Artificial Intelligence

academic year 2023/2024

Giorgio Fumera

Pattern Recognition and Applications Lab
Department of Electrical and Electronic Engineering
University of Cagliari (Italy)

Knowledge Representation and Inference under Uncertainty: Bayesian Networks

Notes on Probability Theory

Suggested textbooks

This course requires a basic knowledge of probability theory, including:

- random events and probability function
- conditional probability
- the concept of *random variable*
- probability distribution function of a discrete random variable

Bayesian neworks, as well as an informal introduction to probability theory, are covered by the course textbook:
S. Russell, P. Norvig, *Artificial Intelligence – A Modern Approach*, 4th Ed., Pearson, 2021 (or a previous edition)

For a formal introduction to probability theory, textbooks like the following one are suggested: A.M. Mood, F.A. Graybill, D.C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, 1991 / 1998

Acting under uncertainty

In real application scenarios, autonomous systems must make decisions and act under **uncertainty**, which can be due to many factors, such as imprecise sensor inputs. Some examples:

- decision support systems: medical diagnosis, etc.
- speech recognition from audio signals
- image recognition (OCR, object detection, object recognition, scene understanding, etc.)
- robot perception (sensors) to enable action (actuators), e.g., driverless cars

Rational decision-making under uncertainty

Decision theory has been used in AI since the 1990s, to allow the design of **rational agents** capable of decision-making in uncertain environments.

Decision theory is the mathematical study of strategies for optimal decision-making between options involving different risks or expectations of gain or loss depending on the outcome.

It involves two components:
Decision theory = **probability** theory + **utility** theory

- probability theory: evaluates how "likely" an event is
- utility theory: evaluates how "desirable" an event is

Main probabilistic approaches used in AI

- **Bayesian networks** (of Belief networks), for **static** decision problems
- Dynamic Bayesian networks, for **sequential** decision problems: Hidden Markov Models, Kalman filter
- Decision networks (including utility theory) for **one-shot** decision problems
- Planning/scheduling under uncertainty, for **sequential** decision problems

Notes on probability theory

The concept of *probability*

The concept of *probability* was formalized in the XVII century to model **long-term** predictability in **games of chance**, such as playing cards.

Probability theory for decision-making under uncertainty has become instrumental in areas such as:

- scientific disciplines like physics (e.g., quantum mechanics)
- economics and finance, e.g.:
  - insurance companies (life, household, car, etc.)
  - pension systems

Classical probability

Classical (**a priori**) probability can be applied to model "experiments" such as tossing a coin and throwing a dice, whose possible **outcomes** are assumed to be:

- mutually exclusive
- equally likely
- random

Under the above assumptions, the probability of an event of interest is **defined** as:

$$\frac{\text{number of favourable outcomes}}{\text{total number of outcomes}}$$

Classical probability: examples

- Tossing a coin: what is the probability of a head (or a tail)?
- Throwing a dice: what is the probability of face 3 up?
- Tossing **two** coins: what is the probability of getting two heads?
- Throwing **two** dice:
  - what is the probability of getting (6,6)?
  - what is the probability that the sum of the faces up is 6?
- Picking a card from a well shuffled deck of 52 cards:
  - what is the probability of picking an ace of hearts (♡)?
  - what is the probability of picking an ace or a spade (♠)?

Limits of classical probability

Classical probability cannot be used to compute the probability of events like the following ones, since its underlying assumptions are **not** valid in the corresponding scenarios:

- face 6 up after throwing a **loaded** dice
- a chip manufactured by a given company is faulty
- a MSc student of the University of Cagliari graduates within 3 years

Frequentist probability

When the assumptions underlying classical probabilty are not valid, but an experiment can be **repeated** under **similar, uniform conditions**, **frequentist** (or **a posteriori**) probability can be applied.

Frequentist proability **estimating** the probability of an event of interest E as its **relative frequency**:

$$\frac{\text{number of repetitions where } E \text{ occurs}}{\text{total number of repetitions}}$$

Limits of classical and frequentist probability

Common requirement of classical and frequentist probability: a **conceptual** experiment in which the outcomes can occur under **uniform conditions**.

This does not fit scenarios requiring the evaluation of the probability of events such as:

- a piano player breaking one of his hands within the next 10 years
- the third World War starting within 2024

This kind of event can be dealt with using **subjective** probability.

Probability model

Computing the probability of **complex** events using the above definitions of "probability" (classical, frequentist or subjective) can be difficult.

A **probability model** allows to compute the probability of complex events in terms of probabilities of simpler ones. It can be defined in terms of the following elements:

- **elementary event** (**sample**): one of the possible **mutually disjoint** outcomes of a conceptual experiment
- **sample space** Ω: the set of all elementary events
- **event**: any subset of the sample space, $A \subseteq \Omega$
- **event space**: the set of all possible events, $\mathcal{A} = \{A : A \subseteq \Omega\}$, including
  - the **impossible** event $\emptyset$
  - the **certain** event Ω

Probability model: examples

Define a **probability model** (i.e., the **elementary events**) for the following experiments:

- outcome of tossing a coin
- outcome of throwing a dice
- outcome of tossing two coins
- outcome of throwing two dice
- number of car accidents in Italy in 2024
- number of hours a lightbulb burns before burning out
- number of rain days in Cagliari during January 2025 and total rainfall (in cm)

Axiomatic definition of probability

Based on a probability model for a scenario of interest, the concept of **probability function** and a set of **axioms** can be defined to compute the probability of **any** event of interest.

A **probability function** is a function mapping from any event to a real value in the interval $[0, 1]$:

$$P : \mathcal{A} \mapsto [0, 1]$$

For any event $A \in \mathcal{A}$, $P[A]$ is the probability that **any** of the **mutually exclusive** elementary events belonging to A occurs.

Axiomatic definition of probability

A set of possible **axioms** for the probability function:

- $P[A] \geq 0$ for every $A \in \mathcal{A}$
- $P[\Omega] = 1$
- if $A_1 \in \mathcal{A}$ and $A_2 \in \mathcal{A}$ are **mutually exclusive** ($A_1 \cap A_2 = \emptyset$), then $P[A_1 \cup A_2] = P[A_1] + P[A_2]$

Some **theorems** that can be **derived** from the above axioms:

- $P[\emptyset] = 0$
- $P[\overline{A}] = 1 - P[A]$
- for any events A and B, $P[A] = P[A \cap B] + P[A \cap \overline{B}]$
- for any events A and B,
  $P[A \cup B] = P[A] + P[B] - P[A \cap B] \leq P[A] + P[B]$
- for any events A and B, if $A \subseteq B$ then $P[A] \leq P[B]$

Conditional probability

Often one needs to evaluate the probability of an event A, **after** another event B has occurred.

This is called **conditional probability**, and is denoted by $P[A|B]$.

For instance, one may be interested in the probability that:

- the sum of the faces up of two dice is greater than 6 (A), if one knows that the face up of one of them is 3 (B)
- a light bulb will last at least 1.000 hours (A), given that it already lasted 100 hours (B)

In principle, conditional probabilities can be directly computed using any of the applicable definitions of probability.

Conditional probability: examples

The probability that the sum of the faces up of two dice is greater than 6 (A), if one knows that the face up of one of them is 3 (B), can be computed using the **classical** definition of probability:

- the **possible outcomes** are 11:
  - six outcomes where the face up of the **second** dice is 3: $(1,3),(2,3),\ldots,(6,3)$
  - five more outcomes when the face up of the **first** dice is 3 (note that $(3,3)$ has already been considered among the previous outcomes): $(3,1),(3,2),(3,4),(3,5),(3,6)$
- the **favourable outcomes** are 6: $(4,3),(5,3),(6,3),(3,4),(3,5),(3,6)$

Therefore, $P[A|B] = \frac{6}{11}$

Conditional probability: examples

The probability that a light bulb will last at least 1.000 hours (A), given that it already lasted 100 hours (B), can be computed using the **frequentist** definition of probability.

To this aim one should draw a sample of N light bulbs, and try to keep them on for 1.000 hours:

- the "total number of repetitions" is the number N' of light bulbs that lasted 100 hours or more
- the number of repetitions where the "event of interest" occurs is the number N'' of light bulbs that lasted at least 1.000 hours, among the ones that lasted 100 hours or more (obviously, $N'' < N'$)

Therefore, $P[A|B] = \frac{N''}{N'}$.

Conditional probability: definition

Using the axiomatic definition of probability, the conditional probability of event A, given event B, is **defined** as:

$$P[A|B] = \begin{cases} \frac{P[A \cap B]}{P[B]}, & \text{if } P[B] > 0 \\ \text{undefined}, & \text{if } P[B] = 0 \end{cases}$$

This allows one to obtain the value of the conditional probability of interest in terms of the probabilities of two events that may be **easier** to compute:

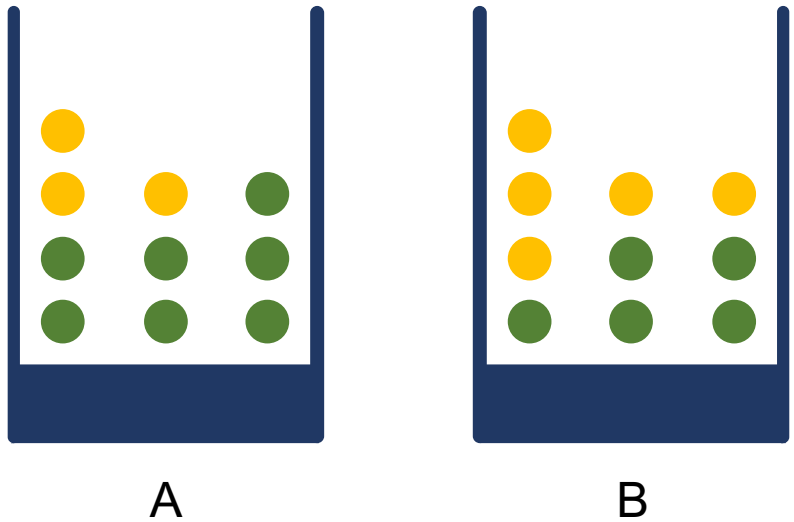
- $P[A \cap B]$: the probability that both A and B occur
- $P[B]$: the probability that B occurs

Conditional probability: definition

The above definition of conditional probability is **compatible** with the classical and frequentist definitions. For instance, in the above examples:

- throwing two dice:
  - $A = \{(4,3),(5,3),(6,3),(3,4),(3,5),(3,6)\}$
  - $B = \{(1,3),(2,3),\ldots,(6,3),(3,1),(3,2),(3,4),(3,5),(3,6)\}$
  - $A \cap B = A$
  - $P[A \cap B] = \frac{6}{36} = \frac{1}{6}$, $P[B] = \frac{11}{36}$, $\frac{P[A \cap B]}{P[B]} = \frac{6}{11}$
- duration of light bulbs:
  - A is the set of the N'' light bulbs, out of a given sample of N light bulbs, that lasted at least 1.000 hours
  - B is the set of the N' light bulbs, out of the same sample, that lasted at least 100 hours
  - $A \cap B = A$
  - $P[A \cap B] = \frac{N''}{N}$, $P[B] = \frac{N'}{N}$, $\frac{P[A \cap B]}{P[B]} = \frac{N''}{N'}$

Conditional probability: exercise



Two boxes A and B contain ten balls each: three yellow and seven green balls in box A, five yellow and five green in box B. Compute the probability that:

- a randomly picked ball from a randomly chosen box is green
- a randomly picked ball from box A is green

Conditional probability: product (chain) rule

From the definition of conditional probability

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \ ,$$

it immediately **follows** that:

$$P[A \cap B] = P[A|B] \cdot P[B]$$

The above equality is known as **product rule** or **chain rule**.

Bayes' formula

The product rule can be written in two **equivalent** ways, taking into account that $A \cap B$ and $B \cap A$ are the **same** event:

$$\begin{aligned} P[A \cap B] &= P[A|B] \cdot P[B] \\ P[B \cap A] &= P[B|A] \cdot P[A] \end{aligned}$$

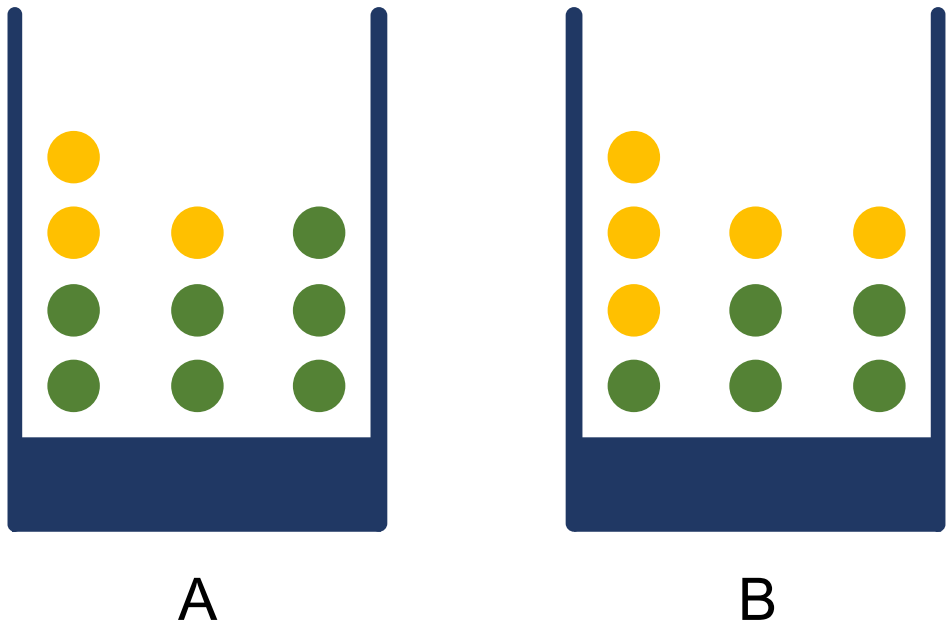
It easily **follows** that:

$$P[B|A] = \frac{P[A|B] \cdot P[B]}{P[A]}, \quad P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B]}$$

This is known as **Bayes' formula**.

Bayes' formula turns out to be very useful in many practical applications, when a conditional probability of interest (e.g., $P[B|A]$) is difficult to compute or estimate, whereas the opposite (e.g., $P[A|B]$) is easier.

Bayes' formula: exercise



Assume that a ball is randomly picked from a randomly chosen box, and its colour turns out to be green: what is the probability that the chosen box was A?

Bayes' formula: exercise

A doctor knows that 50% of the patients suffering from meningitis also have a stiff neck.

She also knows that meningitis affects one out of 50,000 people, whereas stiff neck affects one out of 20 people.

- What is the probability that patient John Smith, who has a stiff neck, also has meningitis?
- What if the doctor also knows that one out of 5,000 patients with stiff neck have meningitis?
- How would the computation of the above probability change, if there is a sudden epidemic of meningitis? (note that $P[S|M]$ is **unaffected** by the epidemic)

Independent events

Definition

Two events A and B are said to be **independent** (or **statistically independent**, or **stochastically independent**), if

$$P[A|B] = P[A]$$

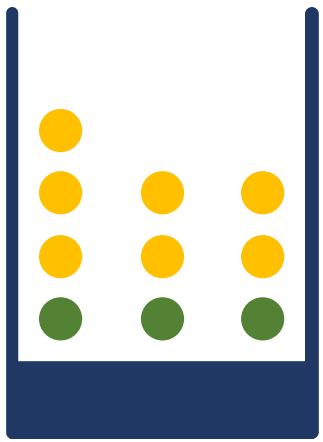
It is easy to prove that the above condition is **equivalent** to the following ones:

$$\begin{aligned} P[B|A] &= P[B] \\ P[A \cap B] &= P[A] \cdot P[B] \end{aligned}$$

Note that **statistical independence** between events **does not** imply the absence of a **cause–effect** relationship between them.

Independent events: exercises

- Consider two possible outcomes (events) after throwing one dice: $A =$ ⚀ or ⚁; $B =$ even number. What is $P[A|B]$?
- Consider three possible outcomes (events) after throwing two dice: $A =$ odd sum; $B = 1$ up on 1st dice; $C =$ sum is 7. Are A and B, A and C, B and C independent?
- What is the probability that two balls drawn **with replacement** from the box below are both green?



Random variables

Random variables are an alternative, **numerical** description of events of interest:

- random variables are denoted by symbols starting with an **uppercase** letter, e.g., X
- random variables have a **numerical** domain, e.g., $\mathbb{N}$, $\mathbb{R}$, $\{2, 3, \ldots, 12\}$, $\{0, 1\}$
- a generic value in the domain of a random variable is denoted with symbols starting with an **lowercase** letter, e.g., x
- every single value of a random variable corresponds to **one** possible **event**

Random variables: examples

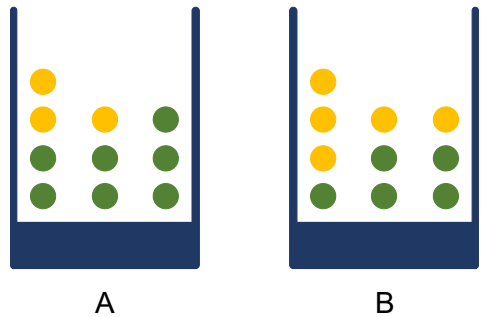
Two dice are thrown. Event $A =$ "the sum of the faces up is 7" corresponds to six elementary events that can be described in different ways, e.g.:

- $\{(1,6),(2,5),\ldots,(6,1)\}$
- $\{(\boxdot,\boxplus),(\boxdot,\boxplus),\ldots,(\boxplus,\boxdot)\}$

Alternatively, the **sum of the faces up of the two dice** can be represented as the value of a random variable $X \in \{2,3,\ldots,12\}$. Accordingly, event A corresponds to $X = 7$.

Random variables: examples

One of the boxes below is randomly chosen and a ball is randomly drawn from it.



Event $E =$ "Box A has been chosen" corresponds to ten elementary events that can be described by, e.g.:

$$\{(\mathtt{A}, \text{orange}), (\mathtt{A}, \text{orange}), \ldots, (\mathtt{A}, \text{green}), \ldots\}$$

Alternatively, the chosen box can be represented as the value of a random variable Y with domain $\{0, 1\}$ (where, e.g., 0 stands for A and 1 stands for B). Accordingly, event E corresponds to $Y = 0$.

Random variables: examples

Consider two events corresponding to the face up of a tossed coin:
$A = \{\text{head}\}$, $B = \{\text{tail}\}$.

The above events can be described by a random variable
$X \in \{0, 1\}$, where, e.g., 0 stands for “head” and 1 stands for “tail”.

Probability density function of a random variable

The **probability density function** (pdf) of a **discrete** random variable X with domain D_X is a function mapping from D_X to the interval $[0, 1]$:

$$f_X : D_X \mapsto [0, 1] \ .$$

Denoting by A the event corresponding to $X = x$, for any given $x \in D_X$, the value of $f_X(x)$ equals **by definition** $P[A]$.

The notation $X = x$ can be used as an alternative way to denote events; accordingly, the corresponding probability can also be denoted by $P[X = x]$.

Probability density function of a random variable

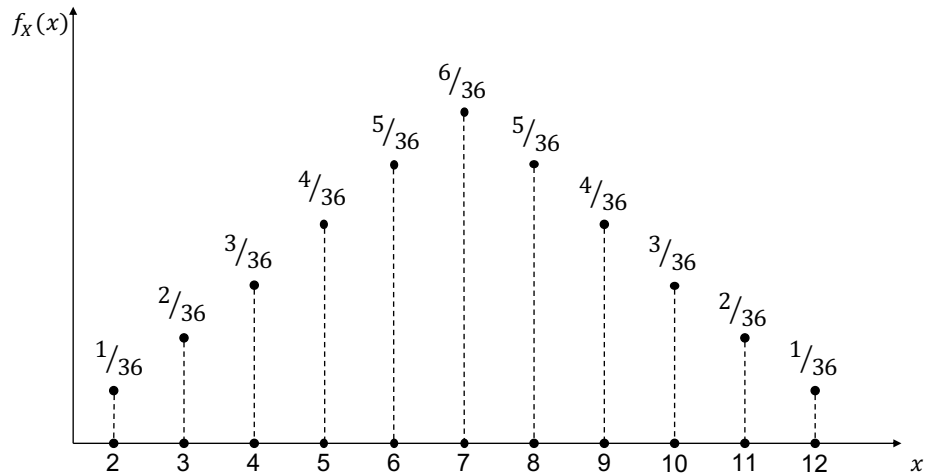
Note that the set of events $\{X = x : x \in D_X\}$ associated to **all** possible values of a random variable X, correspond by definition to the **whole** event space $\mathcal{A}$ of the underlying conceptual experiment.

From the axiomatic definition of probability it follows that

$$\sum_{x \in D_X} P[X = x] = 1 \ .$$

Probability density function: example

It is easy to see that the probability density function of a random variable X representing the sum of the faces up of two dice, where $X \in \{2, 3, \ldots, 12\}$, is the one graphically depicted as follows:



It is also easy to see that:

$$\sum_{x=2}^{12} P[X = x] = 1 \, .$$

Joint probability density function

If $X_1, \ldots, X_n$ are random variables defined over the **same** event space $\mathcal{A}$, then they are called **joint** random variables, and $(X_1, \ldots, X_n)$ is a n–dimensional random variable.

Their **joint probability density function** is defined as:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P[X_1 = x_1 \cap \ldots \cap X_n = x_n]$$

For instance, consider the following joint random variables related to the outcome of throwing two dice:

- X: face up of the first dice
- Y: highest value among the faces up

(X, Y) is a $2-$dimensional random variable with domain:
$\{(6,6), (5,6), (5,5), (4,6), (4,5), (4,4), \ldots, (1,6), \ldots, (1,1)\}$

Marginal probability density function and sum rule

If X, Y are **joint** discrete random variables, then $f_X(\cdot)$ and $f_Y(\cdot)$ are called **marginal** probability density functions.

By definition of the joint density function, the marginal density functions of the corresponding random variables can be obtained through the following **marginalization** operation:

$$\begin{aligned} f_X(x) &= \sum_{y \in D_Y} f_{X,Y}(x, y) \\ f_Y(y) &= \sum_{x \in D_X} f_{X,Y}(x, y) \end{aligned}$$

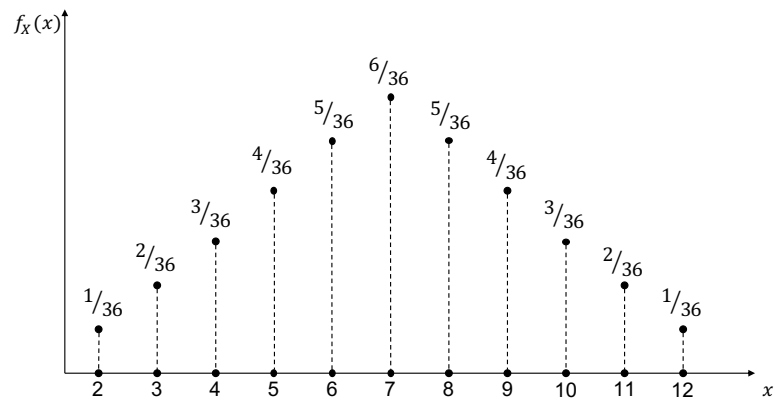
The above operation is also called **sum rule**.

Marginal density function and sum rule: example

Consider two random variables related to the outcome of throwing two dice: $X =$ sum of the faces up, $Y =$ face up of the first dice. Assuming $f_{X,Y}(x, y)$ is known, the marginal density of X can be obtained as follows:

$$f_X(x) = \sum_{y=1}^{6} f_{X,Y}(x, y)$$

This is the same probability distribution function previously shown:



Conditional density function

If X, Y are **joint** discrete random variables, then the **conditional** density function of Y given $X = x$ is defined as:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \ , \quad \text{if } f_X(x) > 0$$

Similarly:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \ , \quad \text{if } f_Y(y) > 0$$

Both functions are undefined, if $f_X(x) = 0$ or $f_Y(y) = 0$, respectively.

Independent random variables

Joint random variables $X_1, \ldots, X_n$ are said to be **independent**, if the following condition holds:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)$$

Exercises

1. An instructor observed from past oral examinations that students **knew** the correct answer to 70% of questions. Now she observes that a student **selected** the correct answer to a multiple-choice test with five alternatives, with only one of them correct. How can she compute the probability that the student **actually knew** the correct answer?
2. A cheater is known to use in about 30% of his games a stacked card deck in which the four "2"s are replaced with the corresponding aces (i.e., there are **eight** aces and **no** "2"s).
   - if, at the beginning of a game, a King is randomly drawn from the shuffled deck, what is the probability that the deck is stacked?
   - what if the drawn card is an ace?

Probability density function: notation

If A is an event defined on a given sample space Ω, and $X = x$ is the corresponding representation in terms of a random variable X, we have seen that the following expressions are equivalent:

$$f_X(x), \quad P[X = x], \quad P[A]$$

For ease of notation, from now on:

- the probability density **function** $f_X(\cdot)$ will be written as $P(X)$
- a **specific** value of the probability density function, $f_X(x)$, will be written as $P(X = x)$, or simply as $P(x)$

Joint density function

If the **joint density function** of the random variables of interest **is known**, it can be used to compute **any** probability involving all or some of them, using the **sum** and **product** rules.

Consider, for instance, four random variables A, B, C and D, and assume that their joint density function $P(A, B, C, D)$ is known:

- any **marginal** probability can be computed using the **sum rule** (marginalisation), e.g.:

$$P(A = a, B = b) \quad = \quad \sum_{c \in \mathcal{D}_C, d \in \mathcal{D}_D} P(A = a, B = b, C = c, D = d)$$

- any **conditional** probability can be computed using a combination of the **sum** and **product** rules, e.g.:

$$P(A = a | B = b) \quad = \quad \frac{P(A = a, B = b)}{P(B = b)} \ ,$$

then the sum rule can be applied as shown above to obtain $P(A = a, B = b)$ and $P(B = b)$.

Joint density function

However, the **computational complexity** of the **sum rule** is very high, since it requires to sum over **all possible combinations** of values of the random variables not involved in the marginal probability of interest.

For instance, given n Boolean random variables $X_1, \ldots, X_n$, the probability $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$ can be computed as:

$$\sum_{X_{k+1} \in \{0,1\}, X_{k+2} \in \{0,1\}, \ldots, X_n \in \{0,1\}} P(X_1 = x_1, \ldots, X_n = x_n) \ ,$$

which involves 2^{n-k} summands.

Joint density function

Moreover, in practical applications, computing or estimating the joint density function of the random variables of interest is often **infeasible**.

For instance, given n Boolean random variables $X_1, \ldots, X_n$, their joint density $P(X_1, \ldots, X_n)$ is defined by $2^n - 1$ values, such as

$$P(X_1 = \text{false}, X_2 = \text{true}, \ldots, X_n = \text{false})$$

(the remaining probability value can be derived by the constraint that all the 2^n probabilities sum up to 1).

Defining these $2^n - 1$ probability values may be feasible, if the **classical** definition of probability can be applied, e.g., if $X_1, \ldots, X_n$ denote the face up of n thrown dice.

It is likely to be infeasible, instead, if the **frequentist** definition of probability has to be used: if so, one should repeat the underlying experiment for a very large number of times to estimate each of the $2^n - 1$ probability values as the observed frequency of the corresponding events.

Joint density function: the *chain* rule

From the definition of conditional density function it is easy to derive the following relationship, known as **chain rule** (or multiplication rule):

$$
\begin{aligned}
& P(X_1, X_2, \ldots, X_n) \\
&= P(X_n|X_{n-1}, \ldots, X_1)P(X_{n-1}|X_{n-2}, \ldots, X_1) \cdots P(X_2|X_1)P(X_1) \\
&= \prod_{k=1}^{n} P(X_k|X_{k-1}, \ldots, X_1)
\end{aligned}
$$

Note that, given n random variables, the chain rule can be written in $n!$ different ways, corresponding to all their possible permutations. For instance, given four random variables A, B, C and D:

$$
\begin{aligned}
P(A, B, C, D) &= P(A|B, C, D)P(B|C, D)P(C|D)P(D) \\
P(A, B, C, D) &= P(B|A, D, C)P(A|D, C)P(C|D)P(D) \\
&\ldots
\end{aligned}
$$

Practical use of the chain rule

Does the chain rule allow to **reduce** the number of proability values required to specify the joint density function? The answer is **no**, as can be seen from the following example.

Consider again four Boolean random variables A, B, C and D, whose joint density $P(A, B, C, D)$ is fully specified by $2^4 - 1 = 15$ values.

If the joint density is rewritten, e.g., as:

$$P(A, B, C, D) = P(A|B, C, D)P(B|C, D)P(C|D)P(D) \ ,$$

then the number of probability values required to specify the density functions in the right-hand side can be obtained as follows.

Practical use of the chain rule

- $P(A|B, C, D)$ represents $2^3 = 8$ density functions of **one** Boolean random variable (A), e.g.:

  $$P(A|B = \text{true}, C = \text{false}, D = \text{true}) \ ;$$

  each one is fully specified by $2^1 - 1 = 1$ probability value, e.g.:

  $$P(A = \text{true}|B = \text{true}, C = \text{false}, D = \text{true}) \ ;$$

  therefore, $1 \times 2^3 = \mathbf{8}$ values are required
- $P(B|C, D)$ is fully specified by $1 \times 2^2 = \mathbf{4}$ values
- $P(C|D)$ is fully specified by $1 \times 2^1 = \mathbf{2}$ values
- $P(D)$ is fully specified by $\mathbf{1}$ value

Therefore, $8 + 4 + 2 + 1 = \mathbf{15}$ probability values have to be specified, as is the case with the joint density $P(A, B, C, D)$.

Practical use of the chain rule

However, if some of the random variables are **conditionally independent** given some other variables, then the chain rule allows their joint density to be defined in terms of a **lower** number of probability values.

For instance, consider again the above example involving the Boolean random variables A, B, C and D, and assume that:

- A is known, or assumed, to be independent of B and D, **given** C, i.e., $P(A|B, C, D) = P(A|C)$
- B is known, or assumed, to be independent of D, **given** C, i.e., $P(B|C, D) = P(B|C)$

Practical use of the chain rule

If, in the above example, the chain rule is applied as follows:

$$P(A, B, C, D) = \mathbf{P}(\mathbf{A}|\mathbf{B}, \mathbf{C}, \mathbf{D})\mathbf{P}(\mathbf{B}|\mathbf{C}, \mathbf{D})P(C|D)P(D) \ ,$$

then the above conditional independence relationships can be exploited to rewrite the joint density as:

$$P(A, B, C, D) = \mathbf{P}(\mathbf{A}|\mathbf{C})\mathbf{P}(\mathbf{B}|\mathbf{C})P(C|D)P(D)$$

How many probability values are required to define the joint density functions in the right-hand side of the above expression?

Practical use of the chain rule

Reasoning as in the previous example, one gets that:

- $P(A|C)$ is fully specified by $1 \times 2^1 = \mathbf{2}$ values
- $P(B|C)$ is fully specified by $1 \times 2^1 = \mathbf{2}$ values
- $P(C|D)$ is fully specified by $1 \times 2^1 = \mathbf{2}$ values
- $P(D)$ is fully specified by $\mathbf{1}$ value

Therefore, only $2 + 2 + 2 + 1 = \mathbf{7}$ values have to be specified, instead of 15.

As a **limit case**, if all four variables were **independent**, by definition their joint density would be given by

$$P(A, B, C, D) = P(A)P(B)P(C)P(D) \ ,$$

which requires only **4** probability values to be fully specified.

Bayesian networks

To sum up, in practical applications the number of probability values to be specified (e.g., estimated using the frequentist definition) to define the joint density of the random variables of interest can be reduced by suitably rewriting it through the **chain rule**, provided that some **conditional independence** assumptions can be made.

Bayesian networks are a very useful **graphical model** to represent conditional independence assumptions among the random variables of interest, as a **directed acyclic graph**.

They also ease the definition of **probabilistic inference algorithms** to compute **any** probability involving all or some of the corresponding random variables, as a function of the marginal density functions involved in the application of the chain rule.

Bayesian networks

See the course textbook for a detailed introduction to Bayesian networks:

S. Russell, P. Norvig, *Artificial Intelligence – A Modern Approach*, 4th Ed., Pearson, 2021 (or a previous edition)

The following topics are part of the course syllabus:

- the semantics of a Bayesian network
- defining the structure of a Bayesian network
- exact inference in Bayesian networks
- approximate inference: direct sampling, rejection sampling