Probability Theory as an Uncertainty Calculus

Florian Meyer

ECE 275A – Fall 2023

Setup of Probability Theory

- A *definite state* ω of the **world** Ω, $\omega \in \Omega$, is observed/measured to be the case once an experiment/observation is performed on the world.

- The "world" Ω is referred to by a variety of names, including the *universe*, the *sample space*, the *state space*, and the space/set of (underlying) *experimental outcomes* …

- If $\omega \in$ **E,** where **E** is a subset of Ω, **E** $\subset \Omega$, then we say that the ***event E*** has occurred. The collection of admissible events forms a Boolean algebra.

- *We are* ***uncertain*** *about whether* ***event E*** (a concept to be defined below) ***occurred, or will occur, in the world*** either because the measured/observed state value ω is hidden from us or is to be performed in the future.

  - This uncertainty is encoded as the ***Probability*** **P**(E).

- **Probabilities Satisfy the Kolmogorov Axioms (given later below).**

Fundamental First Steps in Probability Calculus

- Create a ***model world*** Ω of a domain of interest.

- Assume that ***all*** possible states, ω , of the domain of interest are *unambiguously defined* and contained in the model world, $\omega \in \Omega$.

  – ω = "state of the world"

- M*easurements* are taken of the world Ω **and one and only one *experimental outcome* ω *occurs***. The unique experimental outcome ω is called an ***instantiation*** of the world Ω.

Fundamental First Steps in Probability Calculus

For example, if a toss of a 6-sided die is our experiment/trial and we assume the sample space to be comprised of outcomes corresponding to each of the possible face-up positions,

$$\Omega_1 = \left\{ \boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}, \boxed{5}, \boxed{6} \right\},$$

then we do *not* admit the possibility of the die landing on an edge or corner. If these are to be admitted as experimental possibilities, then the sample space *must* be modified to reflect this fact,

$$\Omega_2 = \left\{ \boxed{1}, \boxed{2}, \boxed{3}, \boxed{4}, \boxed{5}, \boxed{6}, \boxed{\text{E}}, \boxed{\text{C}} \right\}.$$

Thus in order to construct the sample space, Ω, it is assumed in advance of performing an experiment that we know *all* possible outcomes. (Of course, because the experiment is random, we don't know *which* of the known possible outcomes in Ω will arise when the experiment is actually performed.) Note that the sample space is not uniquely defined as it can be larger than strictly needed. For example, both of the sample spaces Ω_1 and Ω_2 will suffice for an experiment for which one, and only one, of the six sides of a die will be observed.

Definition of a Random Event $A \sim \alpha$

$\alpha(\omega)$ = "ω satisfies property α" = 0-1 Boolean random variable

Trivially $\alpha(\omega) = 1$ iff $A = \{ \omega \mid \alpha(\omega) = 1 \} \subset \Omega$

Equivalence between propositional logic & subset algebra:

Logical **OR**: $\alpha \vee \beta$ iff $A \cup B$; Logical **FALSE**: false iff $\varnothing$

Logical **AND**: $\alpha \wedge \beta$ iff $A \cap B$; Logical **TRUE**: true iff Ω

Logical **NOT**: $\neg\alpha$ iff $A^c = \Omega - A$

(More rigorously, $\alpha(\omega) \vee \beta(\omega)$ iff $\omega \in A \cup B$, etc.)

Definition of an Event E – Cont.

*An **event E** is a **subset of possible states** of our model world* Ω.

For the die example a possible event $E \subset \Omega$ is

"Even" = E = { ω | e(ω) = "ω shows an even number of dots"}

If the outcome of a measurement of the world results in **any *single* outcome** $\omega \in E$ we say the event "Even" ***has occurred***.

The set of events forms a ***Boolean algebra*** (see previous slide) containing the ***impossible event*** $\varnothing$ = "no value was measured" and the ***certain event*** Ω = "some value was measured"

Probability Measure as a "Soft" Truth Function

- Regular logic has a "hard" 0-1 truth function **T** which acts on propositions ("did the event A occur") and returns a definite truth value of 1 if the proposition is true (i.e., is the case) and 0 if false

$$\mathbf{T(A) = 0 \text{ or } 1 \text{ for all events } A \subset \Omega}$$

- **T**(A) = 1 if and only if $\omega \in$ A. **T**(A) servers as an ***indicator function*** of the event A.

- However, we are generally *ignorant* about the actual state of affairs (i.e., what is the case) before, or even after, a measurement has been performed on the world. (We are ignorant until we are made aware of the outcome of the experiment.)

- We represent our ignorance about the occurrence (or not) of the event A in terms of a "soft" truth function (or belief function) known as a ***probability***

$$\mathbf{0 \leq P(A) \leq 1 \text{ for all events } A \subset \Omega, \text{ where } P(A) = E\{T(A)\}}$$

Properties of Probability **P**

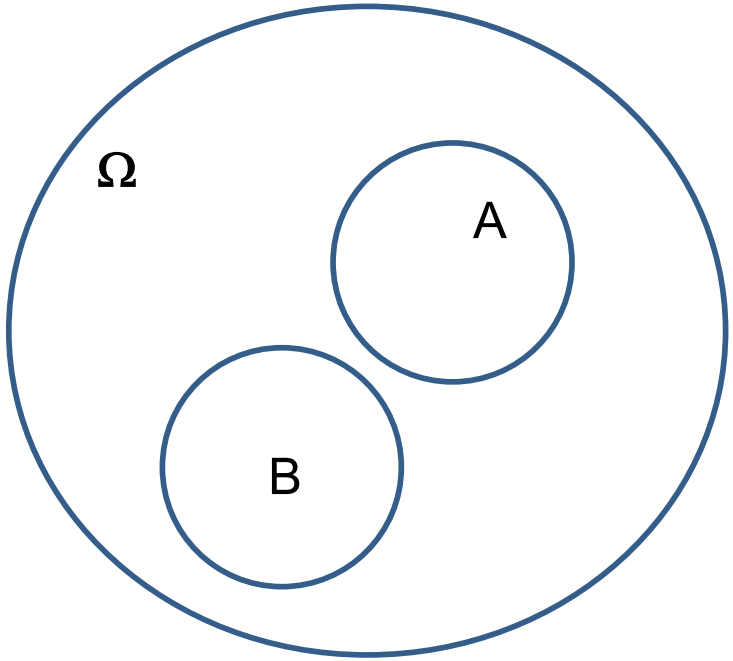
Ω = universal set, of exhaustive, mutually exclusive model world states ω.

A, B, C, D, … are events in the model world Ω (i.e. subsets of Ω)

P is a real-valued function of events satisfying

The 3 Kolmogorov Axioms

1. $0 \leq \mathbf{P}(A)$ for all events $A \subset \Omega$
2. $\mathbf{P}(\Omega) = 1$
3. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ iff $A \cap B = \varnothing$



These Axioms are Deceptively Simple

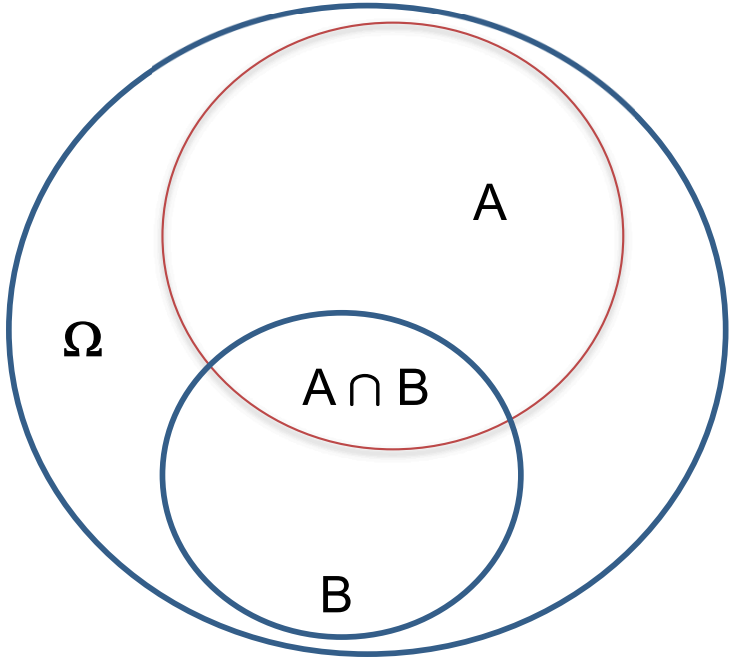
The <u>entire edifice</u> of modern probability theory is built up from these three axioms. An important consequence is

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$$

Proof: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B \setminus A)$

$= \mathbf{P}(A) + \mathbf{P}(B \setminus (A \cap B))$

$\mathbf{P}(B) = \mathbf{P}(B \setminus (A \cap B)) + \mathbf{P}(A \cap B)$



de Finitti (see the "coherence" discussion below, which is taken from the discussion in [Russell 2003]) has shown that if you are betting against someone who uses a probability calculus *violating* this consequence of Axiom 3, ***then you can beat him <u>every single time</u> you bet***, not just in the long run.

Additional Consequences of the Axioms

Here are some additional consequences and properties of the Kolmogorov Probability Axioms which are commonly explained in introductory probability courses. You should be able to readily prove them using the axioms and some basic properties of (naïve) set theory.

A. $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$. Hint: Note that $\Omega = A \cup A^c$ with A and A^c disjoint.

B. $\mathbf{P}(\emptyset) = 0$. Hint: Use the result of A.

C. If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$. Hint: Note that $B = A \cup (B \cap A^c)$.

Kolmogorov Axioms Yield "Coherence"

de Finitti assumes that the degree of belief that an agent has in a proposition ***a*** corresponds to the odds at which it is indifferent to a bet for or against ***a***. (Because in the long-run the agent expects to break even.)

Assume that Agent 1 holds the following degrees of beliefs about events involving propositions *a* and *b*:

$$\mathbf{P}(\boldsymbol{a}) = 0.4 \qquad \mathbf{P}(\boldsymbol{a} \wedge \boldsymbol{b}) = 0.0$$
$$\mathbf{P}(\boldsymbol{b}) = 0.3 \qquad \mathbf{P}(\boldsymbol{a} \vee \boldsymbol{b}) = 0.8$$

Agent 1's beliefs clearly violate the consequence of Axiom 3 discussed before. These beliefs are "incoherent" in the sense that a so-called "Dutch book" betting strategy is possible.

Figure 13.2 taken from [Russell 2003] shows that if Agent 2 bets \$4 on *a*, \$3 on *b*, and \$2 on ¬ (*a* ∨ *b*) then ***Agent 1 always loses regardless of the outcomes*** for events *a* and *b*.

Example from Russell & Norvig 2003

Agent 1		Agent 2		Outcome for Agent 1			
Proposition	Belief	Bet	Stakes	$a \wedge b$	$a \wedge \neg b$	$\neg a \wedge b$	$\neg a \wedge \neg b$
a	0.4	a	4 to 6	–6	–6	4	4
b	0.3	b	3 to 7	–7	3	–7	3
$a \vee b$	0.8	$\neg(a \vee b)$	2 to 8	2	2	2	–8
				–11	–1	–1	–1

Figure 13.2 Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of bets that guarantees a loss for Agent 1, no matter what the outcome of a and b.

Agent 2 bets \$4 on a, \$3 on b, and \$2 on $\neg (a \vee b)$

Agent 1 bets \$6 on $\neg a$, \$7 on $\neg b$, and \$8 on $a \vee b$

Probability Logic & Conditional Probability

Joint Probability of events A and B:

$$\mathbf{P}(A,B) = \mathbf{P}(A \cap B) = \mathbf{P}(B \cap A) = \mathbf{P}(B,A)$$

The ***conditional probabilities*** **P**(A|B) and **P**(B|A) are defined by

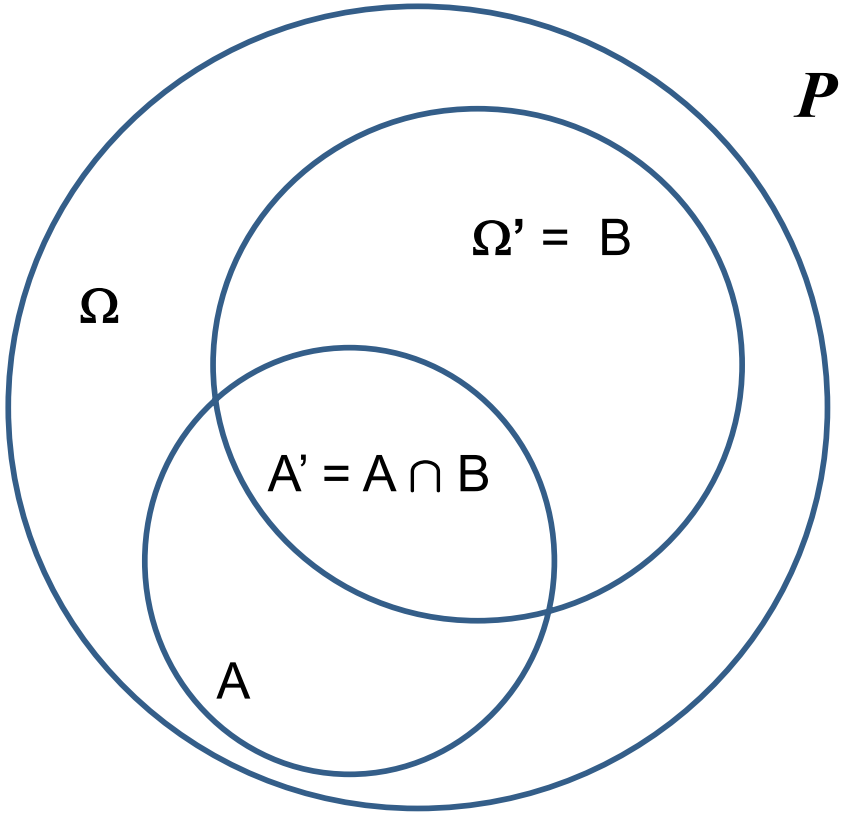
$$\mathbf{P}(A|B)\,\mathbf{P}(B) = \mathbf{P}(A,B) = \mathbf{P}(B,A) = \mathbf{P}(B|A)\,\mathbf{P}(A)$$

Or, assuming that **P**(A) and **P**(B) are nonzero, by simple rearrangement we obtain the two "inverse forms" of

Bayes' Rule for Events:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad \textbf{and} \quad P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Conditional Probability for Events A



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A')}{P(\Omega')} \triangleq P'(A')$$

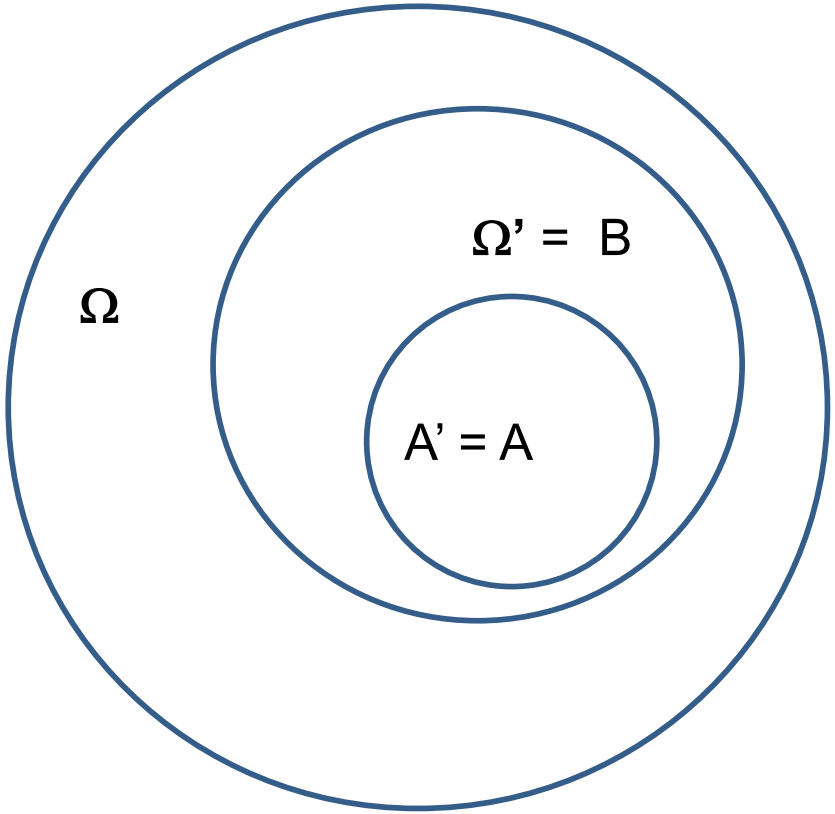
$$0 \le P(A \mid B) = P'(A') \le 1$$

$$P(\varnothing \mid B) = P'(\varnothing) = 0$$

$$P(B \mid B) = P'(\Omega') = 1$$

$$P(C \mid B) = 0 \quad \text{if} \quad C \cap B = \varnothing$$

Conditional Probability for Events A

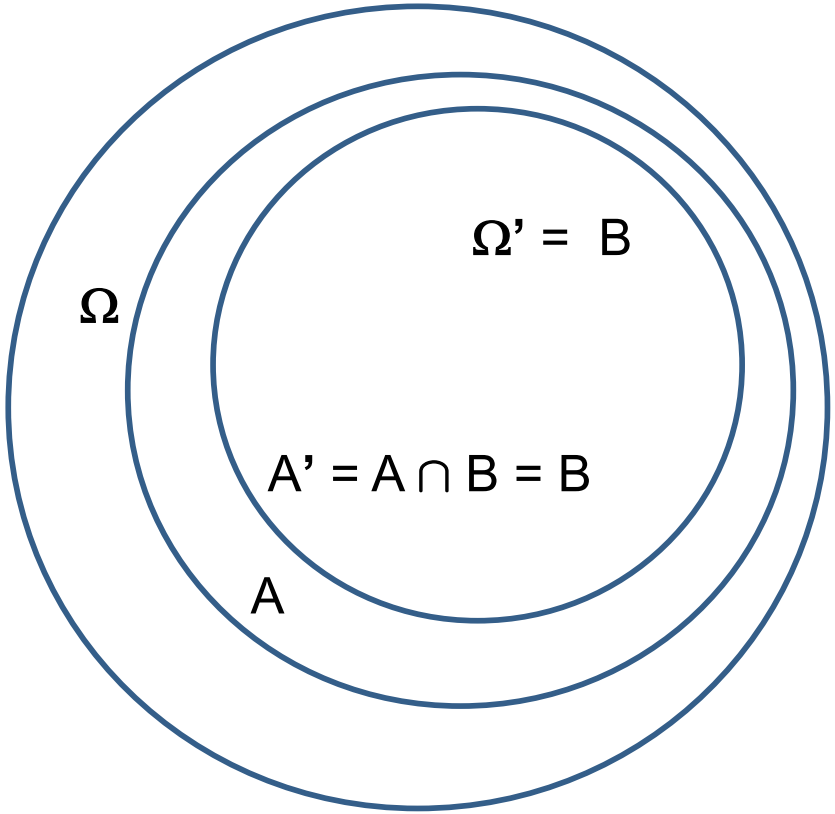


Suppose A ⊂ B (i.e., B is ***necessary***, but **not** sufficient, for A) then, since **P**(B) ≤ 1,

$$P(A \mid B) = \frac{P(A)}{P(B)} \geq \frac{P(A)}{1} = P(A)$$

Showing that although necessity does **<u>not</u> guarantee** A, *it does make it **more likely***.

Conditional Probability for Events A



Suppose now $B \subset A$, so that B is **sufficient** for A (i.e., B guarantees A) .

Then

$$P(A \mid B) = \frac{P(B)}{P(B)} = 1$$

Independence of Events

Two events *A* and *B* are said to be ***independent*** , A $\perp\!\!\!\perp$ B , iff

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

This is ***equivalent*** to the condition,

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

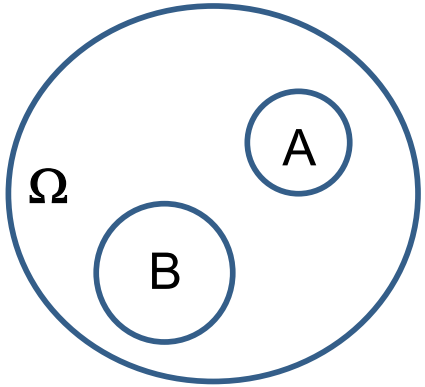
and to the condition,

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

I.e., all three of these conditions are equivalent.

If A and B are not independent, they are said to be ***dependent****.*

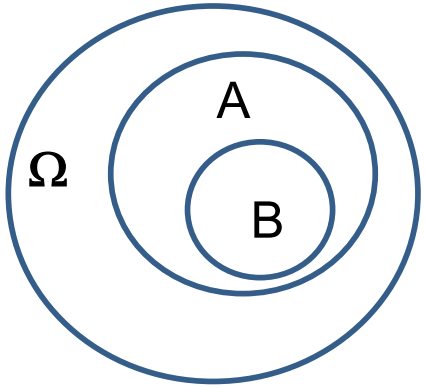
Extremes of Dependence



A and B are dependent

P(A) > 0 while **P**(A|B) = 0

Given B, event A is almost surely *impossible*



A and B are dependent

P(A) < 1 while **P**(A|B) = 1

Given B, event A is almost surely *a certainty*

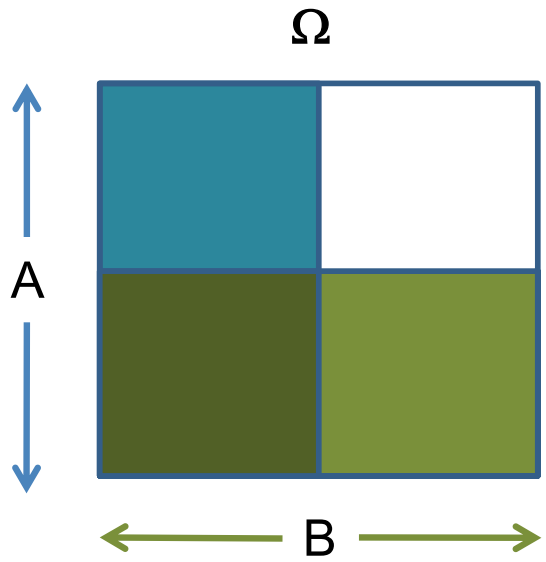
Dependence of A and B means you ***would*** *change a bet based on knowledge of probabilities* regarding the occurrence of A if you were given knowledge of B.

Independence

P(A) = 0.5
P(B) = 0.5

P(A ∩ B) = 0.25 = **P**(A) **P**(B)

P(A|B) = 0.5 = **P**(A)
P(B|A) = 0.5 = **P**(B)



Events A and B are independent.

Independence means you would ***not*** change your bet.

Note that *knowledge of B* ***does*** *give you* ***knowledge of A*** even though B and A are independent (so that you would not change your bet, which is based on probabilities).

Definition of a Random Variable **X**

A *random variable* **X** is a *real-valued (or complex-valued)* ***function*** of experimental outcomes ω:

For all $\omega \in \Omega$, **X (ω) =** x $\in$ Reals (or Imaginaries)

We usually get sloppy and don't carefully distinguish between the random variable **X** and the value x, denoting both by x and relying on context to keep the two very different meanings straight. This can cause beginning students much confusion.

Because it is a *function*, **X** **must** be single-valued. Furthermore, **X** is a ***total function***, meaning that it is defined *everywhere* on its domain Ω.

This means that **X** *induces a disjoint partition* of Ω. In this manner a measurement of **X** yields information which reduces uncertainty about which value of ω is the instantiated experimental outcome.

Random Variables Naturally Define Events

$\{ \mathbf{X} = x \} = \{ \omega \mid \mathbf{X}(\omega) = x \} \subset \Omega$

$\{ \mathbf{X} \leq x \} = \{ \omega \mid \mathbf{X}(\omega) \leq x \} \subset \Omega$

$\{ \mathbf{X} < \infty \} = \{ \mathbf{X} > -\infty \} = \Omega$ = the event “always true”

$\{ \mathbf{X} \geq \infty \} = \varnothing$ = the event “never true”

Etc.

Random Variables (RVs)

It is convenient now to move to events defined by the values taken by *random variables.* (See the first example of the previous slide.)

Since an *event* A *corresponds* to the Boolean 0-1 *random variable* (or indicator function) $\alpha(\omega)$ that takes the value 1 iff A occurs, this results in a more general (and more convenient) form of Bayes Rule.

To simplify the discussion, in the remainder of these slides *assume* that our random variables take on a finite number *n* of distinct, discrete possible values (*n* = 2, in the case of boolean random variables),

$$X(\omega) = x \in \{x_1, \cdots, x_n\}$$

(E.g., in the boolean case the set of admissible values is {0,1}.)

RVs & Probability Distributions

Setting $A_i = \{X = x_i\} = \{\omega \mid X(\omega) = x_i\}$

Define the (marginal = individual) ***probability distribution***,

$$p_i = p(x_i) = p_X(x_i) \triangleq P(A_i) = P(X = x_i)$$

Obviously $0 \le p_i = p(x_i) \le 1$

Probability Distributions

Because the RV ***X*** is a total function on the sample space **Ω**, the events A_i , $i = 1, \ldots, n$, form a *disjoint partition* of **Ω**,

$$\Omega = A_1 \cup \cdots \cup A_n \,, \quad A_i \cap A_j = \varnothing \,, \quad \forall i \neq j$$

By induction on the fact that

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) \text{ iff } A \cap B = \varnothing$$

we see that a probability distribution *must be normalized*.

$$1 = P(\Omega) = P(A_1) + \cdots + P(A_n) = p_1 + \cdots + p_n$$

This is **an important result** which holds even if **X** is a continuous valued random variable (in which case the sum goes over to an integral).

Jointly Random Variables

Suppose we have two random variables **X** and **Y**. With,

$$A_i = \{X = x_i\}, \quad B_j = \{Y = y_j\}, \quad \textbf{and} \quad C_{ij} = A_i \cap B_j$$

we define the *joint distribution* $p(x_i, y_j)$ by

$$p(x_i, y_j) \triangleq P(X = x_i, Y = y_j) = P\left(\{X = x_i\} \cap \{Y = y_j\}\right) = P(A_i \cap B_j) = P(C_{ij})$$

The sets C_{ij} (induced by a measurement of **X** ***and*** **Y**) form a disjoint partition of the sample space Ω which is generally finer than that induced by **X** ***or*** **Y** ***alone***. I.e., measurements of two RVs generally provide more information (in the sense of being able to localize the instantiated value of ω) than a measurement of just one RV alone.

Marginalization

We have

$$\sum_j p(x_i, y_j) = \sum_j P\left(\{X = x_i\} \cap \{Y = y_j\}\right) = P\left(\cup_j \{X = x_i\} \cap \{Y = y_j\}\right)$$
$$= P\left(\{X = x_i\} \cap \cup_j \{Y = y_j\}\right) = P\left(\{X = x_i\} \cap \Omega\right)$$
$$= P\left(\{X = x_i\}\right) = p(x_i)$$

A process known as ***marginalization.***

Similarly,

$$p(y_j) = \sum_i p(x_i, y_j)$$

Marginalization

Suppose that 3 batteries are randomly chosen from a group of 3 new, 4 used but still working, and 5 defective batteries. If we let X and Y denote, respectively, the number of new and used but still working batteries that are chosen, then the joint probability mass function of X and Y, $p(i,j) = P\{X = i, Y = j\}$, is given by

TABLE 4.1 $P\{X = i, Y = j\}$

i \ j	0	1	2	3	Row Sum $= P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column Sums $= P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

[Ross 2004]

Note that the sums are placed in the bottom and side ***margins***, hence ***marginalization***.

Conditional Probability for RVs

We now define conditional probability for ***random variables***,

$$p(x,y) = P(X = x, Y = y) = P(X = x \mid Y = y)P(Y = y) = p(x \mid y)p(y)$$

where

$$p(x \mid y) \triangleq P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x,y)}{p(y)}, \quad p(y) \neq 0$$

Note that by induction we have *the product rule of conditional probabilities,*

$$P(A \cap B \cap C) = P(A \mid B \cap C)P(B \cap C) = P(A \mid B \cap C)P(B \mid C)P(C)$$

For random variables, the product rule becomes,

$$p(x,y,z) = p(x \mid y,z)p(y,z) = p(x \mid y,z)p(y \mid z)p(z)$$

Independence of Random Variables

Two random variables ***X*** and ***Y*** are said to be ***independent*** iff

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})\, p(\mathbf{y})$$

We write $X \perp\!\!\!\perp Y$. Equivalent statements are

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}) \quad \text{and} \quad p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})$$

If **X** and **Y** are not independent, they are said to be ***dependent,*** and we write $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}$

Bayes' Rule For Random Variables

Note that marginalization can be written as

$$p(y) = \sum_x p(x, y) = \sum_x p(y \mid x) p(x)$$

Bayes' Rule for Random Variables is then given by

$$p(x \mid y) = \frac{p(y \mid x) p(x)}{p(y)} = \frac{p(y \mid x) p(x)}{\sum_{x'} p(y \mid x') p(x')}$$

$p(x) =$ ***prior probability*** **of** x

$p(x \mid y) =$ ***posterior probability*** **of** x **given** ***evidence*** y

$p(y \mid x) =$ ***likelihood*** **of** y **given** x

$p(y) =$ ***evidence for probability model*** $p(\cdot)$ **given** y

Example of Inference via Bayes' Rule

Classic medical test example

An fast and inexpensive medical test is 99% effective in detecting a very rare, but fatal, disease if, in fact, a patient has it. However, the test has a false positive rate of 2%. If a patient tests positive then he/she has to take a much more expensive, but definitive, test which takes about two weeks to process, during which time the patient (understandably) is experiencing some anxiety.

If 0.001% of the general population is known to have the disease, what is the probability that a patient chosen at random who tests positive on the initial test actually has it?

Solution

Let x = d be the 0-1 random variable that indicates if the patient has the disease and y = ***t*** be the 0-1 random variable that the test result is positive. Then from Bayes' rule:

$$p(d=0 \mid t=1) = \frac{p(t=1 \mid d=0)\,p(d=0)}{p(t=1 \mid d=0)\,p(d=0) + p(t=1 \mid d=1)\,p(d=1)} = \frac{(0.02)(0.99999)}{(0.02)(0.99999)+(0.99)(0.00001)}$$

= 99.95% ***posterior probability*** that the patient does ***not*** have the disease even thought the ***likelihood*** of the disease given the positive test is 99%.

References

[Bishop 2006] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[Charniak 1991] E. Charniak, "Bayesian Networks without Tears," *AI Magazine*, pp. 50-53, Winter 1991.

[Ross 2004] S. Ross, Intro. To Prob. & Statistics for Engineers & Scientists, Academic Press, 2004.

[Russell 2003] S. Russell & P. Norvig, *Artificial Intelligence: A Modern Approach*, 2e, Prentice-Hall, 2003.

[Wuketits 1990] F. Wuketits, *Evolutionary Epistemology and its Implications for Humankind,* SUNY Press, 1990.