Lecture 4: Introduction to Probability Theory

Sarah Erfani and Karin Verspoor, CIS

Semest

Probability Theory

"The calculus of probability theory provides us with a formal framework

without reducing our conclusions to conte

From Probabilistic Graphical Models Pr
Koller and Friedman) `http://pgm.s`

(Very) Basics of Probability Theory

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

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

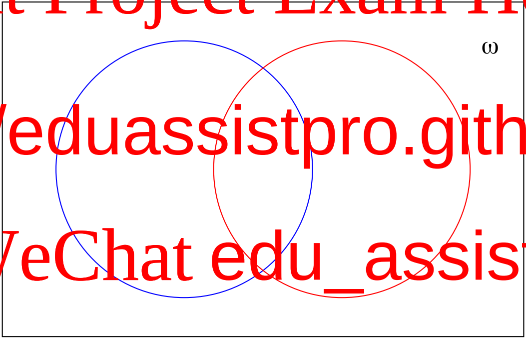
$$P(\texttt{ace}) = ?,\ P(\texttt{red}) = ?$$

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

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

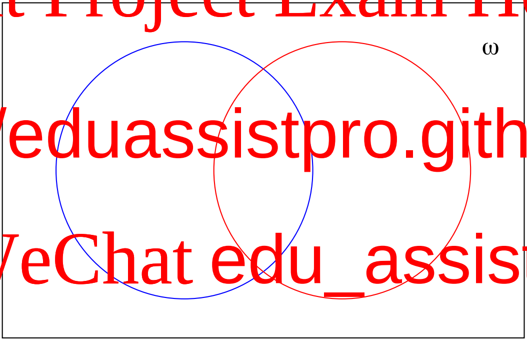
$$P(\texttt{ace}) = \frac{1}{13}, P(\texttt{red}) = -$$

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



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

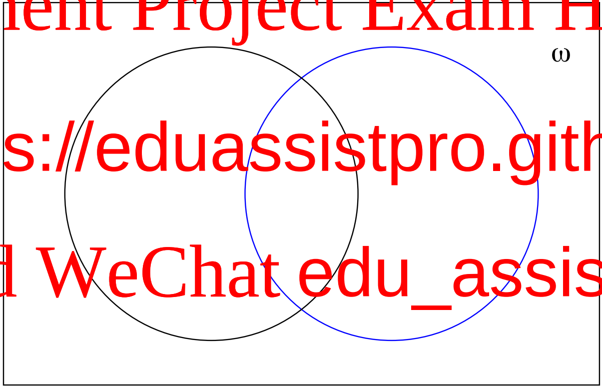
- *Joint probability* ($P(A, B)$):
  the probability of both A and B occurring $= P(A \cap B)$



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

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*: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
-
- $\ldots |\cap_{i=1}^{n-1} A_i)$

- *Prior probability* ($P(A)$): the proba o additional knowledge about A
- *Posterior probability* ($P(A|B)$): th iven background knowledge about eve
- *Independence: A* and *B* are independent iff $P(A \cap B) = P(A)P(B)$

Bayes Rule

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

For proposition A and evidence B,

Bayes' Rule is important because it allows u) given knowledge of the 'inverse' probability

For instance, imagine we believe (from prio $P(H1|Smart) = 0.6$, $P(Smart) = 0.3$, 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 ($\langle H, T, H, H, ..., T \rangle$)

- xactly m

  $$(\ , \ , \ ) \ = \ (\quad)$$

  $$P(m, n, p) \ = \ \underline{\qquad\qquad}$$

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

Lecture 4: Introduction to Probability Theory

COMP90049 Knowledge Technologies

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; $P(X = 2)$

ch possible

$(\ , \ ) = \frac{}{2!1!} =$

So, 3 possible outcomes, $\frac{1}{8}$ for each,

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

- 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);

- $p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ ti

$$P(X_1 = x_1, X_2 = x_2, ..., X$$

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

$$E = \sum_i -f$$

is the *entropy* of the message; this is the th 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'.

- 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*
-

- ... ales $1, \dots, n$

$$H(x) = -\sum_{i=1}^{n} P(i)\log_2 \; \dots$$

$$= \frac{freq(*)\log_2(fr \; \dots \; \sum_{i=1} \; \dots \; {}_2 \; q(i))}{freq(*)}$$

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

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.
  - e

$$g_2 P(X = t))$$

$$= -((0.5 * -1) + (0.5 \quad 1)) = (\quad 1) = 1$$

  - Two possible outcomes with equal
    Learning the outcome contains on

- 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

$P(X = i))$

$= \quad 0.47$

Entropy values

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

- Entropy=0 (minimum entropy) when one probability is 1, others 0
- Entropy=log(n) (maximum entrop equal values of 1/n

Summary

Probability forms the foundation of many knowledge technologies.

- What is entropy, and how should you in

Next: Approximate matching