\frac{n!}{m!(n-m)!}}_{\substack{\text{possible distributions} \\ \text{of } m \text{ successes} \\ \text{over } n \text{ trials}}} \underbrace{p^m}_{m \text{ successes}} \underbrace{(1-p)^{n-m}}_{n-m \text{ failures}}$$

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes e from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes e from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 out of 3: $C(3, 2) = \frac{3!}{2!1!} = 3$

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$

2. number of possible outcomes e from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 out of 3: $C(3, 2) = \frac{3!}{2!1!} = 3$

4. 3 possible outcomes, $\frac{1}{8}$ for each: $P(X = 2) = \frac{3}{8}$

$$P\left(2, 3, \frac{1}{2}\right) = \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} = 3\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)$$

Binomial Example: Coin Toss

What is the probability that if we toss a fair coin 3 times, we will get 2 heads?

1. $m = 2$ successes (heads) when flipping coin $n = 3$ times; $P(X = 2)$
2. number of possible outcomes e from 3 coin flips:

$$2 * 2 * 2 = 2^3 = 8 \qquad \text{each with } P(e) = \frac{1}{8}$$

3. Choose 2 out of 3: $C(3,2) = \frac{3!}{2!1!} = 3$
4. 3 possib

$$P(m,n,p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

$$P\left(2, 3, \frac{1}{2}\right) = \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} = 3\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)$$

Multinomial Distributions

- A **multinomial distribution** results from a series of independent trials with more than two outcomes, e.g.,

  - players A and B in a chess tournament
  - 3 possible outcomes:

$$P(A\ wins) = 0.4 \qquad P(B\ wins) = 0.35 \qquad P(draw) = 0.25$$

Multinomial Distributions

- A **multinomial distribution** results from a series of independent trials with more than two outcomes, e.g.,

  - players A and B in a chess tournament
  - 3 possible outcomes:

  $$P(A\ wins) = 0.4 \qquad P(B\ wins) = 0.35 \qquad P(draw) = 0.25$$

- The probability of events $X_1, X_2, ..., X_n$ with probabilities $p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ times, respectively, is given by

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \left(\sum_i x_i\right)! \prod_i \frac{p_i^{x_i}}{x_i!}$$

If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?

Activity!

Please navigate to

`https://pollev.com/kt19`

and take the small quiz on probabilities.[1]

You are welcome to work in pairs!

[1] It's 100% anonymous!

Information theory – Intuition and Motivation

Claude Shannon (1948) "A Mathematical Theory of Communication"

What is information

- loss of uncertainty
- transmitted through sequence of symbols (letters, bits, pixels, ...)
- e.g., reading a troubleshooting forum
  e.g., listening to the news

Optimal encoding and transmission of information

- data compression
- cryptography
- linguistics and natural language processing

Core Concepts in Information Theory

1. Digital Representation

- Content of the message is irrelevant
- it's all just bits (sequences of 0 and 1)
- we can encode text, images, sound, ...
- **how can we minimize the # bits to communicate successfully?**

2. Communication is never perfect

- channels are noisy
- every channel has a maximum capacity (bits/s)
- above the limit: no error-free communication
- minimal **reduncancy** for **fast, error-free communication** under noise

3. Efficient Representation (aka Data compression)

- minimal sequence of bits for exact recoverability

- set of symbols (e.g., words) $w_1, w_2 \ldots . w_N$
- a message as a sequence of symbols

$$M = \{w_1, w_5, w_{27}, \ldots\} \qquad |M| = \sum_i f(w_i)$$

- optimal (minimal) code length of each symbol (in bits)

$$E(w_i) = -log_2 \frac{f(w_i)}{|M|}$$

- **frequent symbols need fewer bits**

3. Efficient Representation (aka Data compression)

- minimal sequence of bits for exact recoverability

- set of symbols (e.g., words) $w_1, w_2 \dots w_N$
- a message as a sequence of symbols

$$M = \{w_1, w_5, w_{27}, \dots\} \qquad |M| = \sum_i f(w_i)$$

- optimal (minimal) code length of each symbol (in bits)

$$E(w_i) = -log_2 \frac{f(w_i)}{|M|}$$

- **frequent symbols need fewer bits**
- same holds for events from probability distribution $P(X)$

$$E(x_i) = -log_2 P(x_i)$$

3. Efficient Representation (aka Data compression)

- minimal sequence of bits for exact recoverability
- set of symbols (e.g., words) $w_1, w_2 \ldots w_N$
- a message as a sequence of symbols

$$M = \{w_1 \ldots \sum f(w_i)$$

- optimal (minimal) ...s)

$$E(w_i) = -log_2 \frac{\ldots}{|M|}$$

- **frequent symbols need fewer bits**
- same holds for events from probability distribution $P(X)$

$$E(x_i) = -log_2 P(x_i)$$

Significance for knowledge technologies (like information retrieval)?

4. Entropy

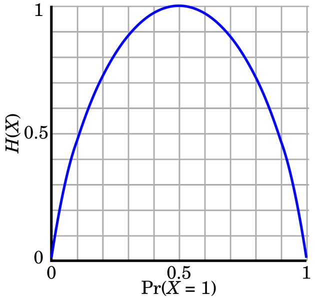
- average message length under the best possible code
- given **probability distribution** $P(X)$, the information (bits) required to predict an **event** x_i

$$H(X) = -\sum_i P(x_i) log_2 P(x_i)$$

- frequent events need fewer bits
- amount of **unexpected** content in a message
  the closer to random distribution, the higher the entropy

Entropy – Examples

https://en.wikipedia.org/wiki/Entropy“˙(information“˙theory)“#/media/File:Binary“˙entropy“˙plot.svg



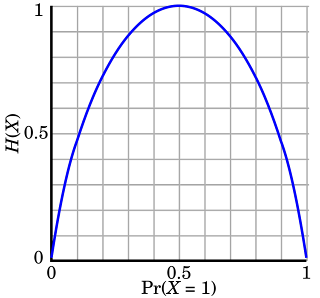
Fair coin $P(X) = [0.5, 0.5]$

$$
\begin{aligned}
H(x) &= -(P(X=h) \times \log P(X=h) + P(X=t) \times \log P(X=t)) \\
&= -(0.5 \times \log 0.5 + 0.5 \times \log 0.5) \\
&= -(0.5 \times -1 + 0.5 \times -1) \\
&= -(-1) = 1
\end{aligned}
$$

We learnt 1 bit of information!

Entropy – Examples

https://en.wikipedia.org/wiki/Entropy“˙(information“˙theory)“#/media/File:Binary“˙entropy“˙plot.svg



log 0 = 0 (def.)

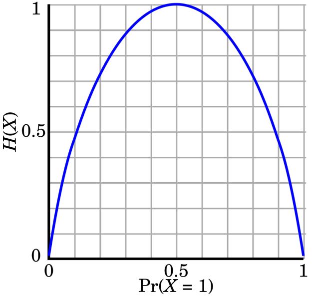
Deterministic coin $P(X) = [1.0, 0.0]$

$$
\begin{array}{rlllll}
H(x) = & -(P(X=h) & \times \log P(X=h) & + P(X=t) & \times \log P(X=t)) \\
= & -(1.0 & \times \log 1.0 & + 0.0 & \times \log 0.0) \\
= & -(1.0 & \times -0 & + 0.0 & \times 0) \\
= & -(0) = 0 & & &
\end{array}
$$

We learnt nothing at all!

Entropy – Examples

https://en.wikipedia.org/wiki/Entropy“˙(information“˙theory)“#/media/File:Binary“˙entropy“˙plot.svg



Unfair coin $P(X) = [0.8, 0.2]$

$$\begin{array}{llllll} H(x) = & -(P(X=h) & \times \log P(X=h) & +P(X=t) & \times \log P(X = \\ = & -(0.8 & \times \log 0.8 & +0.2 & \times \log 0.2) \\ = & -(0.8 & \times -0.32 & +0.2 & \times -2.32) \\ = & -(-0.72) = 0.72 & & & \end{array}$$

$\boxed{\text{We learnt} < 1 \text{ bit!}}$

Value range of entropy

$0 \leq H(X) \leq log\ n$ $\quad\quad$ n : number of outcomes

- coin toss: 2 outcomes, maximum entropy = log 2 = 1 bit
- dice roll: 6 outcomes, maximum entropy = log 6 = 2.58 bits
- English letters: 26 outcomes, maximum entropy = log 26 = 4.7 bits
- **English text**: **What do you think? Why?**

Entropy

Value range of entropy

$$0 \leq H(X) \leq log\ n \qquad n: \text{ number of outcomes}$$

- coin toss: 2 outcomes, maximum entropy = log 2 = 1 bit
- dice roll: 6 outcomes, maximum entropy = log 6 = 2.58 bits
- English letters: 26 outcomes, maximum entropy = log 26 = 4.7 bits
- **English text**: $\approx$ **1 bit**

Summary

Probability underlies many modern knowledge technologies

- estimate the (conditional, joint) probability of observations
- rare events
- update knowledge using Bayes Rule
- high entropy (=informative) information
- ...

Next: Approximate matching