Lecture 4: Introduction to Probability Theory

COMP90049
Knowledge Technologies

Sarah Erfani and Karin Verspoor, CIS

Semester 2, 2018

THE UNIVERSITY OF MELBOURNE

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

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

Given a deck of 52 cards;
13 ranks (ace, king, queen, jack, 2-10)
of each of four suits (clubs, spades = black; hearts, diamonds = red)

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

(Very) Basics of Probability Theory

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

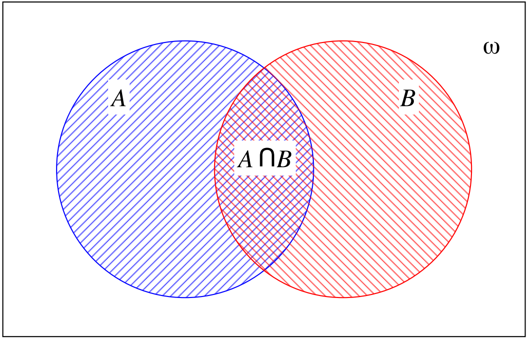
$$0 <= P(A) <= 1$$
$$P(True) = 1$$
$$P(False) = 0$$

Given a deck of 52 cards;
13 ranks (ace, king, queen, jack, 2-10)
of each of four suits (clubs, spades = black; hearts, diamonds = red)

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

Basics of Probability Theory

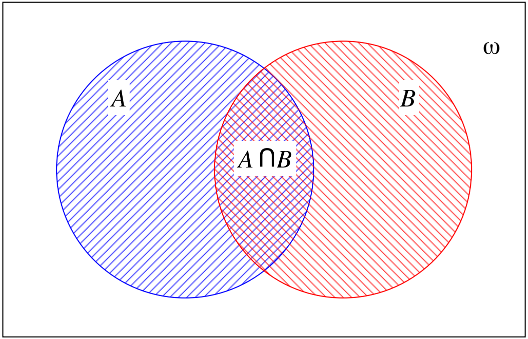
- *Joint probability* ($P(A, B)$):
  the probability of both A and B occurring = $P(A \cap B)$



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

Basics of Probability Theory

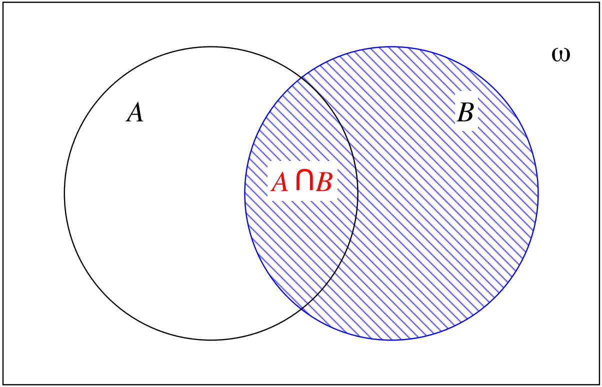
- *Joint probability* ($P(A, B)$):
  the probability of both A and B occurring = $P(A \cap B)$



$$P(\mathtt{ace}, \mathtt{heart}) = \frac{1}{52},\ P(\mathtt{heart}, \mathtt{red}) = \frac{1}{4}$$

Conditional Probability

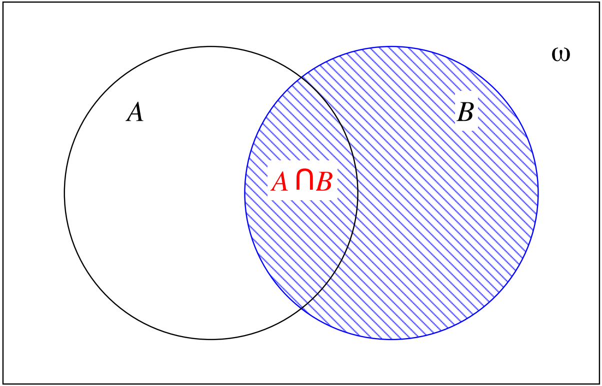
- *Conditional probability* ($P(A|B)$):
  the probability of A occurring given the occurrence of $B = \frac{P(A \cap B)}{P(B)}$



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

Conditional Probability

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



$$P(\texttt{ace}|\texttt{heart}) = \frac{1}{13},\ P(\texttt{heart}|\texttt{red}) = \frac{1}{2}$$

Conditional Probability

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

- *Prior probability* ($P(A)$): the probability of A occurring, given no additional knowledge about A
- *Posterior probability*($P(A|B)$): the probability of A occurring, given background knowledge about event(s) B leading up to A
- *Independence: A* and B are independent iff $P(A \cap B) = P(A)P(B)$

Bayes Rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

- $P(A|B)$, the posterior, is the degree of belief having accounted for B.
- $P(A)$, the prior, is the initial degree of belief in A.
- the quotient $\frac{P(B|A)}{P(B)}$ represents the support B provides for A.

Bayes' Rule is important because it allows us to compute $P(A|B)$ given knowledge of the 'inverse' probability $P(B|A)$.

For instance, imagine we believe (from prior data), that $P(H1|Smart) = 0.6$, $P(Smart) = 0.3$, and $P(H1) = 0.2$.

Now we learn that a particular student received a mark of H1. Can we estimate $P(Smart)$ for that student, e.g. $P(Smart|H1)$?

(What if the $P(H1) = 0.4$?)

Binomial Distributions

- A **binomial distribution** results from a series of independent trials with only two outcomes
(i.e. *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

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

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

Intuition: we want m successes (p^m) and $n-m$ failures ($(1-p)^{n-m}$). However, the m successes can occur anywhere among the n trials, and there are C(n, m) different ways of distributing m successes in a sequence of n trials.

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?

X=number of heads when flipping coin 3 times; $P(X = 2)$

Possible outcomes from 3 coin flips $= 2 * 2 * 2 = 2^3 = 8$. Each possible outcome has $\frac{1}{8}$ probability.

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

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

$$P\left(2, 3, \frac{1}{8}\right) = \frac{3!}{2!(3-2)!}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{3-2} = 3\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)$$

Multinomial Distributions

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

  e.g. two players in a tournament, 3 outcomes: (Player A winner, Player B winner, draw);
  probability that Player A wins is 0.4, that player B wins is 0.35, probability of draw is 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

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

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?

Information theory

Consider a message M composed of distinct symbols $w_1, \ldots, w_n$, where each symbol w_i has a frequency f_i. The total length of the message is $|M| = \sum_i f_i$.

Information theory tells us that the minimum length encoding of the message is to allocate $-\log_2 \frac{f_i}{|M|}$ bits to symbol w_i.

That is, common symbols (high f_i) get a small number of bits and rare symbols get a large number of bits. The sum

$$E = \sum_i -f_i \times \log_2 \frac{f_i}{|M|}$$

is the *entropy* of the message; this is the theoretical minimum length of the message in the context of the provided information.

Relationship to information retrieval: we are interested in terms that have high entropy in a document collection (bursty), and documents in which these terms are a significant component of the document's 'message'.

Entropy (Information Theory)

- A measure of *unpredictability*
- Given a probability distribution, the information (in bits) required to predict an event is the distribution's *entropy* or *information value*
- (The average information required to specify the outcome x when the receiver knows the distribution p)
- The entropy of a discrete random event x with possible states $1, ..n$ is:

$$\begin{aligned} H(x) &= -\sum_{i=1}^{n} P(i) \log_2 P(i) \\ &= \frac{freq(*) \log_2(freq(*)) - \sum_{i=1}^{n} freq(i) \log_2(freq(i))}{freq(*)} \end{aligned}$$

where $0 \log_2 0 =^{def} 0$

Interpreting Entropy Values

entropy = information content
Measures the average missing information on a random source, or the *unevenness* of a probability distribution.

- A high entropy value means x is unpredictable.
  - fair coin $\rightarrow$ impossible to predict outcome of coin toss ahead of time

$$\begin{aligned} H(x) &= -(P(X=h)\log_2 P(X=h) + P(X=t)\log_2 P(X=t)) \\ &= -(0.5\log_2 0.5 + 0.5\log_2 0.5) \\ &= -((0.5 * -1) + (0.5 * -1)) = -(-1) = 1 \end{aligned}$$

  - Two possible outcomes with equal probability;
    Learning the outcome contains one bit of information

- A low entropy value means x is predictable.
  - A coin toss with two heads is perfectly predictable.

$$H(x) = -(1\log_2 1 + 0\log_2 0) = -(0+0) = 0$$

  - We don't learn anything once we see the outcome.

Entropy of an unfair coin

Let's say $P(X = h) = 0.9$ and $P(X = t) = 0.1$

$$\begin{aligned} H(x) &= -(P(X = h)\log_2 P(X = h) + P(X = t)\log_2 P(X = t)) \\ &= -(0.9\log_2 0.9 + 0.1\log_2 0.1) \\ &= 0.47 \end{aligned}$$

Entropy values

NB: The range of the entropy values is not [0, 1].

- The range is determined by the possible number of outcomes.
- $0 \leq$ Entropy $\leq \log(n)$, where n is number of outcomes
- Entropy=0 (minimum entropy) when one probability is 1, others 0
- Entropy=log(n) (maximum entropy): when all probabilities have equal values of 1/n

Summary

Probability forms the foundation of many knowledge technologies.

- What are joint and conditional probabilities?
- What are prior and posterior probabilities?
- What is entropy, and how should you interpret entropy values?

Next: Approximate matching