1

Bayesian Reasoning

1.1 Reasoning under uncertainty

Artificial intelligence (AI), should it ever exist, will be an intelligence developed by humans, implemented as an artifact. The level of intelligence demanded by Alan Turing's famous test (1950) — the ability to fool ordinary (unfoolish) humans about whether the other end of a dialogue is being carried on by a human or by a computer — is some indication of what AI researchers are aiming for. Such an AI would surely transform our technology and economy. We would be able to automate a great deal of human drudgery and paperwork. Since computers are universal, programs can be effortlessly copied from one system to another (to the consternation of those worried about intellectual property rights!), and the labor savings of AI support for bureaucratic applications of rules, medical diagnosis, research assistance, manufacturing control, etc. promises to be enormous. If a serious AI is ever developed.

There is little doubt that an AI will need to be able to reason logically. An inability to discover, for example, that a system's conclusions have reached inconsistency is more likely to be debilitating than the discovery of an inconsistency itself. For a long time there has also been widespread recognition that practical AI systems shall have to cope with **uncertainty** — that is, they shall have to deal with incomplete evidence leading to beliefs that fall short of knowledge, with fallible conclusions and the need to recover from error, called **non-monotonic reasoning**. Nevertheless, the AI community has been slow to recognize that any serious, general-purpose AI will need to be able to reason probabilistically, what we call here **Bayesian reasoning**.

There are at least three distinct forms of uncertainty which an intelligent system operating in anything like our world shall need to cope with:

1. **Ignorance.** The limits of our knowledge lead us to be uncertain about many things. Does our poker opponent have a flush or is she bluffing?
2. **Physical randomness or indeterminism.** Even if we know everything that we might care to investigate about a coin and how we impart spin to it when we toss it, there will remain an inescapable degree of uncertainty about whether it will land heads or tails when we toss it. A die-hard determinist might claim otherwise, that some unimagined amount of detailed investigation might someday reveal which way the coin will fall; but such a view is for the foreseeable future a mere act of scientistic faith. We are all practical indeterminists.
3. **Vagueness.** Many of the predicates we employ appear to be vague. It is often

unclear whether to classify a dog as a spaniel or not, a human as brave or not, a thought as knowledge or opinion.

Bayesianism is the philosophy that asserts that in order to understand human opinion as it ought to be, constrained by ignorance and uncertainty, the probability calculus is the single most important tool for representing appropriate strengths of belief. In this text we shall present Bayesian computational tools for reasoning with and about strengths of belief as probabilities; we shall also present a Bayesian view of physical randomness. In particular we shall consider a probabilistic account of causality and its implications for an intelligent agent's reasoning about its physical environment. We will not address the third source of uncertainty above, vagueness, which is fundamentally a problem about semantics and one which has no good analysis so far as we are aware.

1.2 Uncertainty in AI

The successes of formal logic have been considerable over the past century and have been received by many as an indication that logic should be the primary vehicle for **knowledge representation** and reasoning within AI. **Logicism** in AI, as this has been called, dominated AI research in the 1960s and 1970s, only losing its grip in the 1980s when artificial neural networks came of age. Nevertheless, even during the heyday of logicism, any number of practical problems were encountered where logic would not suffice, because uncertain reasoning was a key feature of the problem. In the 1960s, medical diagnosis problems became one of the first attempted application areas of AI programming. But there is no symptom or prognosis in medicine which is strictly logically implied by the existence of any particular disease or syndrome; so the researchers involved quickly developed a set of "probabilistic" relations. Because probability calculations are hard — in fact, NP hard in the number of variables (Cooper, 1990) (i.e., computationally intractable; see §1.11) — they resorted to implementing what has subsequently been called "naive Bayes" (or, "Idiot Bayes"), that is, probabilistic updating rules which assume that symptoms are independent of each other given diseases.[1]

The independence constraints required for these systems were so extreme that the systems were received with no wide interest. On the other hand, a very popular set of expert systems in the 1970s and 1980s were based upon Buchanan and Shortliffe's MYCIN, or the uncertainty representation within MYCIN which they called **certainty factors** (Buchanan and Shortliffe, 1984). Certainty factors (CFs) were obtained by first eliciting from experts a "degree of increased belief" which some evidence e should imply for a hypothesis h, $MB(h,e) \in [0,1]$, and also a corresponding

[1] We will look at naive Bayes models for prediction in Chapter 7.

"degree of increased disbelief," $MD(h,e) \in [0,1]$. These were then combined:

$$CF(h,e) = MB(h,e) - MD(h,e) \in [-1,1]$$

This division of changes in "certainty" into changes in belief and disbelief reflects the curious notion that belief and disbelief are not necessarily related to one another (cf. Buchanan and Shortliffe, 1984, section 11.4). A popular AI text, for example, sympathetically reports that "it is often the case that an expert might have confidence 0.7 (say) that some relationship is true and have no feeling about it being not true" (Luger and Stubblefield, 1993, p. 329). The same point can be put more simply: experts are often inconsistent. Our goal in Bayesian modeling is, at least largely, to find the most accurate representation of a real system about which we may be receiving inconsistent expert advice, rather than finding ways of modeling the inconsistency itself.

Regardless of how we may react to this interpretation of certainty factors, no operational semantics for CFs were provided by Buchanan and Shortliffe. This meant that no real guidance could be given to experts whose opinions were being solicited. Most likely, they simply assumed that they were being asked for conditional probabilities of h given e and of $\neg h$ given e. And, indeed, there finally was a probabilistic semantics given for certainty factors: David Heckerman (1986) proved that a consistent probabilistic interpretation of certainty factors[2] would once again require strong independence assumptions: in particular that, when combining multiple pieces of evidence, the different pieces of evidence must always be independent of each other. Whereas this appears to be a desirable simplification of **rule-based** systems, allowing rules to be "modular," with the combined impact of diverse evidence being a compositional function of their separate impacts it is easy to demonstrate that the required independencies are frequently unavailable. The price of rule-based simplicity is irrelevance.

Bayesian networks provide a natural representation of probabilities which allow for (and take advantage of, as we shall see in Chapter 2) any independencies that may hold, while not being limited to problems satisfying strong independence requirements. The combination of substantial increases in computer power with the Bayesian network's ability to use any existing independencies to computational advantage make the approximations and restrictive assumptions of earlier uncertainty formalisms pointless. So we now turn to the main game: understanding and representing uncertainty with probabilities.

1.3 Probability calculus

The probability calculus allows us to represent the independencies which other systems require, but also allows us to represent any dependencies which we may need.

[2] In particular, a mapping of certainty factors into likelihood ratios.

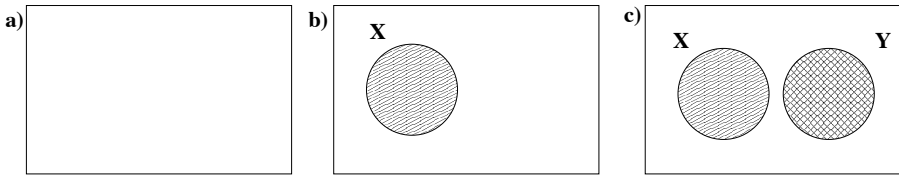

FIGURE 1.1: (a) The event space U; (b) $P(X)$; (c) $P(X \cup Y)$.

The probability calculus was specifically invented in the 17th century by Fermat and Pascal in order to deal with the problems of physical uncertainty introduced by gambling. But it did not take long before it was noticed that the concept of probability could be used in dealing also with the uncertainties introduced by ignorance, leading Bishop Butler to declare in the 18th century that "probability is the very guide to life." So now we introduce this formal language of probability, in a very simple way using Venn diagrams.

Let U be the universe of possible events; that is, if we are uncertain about which of a number of possibilities is true, we shall let U represent all of them collectively (see Figure 1.1(a)). Then the maximum probability must apply to the true event lying within U. By convention we set the maximum probability to 1, giving us Kolmogorov's first axiom for probability theory (Kolmogorov, 1933):

Axiom 1.1 $P(U) = 1$

This probability mass, summing or integrating to 1, is distributed over U, perhaps evenly or perhaps unevenly. For simplicity we shall assume that it is spread evenly, so that the probability of any region is strictly proportional to its area. For any such region X its area cannot be negative, even if X is empty; hence we have the second axiom (Figure 1.1(b)):

Axiom 1.2 *For all* $X \subseteq U, P(X) \geq 0$

We need to be able to compute the probability of combined events, X and Y. This is trivial if the two events are mutually exclusive, giving us the third and last axiom (Figure 1.1(c)), known as **additivity**:

Axiom 1.3 *For all* $X,Y \subseteq U$, if $X \cap Y = \emptyset$, then $P(X \cup Y) = P(X) + P(Y)$

Any function over a field of subsets of U satisfying the above axioms will be a probability function.[3]

A simple theorem extends addition to events which overlap (i.e., sets which intersect):

Theorem 1.1 *For all* $X,Y \subseteq U, P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$.

[3] A set-theoretic field is a set of sets containing U and $\emptyset$ and is closed under union, intersection and complementation.

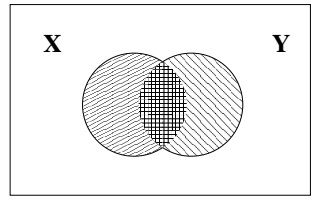

FIGURE 1.2: Conditional probability: $P(X|Y) = P(X \cap Y)/P(Y)$.

This can be intuitively grasped from Figure 1.2: the area of $X \cup Y$ is less than area of X plus the area of Y because when adding the area of intersection $X \cap Y$ has been counted twice; hence, we simply remove the excess to find $P(X \cup Y)$ for any two events X and Y.

The concept of **conditional probability** is crucial for the useful application of the probability calculus. It is usually introduced by definition:

Definition 1.1 Conditional probability

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

That is, given that the event Y has occurred, or will occur, the probability that X will also occur is $P(X|Y)$. Clearly, if Y is an event with zero probability, then this conditional probability is undefined. This is not an issue for probability distributions which are **positive**, since, by definition, they are non-zero over every event. A simple way to think about probabilities conditional upon Y is to imagine that the universe of events U has shrunk to Y. The conditional probability of X on Y is just the measure of what is left of X relative to what is left of Y; in Figure 1.2 this is just the ratio of the darker area (representing $X \cap Y$) to the area of Y. This way of understanding conditional probability is justified by the fact that the conditional function $P(\cdot|Y)$ is itself a probability function[4] — that is, it provably satisfies the three axioms of probability.

Another final probability concept we need to introduce is that of **independence** (or, marginal independence). Two events X and Y are probabilistically independent (in notation, $X \perp\!\!\!\perp Y$) whenever conditioning on one leaves the probability of the other unchanged:

Definition 1.2 Independence $X \perp\!\!\!\perp Y \equiv P(X|Y) = P(X)$

This is provably symmetrical: $X \perp\!\!\!\perp Y \equiv Y \perp\!\!\!\perp X$. The simplest examples of independence come from gambling. For example, two rolls of dice are normally independent. Getting a one with the first roll will neither raise nor lower the probability of getting a one the second time. If two events are **dependent**, then one coming true will *alter* the probability of the other. Thus, the probability of getting a diamond flush in poker

[4] $P(\cdot|Y)$ is just the function equal to $P(X|Y)$ for all $X \subset U$.

(five diamonds in five cards drawn) is *not* simply $(1/4)^5 = 1/1024$: the probability that the first card drawn being a diamond is $1/4$, but the probability of subsequent cards being diamonds is influenced by the fact that there are then fewer diamonds left in the deck.

Conditional independence generalizes this concept to X and Y being independent given some additional event Z:

Definition 1.3 Conditional independence $X \perp\!\!\!\perp Y|Z \equiv P(X|Y,Z) = P(X|Z)$

This is a true generalization because, of course, Z can be the empty set $\emptyset$, when it reduces to marginal independence. Conditional independence holds when the event Z tells us everything that Y does about X and possibly more; once you know Z, learning Y is uninformative. For example, suppose we have two diagnostic tests for cancer X, an inexpensive but less accurate one, Y, and a more expensive and more accurate one, Z. If Z is more accurate partly because it effectively incorporates all of the diagnostic information available from Y, then knowing the outcome of Z will render an additional test of Y irrelevant — Y will be "screened off" from X by Z.

1.3.1 Conditional probability theorems

We introduce without proof two theorems on conditional probability which will be of frequent use:

Theorem 1.2 Total Probability *Assume the set of events $\{A_i\}$ is a partition of U; i.e., $\bigcup_i A_i = U$ and for any distinct i and j $A_i \cap A_j = \emptyset$. Then*

$$P(U) = \sum_i P(A_i)$$

We can equally well partition the probability of any particular event B instead of the whole event space. In other words, under the above conditions (and if $\forall i A_i \neq \emptyset$),

$$P(B) = \sum_i P(B \cap A_i)$$

We shall refer to either formulation under the title "Total Probability".

Theorem 1.3 The Chain Rule *Given three events A,B,C in a chain of influence (i.e., A and C independent given B),*

$$P(C|A) = P(C|B)P(B|A) + P(C|\neg B)P(\neg B|A)$$

assuming the conditional probabilities are defined. This allows us to divide the probabilistic influence of C on A across the different states of a third variable. (Here, the third variable is binary, but the theorem is easily generalized to variables of arbitrary arity.)

1.3.2 Variables

Although we have introduced probabilities over events, in most of our discussion we shall be concerned with probabilities over **random variables**. A random variable is a variable which reports the outcome of some measurement process. It can be related to events, to be sure. For example, instead of talking about which event in a partition $\{A_i\}$ turns out to be the case, we can equivalently talk about which state x_i the random variable X takes, which we write $X = x_i$. The set of states a variable X can take form its state space, written Ω_X, and its size (or arity) is $|\Omega_X|$.

The discussion thus far has been implicitly of discrete variables, those with a finite state space. However, we need also to introduce the concept of probability distributions over **continuous variables**, that is, variables which range over real numbers, like *Temperature*. For the most part in this text we shall be using probability distributions over discrete variables (events), for two reasons. First, the Bayesian network technology is primarily oriented towards handling discrete state variables, for example the inference algorithms of Chapter 3. Second, for most purposes continuous variables can be **discretized**. For example, temperatures can be divided into ranges of ± 5 degrees for many purposes; and if that is too crude, then they can be divided into ranges of ± 1 degree, etc.

Despite our ability to evade probabilities over continuous variables much of the time, we shall occasionally need to discuss them. We introduce these probabilities by first starting with a **density function** $f(X)$ defined over the continuous variable X. Intuitively, the density assigns a weight or measure to each possible value of X and can be approximated by a finely partitioned histogram reporting samples from X. Although the density is not itself a probability function, it can be used to generate one so long as $f(\cdot)$ satisfies the conditions:

$$f(x) \geq 0 \tag{1.1}$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{1.2}$$

In words: each point value is positive or zero and all values integrate to 1. In that case we can define the **cumulative probability distribution** $F(\cdot)$ by

$$F(x) = P(X \leq x) = \int_{x' \leq x} f(x')dx' \tag{1.3}$$

This function assigns probabilities to ranges from each possible value of x down to negative infinity. Note that we can analogously define probabilities over any continuous interval of values for X, so long as the interval is not degenerate (equal to a point). In effect, we obtain a probability distribution by discretizing the continuous variable — i.e., by looking at the mass of the density function over intervals.

1.4 Interpretations of probability

There have been two main contending views about how to understand probability. One asserts that probabilities are fundamentally dispositional properties of non-deterministic physical systems, the classical such systems being gambling devices, such as dice. This view is particularly associated with **frequentism**, advocated in the 19th century by John Venn (1866), identifying probabilities with long-run frequencies of events. The obvious complaint that short-run frequencies clearly do not match probabilities (e.g., if we toss a coin only once, we would hardly conclude that its probability of heads is either one or zero) does not actually get anywhere, since no claim is made identifying short-run frequencies with probabilities. A different complaint does bite, however, namely that the distinction between short-run and long-run is vague, leaving the commitments of this frequentist interpretation unclear. Richard von Mises in the early 20th century fixed this problem by formalizing the frequency interpretation (von Mises, 1919), identifying probabilities with frequency limits in infinite sequences satisfying certain assumptions about randomness. Some version of this frequency interpretation is commonly endorsed by statisticians.

A more satisfactory theoretical account of physical probability arises from Karl Popper's observation (1959) that the frequency interpretation, precise though it was, fails to accommodate our intuition that probabilities of singular events exist and are meaningful. If, in fact, we do toss a coin once and once only, and if this toss should *not* participate in some infinitude (or even large number) of appropriately similar tosses, it would not for that reason fail to have some probability of landing heads. Popper identified physical probabilities with the **propensities** (dispositions) of physical systems ("chance setups") to produce particular outcomes, whether or not those dispositions were manifested repeatedly. An alternative that amounts to much the same thing is to identify probabilities with counterfactual frequencies generated by hypothetically infinite repetitions of an experiment (van Fraassen, 1989).

Whether physical probability is relativized to infinite random sequences, infinite counterfactual sequences or chance setups, these accounts all have in common that the assertion of a probability is relativized to *some* definite physical process or the outcomes it generates.

The traditional alternative to the concept of physical probability is to think of probabilities as reporting our subjective degrees of belief. This view was expressed by Thomas Bayes (1958) (Figure 1.3) and Pierre Simon de Laplace (1951) two hundred years ago. This is a more general account of probability in that we have subjective belief in a huge variety of propositions, many of which are not at all clearly tied to a physical process capable even in principle of generating an infinite sequence of outcomes. For example, most of us have a pretty strong belief in the Copernican hypothesis that the earth orbits the sun, but this is based on evidence not obviously the same as the outcome of a sampling process. We are not in any position to generate solar systems repeatedly and observe the frequency with which their planets revolve around the sun, for example. Bayesians nevertheless are prepared to talk

about the probability of the truth of the Copernican thesis and can give an account of the relation between that probability and the evidence for and against it. Since these probabilities are typically subjective, not clearly tied to physical models, most frequentists (hence, most statisticians) deny their meaningfulness. It is not insignificant that this leaves their (usual) belief in Copernicanism unexplained.

The first thing to make clear about this dispute between physicalists and Bayesians is that Bayesianism can be viewed as *generalizing* physicalist accounts of probability. That is, it is perfectly compatible with the Bayesian view of probability as measuring degrees of subjective belief to adopt what David Lewis (1980) dubbed the **Principal Principle** whenever you learn that the physical probability of an outcome is r, set your subjective probability for that outcome to r. This is really just common sense: you may think that the probability of a friend shaving his head is 0.01, but if you learn that he will do so if and only if a fair coin yet to be flipped lands heads, you'll revise your opinion accordingly.

So, the Bayesian and physical interpretations of probability are compatible, with the Bayesian interpretation *extending* the application of probability beyond what is directly justifiable in physical terms. That is the view we adopt here. But what justifies this extension?

FIGURE 1.3: Reverend Thomas Bayes (1702–1761).

1.5 Bayesian philosophy

1.5.1 Bayes' theorem

The origin of Bayesian philosophy lies in an interpretation of **Bayes' Theorem:**

Theorem 1.4 Bayes' Theorem

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

This is a non-controversial (and simple) theorem of the probability calculus. Under its usual Bayesian interpretation, it asserts that the probability of a hypothesis h conditioned upon some evidence e is equal to its **likelihood** $P(e|h)$ times its probability prior to any evidence $P(h)$, **normalized** by dividing by $P(e)$ (so that the conditional probabilities of all hypotheses sum to 1). Proof is trivial.

The further claim that this is a right and proper way of adjusting our beliefs in our hypotheses given new evidence is called **conditionalization**, and it is controversial.

Definition 1.4 Conditionalization *After applying Bayes' theorem to obtain $P(h|e)$ adopt that as your posterior degree of belief in h — or, $Bel(h) = P(h|e)$.*

Conditionalization, in other words, advocates belief updating via probabilities conditional upon the available evidence. It identifies **posterior probability** (the probability function after incorporating the evidence, which we are writing $Bel(\cdot)$) with **conditional probability** (the prior probability function conditional upon the evidence, which is $P(\cdot|e)$). Put thus, conditionalization may also seem non-controvertible. But there are certainly situations where conditionalization very clearly does not work. The two most basic such situations simply violate what are frequently explicitly stated as assumptions of conditionalization: (1) There must exist joint priors over the hypothesis and evidence spaces. Without a joint prior, Bayes' theorem cannot be used, so conditionalization is a non-starter. (2) The evidence conditioned upon, e, is all and only the evidence learned. This is called the **total evidence condition**. It is a significant restriction, since in many settings it cannot be guaranteed.

The first assumption is also significant. Many take it as the single biggest objection to Bayesianism to raise the question "Where do the numbers come from?" For example, the famous anti-Bayesian Clark Glymour (1980) doesn't complain about Bayesian reasoning involving gambling devices, when the outcomes are engineered to start out equiprobable, but doubts that numbers can be found for more interesting cases. To this kind of objection Bayesians react in a variety of ways. In fact, the different varieties of response pretty much identify the different schools of Bayesianism. Objectivists, such as Rudolf Carnap (1962) and Ed Jaynes (1968), attempt to define prior probabilities based upon the structure of language. Extreme subjectivists, such as de Finetti (1964), assert that it makes no difference what source your priors have:

given that de Finetti's representation theorem shows that non-extreme priors converge in the limit (under reasonable constraints), it just doesn't matter what priors you adopt.

The practical application of Bayesian reasoning does not appear to depend upon settling this kind of philosophical problem. A great deal of useful application can be done simply by refusing to adopt a dogmatic position and accepting commonsense prior probabilities. For example, if there are ten possible suspects in a murder mystery, a fair starting point for any one of them is a 1 in 10 chance of guilt; or, again, if burglaries occur in your neighborhood of 10,000 homes about once a day, then the probability of your having been burglarized within the last 24 hours might reasonably be given a prior probability of 1/10000.

Colin Howson points out that conditionalization is a valid rule of inference if and only if $Bel(e|h) = P(e|h)$, that is, if and only if your prior and posterior probability functions share the relevant conditional probabilities (cf. Howson, 2001). This is certainly a pertinent observation, since encountering some possible evidence may well inform us more about defects in our own conditional probability structure than about the hypothesis at issue. Since Bayes' theorem has $P(h|e)$ being proportional to $P(e|h)$, if the evidence leads us to revise $P(e|h)$, we will be in no position to conditionalize.

How to generate prior probabilities or new conditional probability structure is not dictated by Bayesian principles. Bayesian principles advise how to update probabilities once such a conditional probability structure has been adopted, given appropriate priors. Expecting Bayesian principles to answer all questions about reasoning is expecting too much. Nevertheless, we shall show that Bayesian principles implemented in computer programs can deliver a great deal more than the nay-sayers have ever delivered.

Definition 1.5 Jeffrey conditionalization *Suppose your observational evidence does not correspond specifically to proposition e, but can be represented as a posterior shift in belief about e. In other words, posterior belief in e is not full but partial, having shifted from* $P(e)$ *to* $Bel(e)$*. Then, instead of Bayesian conditionalization, apply Jeffrey's update rule for* **probability kinematics***:* $Bel(h) = P(h|e)Bel(e) + P(h|\neg e)Bel(\neg e)$ *(Jeffrey, 1983).*

Jeffrey's own example is one where your hypothesis is about the color of a cloth, the evidence proposition e describes the precise quality of your visual experience under good light, but you are afforded a view of the cloth only under candlelight, in such a way that you cannot exactly articulate what you have observed. Nevertheless, you have learned *something*, and this is reflected in a shift in belief about the quality of your visual experience. Jeffrey conditionalization is very intuitive, but again is not strictly valid. As a practical matter, the need for such partial updating is common in Bayesian modeling.

1.5.2 Betting and odds

Odds are the ratio between the cost of a bet in favor of a proposition and the reward should the bet be won. Thus, assuming a stake of \$1 (and otherwise simply rescaling the terms of the bet), a bet at 1:19 odds costs \$1 and returns \$20 should the proposition come true (with the reward being \$20 minus the cost of the bet).[5] The odds may be set at any ratio and may, or may not, have something to do with one's probabilities. Bookies typically set odds for and against events at a slight discrepancy with their best estimate of the probabilities, for their profit lies in the difference between the odds for and against.

While odds and probabilities may deviate, probabilities and **fair odds** $O(\cdot)$ are strictly interchangeable concepts. The fair odds in favor of h are defined simply as the ratio of the probability that h is true to the probability that it is not:

Definition 1.6 Fair odds

$$O(h) = \frac{P(h)}{1 - P(h)}$$

Given this, it is an elementary matter of algebraic manipulation to find $P(h)$ in terms of odds:

$$P(h) = \frac{O(h)}{1 + O(h)} \tag{1.4}$$

Thus, if a coin is fair, the probability of heads is 1/2, so the odds in favor of heads are 1:1 (usually described as "50:50"). Or, if the odds of getting "snake eyes" (two 1's) on the roll of two dice are 1:35, then the probability of this is:

$$\frac{1/35}{1 + 1/35} = \frac{1/35}{36/35} = 1/36$$

as will always be the case with fair dice. Or, finally, suppose that the probability an agent ascribes to the Copernican hypothesis (CH) is zero; then the odds that agent is giving to Copernicus having been wrong ($\neg CH$) are *infinite*:

$$O(\neg CH) = \frac{1}{0} = \infty$$

At these odds, incidentally, it is trivial that the agent can never reach a degree of belief in CH above zero on any finite amount of evidence, if relying upon conditionalization for updating belief.

With the concept of fair odds in hand, we can reformulate Bayes' theorem in terms of (fair) odds, which is often useful:

Theorem 1.5 Odds-Likelihood Bayes' Theorem

$$O(h|e) = \frac{P(e|h)}{P(e|\neg h)} O(h)$$

[5] It is common in sports betting to invert the odds, quoting the odds *against* a team winning, for example. This makes no difference; the ratio is simply reversed.

This is readily proven to be equivalent to Theorem 1.4. In English it asserts that the odds on h conditional upon the evidence e are equal to the prior odds on h times the **likelihood ratio** $P(e|h) : P(e|\neg h)$. Clearly, the fair odds in favor of h will rise if and only if the likelihood ratio is greater than one.

1.5.3 Expected utility

Generally, agents are able to assign utility (or, value) to the situations in which they find themselves. We know what we like, we know what we dislike, and we also know when we are experiencing neither of these. Given a general ability to order situations, and bets with definite probabilities of yielding particular situations, Frank Ramsey (1931) demonstrated that we can identify particular utilities with each possible situation, yielding a **utility function**.

If we have a utility function $U(O_i|A)$ over every possible outcome of a particular action A we are contemplating, and if we have a probability for each such outcome $P(O_i|A)$, then we can compute the probability-weighted average utility for that action — otherwise known as the **expected utility** of the action:

Definition 1.7 Expected utility

$$EU(A) = \sum_i U(O_i|A) \times P(O_i|A)$$

It is commonly taken as axiomatic by Bayesians that agents ought to *maximize their expected utility*. That is, when contemplating a number of alternative actions, agents ought to decide to take that action which has the maximum expected utility. If you are contemplating eating strawberry ice cream or else eating chocolate ice cream, presumably you will choose that flavor which you prefer, other things being equal. Indeed, if you chose the flavor you liked *less*, we should be inclined to think that other things are *not* equal — for example, you are under some kind of external compulsion — or perhaps that you are not being honest about your preferences. Utilities have behavioral consequences *essentially*: any agent who consistently ignores the putative utility of an action or situation arguably does not have that utility.

Regardless of such foundational issues, we now have the conceptual tools necessary to understand what is fair about fair betting. **Fair bets** are fair because their expected utility is zero. Suppose we are contemplating taking the fair bet B on proposition h for which we assign probability $P(h)$. Then the expected utility of the bet is:

$$EU(B) = U(h|B)P(h|B) + U(\neg h|B)P(\neg h|B)$$

Typically, betting on a proposition has no effect on the probability that it is true (although this is not necessarily the case!), so $P(h|B) = P(h)$. Hence,

$$EU(B) = U(h|B)P(h) + U(\neg h|B)(1 - P(h))$$

Assuming a stake of 1 unit for simplicity, then by definition $U(h|B) = 1 - P(h)$ (i.e., this is the utility of h being true given the bet for h) while $U(\neg h|B) = -P(h)$, so,

$$EU(B) = (1 - P(h))P(h) - P(h)(1 - P(h)) = 0$$

Given that the bet has zero expected utility, the agent should be no more inclined to take the bet in favor of h than to take the opposite bet against h.

1.5.4 Dutch books

The original Dutch book argument of Ramsey (1931) (see also de Finetti, 1964) claims to show that subjective degrees of belief, if they are to be rational, *must* obey the probability calculus. It has the form of a *reductio ad absurdum* argument:

1. A rational agent should be willing to take either side of any combination of fair bets.
2. A rational agent should never be willing to take a combination of bets which guarantees a loss.
3. Suppose a rational agent's degrees of belief violate one or more of the axioms of probability.
4. Then it is provable that some combination of fair bets will lead to a guaranteed loss.
5. Therefore, the agent is both willing and not willing to take this combination of bets.

Now, the inferences to (4) and (5) in this argument are not in dispute (see §1.11 for a simple demonstration of (4) for one case). A *reductio* argument needs to be resolved by finding a prior assumption to blame, and concluding that it is false. Ramsey, and most Bayesians to date, supposed that the most plausible way of relieving the contradiction of (5) is by refusing to suppose that a rational agent's degrees of belief may violate the axioms of probability. This result can then be generalized beyond settings of explicit betting by taking "bets with nature" as a metaphor for decision-making generally. For example, walking across the street is in some sense a bet about our chances of reaching the other side.

Some anti-Bayesians have preferred to deny (1), insisting for example that it would be uneconomic to invest in bets with zero expected value (e.g., Chihara and Kennedy, 1979). But the ascription of the radical incoherence in (5) simply to the willingness of, say, bored aristocrats to place bets that will net them nothing clearly will not do: the effect of incoherence is entirely out of proportion with the proposed cause of effeteness.

Alan Hájek (2008) has pointed out a more plausible objection to (2). In the scenarios presented in Dutch books there is always some combination of bets which guarantees a net loss whatever the outcomes on the individual bets. But equally there is always some combination of bets which guarantees a net gain — a "Good Book." So, one agent's half-empty glass is another's half-full glass! Rather than dismiss the Dutch-bookable agent as irrational, we might commend it for being open to a guaranteed win! So, Hájek's point seems to be that there is a fundamental symmetry in Dutch book arguments which leaves open the question whether violating probability axioms is rational or not. Certainly, when metaphorically extending betting to a "struggle" with Nature, it becomes rather implausible that She is really out to Dutch book us!

Hájek's own solution to the problem posed by his argument is to point out that whenever an agent violates the probability axioms there will be some variation of its system of beliefs which is guaranteed to win money whenever the original system is guaranteed to win, and which is also capable of winning in some situations when the original system is not. So the variant system of belief in some sense dominates the original: it is everywhere at least as good as the original and in some places better. In order to guarantee that your system of beliefs cannot be dominated, you must be probabilistically coherent (see §1.11). This, we believe, successfully rehabilitates the Dutch book in a new form.

Rather than rehabilitate, a more obviously Bayesian response is to consider the probability of a bookie hanging around who has the smarts to pump our agent of its money and, again, of a simpleton hanging around who will sign up the agent for guaranteed winnings. In other words, for rational choice surely what matters is the relative expected utility of the choice. Suppose, for example, that we are offered a set of bets which has a guaranteed loss of $10. Should we take it? The Dutch book assumes that accepting the bet is irrational. But, if the one and only alternative available is another bet with an expected loss of $1,000, then it no longer seems so irrational. An implicit assumption of the Dutch book has always been that betting is voluntary and when all offered bets are turned down the expected utility is zero. The further implicit assumption pointed out by Hájek's argument is that there is always a shifty bookie hanging around ready to take advantage of us. No doubt that is not always the case, and instead there is only some probability of it. Yet referring the whole matter of justifying the use of Bayesian probability to expected utility smacks of circularity, since expectation is understood in terms of Bayesian probability.

Aside from invoking the rehabilitated Dutch book, there is a more pragmatic approach to justifying Bayesianism, by looking at its importance for dealing with cases of practical problem solving. We take Bayesian principles to be normative, and especially to be a proper guide, under some range of circumstances, to evaluating hypotheses in the light of evidence. The form of justification that we think is ultimately most compelling is the "method of reflective equilibrium," generally attributed to Goodman (1973) and Rawls (1971), but adumbrated by Aristotle in his *Nicomachian Ethics*. In a nutshell, it asserts that the normative principles to accept are those which best accommodate our basic, unshakable intuitions about what is good and bad (e.g., paradigmatic judgments of correct inference in simple domains, such as gambling) and which best integrate with relevant theory and practice. We now present some cases which Bayesian principle handles readily, and better than any alternative normative theory.

1.5.5 Bayesian reasoning examples

1.5.5.1 Breast cancer

Suppose the women attending a particular clinic show a long-term chance of 1 in 100 of having breast cancer. Suppose also that the initial screening test used at the clinic has a false positive rate of 0.2 (that is, 20% of women without cancer will test

positive for cancer) and that it has a false negative rate of 0.1 (that is, 10% of women with cancer will test negative). The laws of probability dictate from this last fact that the probability of a positive test given cancer is 90%. Now suppose that you are such a woman who has just tested positive. What is the probability that you have cancer?

This problem is one of a class of probability problems which has become notorious in the cognitive psychology literature (cf. Tversky and Kahneman, 1974). It seems that very few people confronted with such problems bother to pull out pen and paper and compute the right answer via Bayes' theorem; even fewer can get the right answer without pen and paper. It appears that for many the probability of a positive test (which is observed) given cancer (i.e., 90%) dominates things, so they figure that they have quite a high chance of having cancer. But substituting into Theorem 1.4 gives us:

$$P(Cancer|Pos) = \frac{P(Pos|Cancer)P(Cancer)}{P(Pos)}$$

Note that the probability of *Pos* given *Cancer* — which is the likelihood 0.9 — is only *one* term on the right hand side; the other crucial term is the prior probability of cancer. Cognitive psychologists studying such reasoning have dubbed the dominance of likelihoods in such scenarios "base-rate neglect," since the base rate (prior probability) is being suppressed (Kahneman and Tversky, 1973). Filling in the formula and computing the conditional probability of *Cancer* given *Pos* gives us quite a different story:

$$\begin{aligned} P(Cancer|Pos) &= \frac{P(Pos|Cancer)P(Cancer)}{P(Pos)} \\ &= \frac{P(Pos|Cancer)P(Cancer)}{P(Pos|Cancer)P(Cancer) + P(Pos|\neg Cancer)P(\neg Cancer)} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.2 \times 0.99} \\ &= \frac{0.009}{0.009 + 0.198} \\ &\approx 0.043 \end{aligned}$$

Now the discrepancy between 4% and 80 or 90% is no small matter, particularly if the consequence of an error involves either unnecessary surgery or (in the reverse case) leaving a cancer untreated. But decisions similar to these are constantly being made based upon "intuitive feel" — i.e., without the benefit of paper and pen, let alone Bayesian networks (which are simpler to use than paper and pen!).

1.5.5.2 People v. Collins

The legal system is replete with misapplications of probability and with incorrect claims of the irrelevance of probabilistic reasoning as well.

In 1964 an interracial couple was convicted of robbery in Los Angeles, largely on the grounds that they matched a highly improbable profile, a profile which fit witness reports (Sullivan, Sullivan). In particular, the two robbers were reported to be

- A man with a mustache

- Who was black and had a beard
- And a woman with a ponytail
- Who was blonde
- The couple was interracial
- And were driving a yellow car

The prosecution suggested that these characteristics had the following probabilities of being observed at random in the LA area:

1. A man with a mustache 1/4
2. Who was black and had a beard 1/10
3. And a woman with a ponytail 1/10
4. Who was blonde 1/3
5. The couple was interracial 1/1000
6. And were driving a yellow car 1/10

The prosecution called an instructor of mathematics from a state university who apparently testified that the "product rule" applies to this case: where mutually independent events are being considered jointly, the joint probability is the product of the individual probabilities.[6] This last claim is, in fact, correct (see Problem 2 below); what is false is the idea that the product rule is relevant to this case. If we label the individual items of evidence e_i $(i = 1, \ldots, 6)$, the joint evidence e, and the hypothesis that the couple was guilty h, then what is claimed is

$$P(e|\neg h) = \prod_i P(e_i|\neg h) = 1/12000000$$

The prosecution, having made this inference, went on to assert that the probability the couple were innocent was no more than 1/12000000. The jury convicted.

As we have already suggested, the product rule does *not* apply in this case. Why not? Well, because the individual pieces of evidence are obviously *not* independent. If, for example, we know of the occupants of a car that one is black and the other has blonde hair, what then is the probability that the occupants are an interracial couple? Clearly not 1/1000! If we know of a man that he has a mustache, is the probability of having a beard unchanged? These claims are preposterous, and it is simply shameful that a judge, prosecutor and defence attorney could not recognize how preposterous they are — let alone the mathematics "expert" who testified to them. Since e_2 implies e_1, while e_2, e_3, e_4 jointly imply e_5 (to a fair approximation), a far better estimate for $P(e|\neg h)$ is $P(e_2|\neg h)P(e_3|\neg h)P(e_4|\neg h)P(e_6|\neg h) = 1/3000$.

To be sure, if we accepted that the probability of innocence were a mere 1/3000 we might well accept the verdict. But there is a more fundamental error in the prosecution reasoning than neglecting the conditional dependencies in the evidence. If,

[6] Coincidentally, this is just the kind of independence required for certainty factors to apply.

unlike the judge, prosecution and jury, we take a peek at Bayes' theorem, we discover that the probability of guilt $P(h|e)$ is *not* equal to $1 - P(e|\neg h)$; instead

$$P(h|e) = \frac{P(e|h)P(h)}{P(e|h)P(h) + P(e|\neg h)P(\neg h)}$$

Now if the couple in question *were* guilty, what are the chances the evidence accumulated would have been observed? That's a rather hard question to answer, but feeling generous towards the prosecution, let us simplify and say 1. That is, let us accept that $P(e|h) = 1$. Plugging in our assumptions we have thus far:

$$P(h|e) = \frac{P(h)}{P(h) + P(\neg h)/3000}$$

We are missing the crucial prior probability of a random couple being guilty of the robbery. Note that we cannot here use the prior probability of, for example, an interracial couple being guilty, since the fact that they are interracial is a piece of the evidence. The most plausible approach to generating a prior of the needed type is to count the number of couples in the LA area and give them an equal prior probability. In other words, if N is the number of possible couples in the LA area, $P(h) = 1/N$. So, what is N? The population at the time was about 6.5 million people (Demographia, Demographia). If we conservatively take half of them as being eligible to be counted (e.g., being adult humans), this gives us 1,625,000 eligible males and as many females. If we simplify by supposing that they are all in heterosexual partnerships, that will introduce a slight bias in favor of innocence; if we also simplify by ignoring the possibility of people traveling in cars with friends, this will introduce a larger bias in favor of guilt. The two together give us 1,625,000 available couples, suggesting a prior probability of guilt of 1/1625000. Plugging this in we get:

$$P(h|e) = \frac{1/1625000}{1/1625000 + (1 - 1/1625000)/3000} \approx 0.002$$

In other words, even ignoring the huge number of trips with friends rather than partners, we obtain a 99.8% chance of innocence and so a very large probability of a nasty error in judgment. The good news is that the conviction (of the man only!) was subsequently overturned, partly on the basis that the independence assumptions are false. The bad news is that the appellate court finding also suggested that probabilistic reasoning is just irrelevant to the task of establishing guilt, which is a nonsense. One right conclusion about this case is that, assuming the likelihood has been *properly* worked out, a sensible prior probability must also be taken into account. In some cases judges have specifically ruled out all consideration of prior probabilities, while allowing testimony about likelihoods! Probabilistic reasoning which simply ignores half of Bayes' theorem is dangerous indeed!

Note that we do not claim that 99.8% is the best probability of innocence that can be arrived at for the case of People v. Collins. What we *do* claim is that, for the particular facts represented as having a particular probabilistic interpretation, this is far closer to a reasonable probability than that offered by the prosecution, namely

1/12000000. We also claim that the forms of reasoning we have here illustrated are *crucial* for interpreting evidence in general: namely, whether the offered items of evidence are conditionally independent and what the prior probability of guilt may be.

1.6 The goal of Bayesian AI

The most commonly stated goal for artificial intelligence is that of producing an artifact which performs difficult intellectual tasks at or beyond a human level of performance. Of course, machine chess programs have satisfied this criterion for some time now. Although some AI researchers have claimed that therefore an AI has been produced — that denying this is an unfair shifting of the goal line — it is absurd to think that we ought to be satisfied with programs which are strictly special-purpose and which achieve their performance using techniques that deliver nothing when applied to most areas of human intellectual endeavor.

Turing's test for intelligence appears to be closer to satisfactory: fooling ordinary humans with verbal behavior not restricted to any domain would surely demonstrate some important *general* reasoning ability. Many have pointed out that the conditions for Turing's test, strictly verbal behavior without any afferent or efferent nervous activity, yield at best some kind of disembodied, ungrounded intelligence. John Searle's Chinese Room argument (Searle, 1980) for example, can be interpreted as making such a case; for this kind of interpretation of Searle see Harnad (1989) and Korb (1991). A more convincing criterion for human-like intelligence is to require of an artificial intelligence that it be capable of powering a robot-in-the-world in such a way that the robot's performance cannot be distinguished from human performance in terms of behavior (disregarding, for example, whether the skin can be so distinguished). The program that can achieve this would surely satisfy any sensible AI researcher, or critic, that an AI had been achieved.

We are not, however, actually motivated by the idea of behaviorally cloning humans. If all we wish to do is reproduce humans, we would be better advised to employ the tried and true methods we have always had available. Our motive is to understand *how* such performance can be achieved. We are interested in knowing how humans perform the many interesting and difficult cognitive tasks encompassed by AI — such as, natural language understanding and generation, planning, learning, decision making — but we are also interested in knowing how they might be performed otherwise, and in knowing how they might be performed optimally. By building artifacts which model our best understanding of how humans do these things (which can be called **descriptive artificial intelligence**) and also building artifacts which model our best understanding of what is optimal in these activities (**normative artificial intelligence**), we can further our understanding of the nature of intelligence and also produce some very useful tools for science, government and industry.

As we have indicated through example, medical, legal, scientific, political and

most other varieties of human reasoning either consider the relevant probabilistic factors and accommodate them or run the risk of introducing egregious and damaging errors. The goal of a Bayesian artificial intelligence is to produce a thinking agent which does as well or better than humans in such tasks, which can adapt to stochastic and changing environments, recognize its own limited knowledge and cope sensibly with these varied sources of uncertainty.

1.7 Achieving Bayesian AI

Given that we have this goal, how can we achieve it? The first step is to develop algorithms for doing Bayesian conditionalization properly and, insofar as possible, efficiently. This step has already been achieved, and the relevant algorithms are described in Chapters 2 and 3. The next step is to incorporate methods for computing expected utilities and develop methods for maximizing utility in decision making. We describe algorithms for this in Chapter 4. We would like to test these ideas in application: we describe some Bayesian network applications in Chapter 5.

These methods for probability computation are fairly well developed and their improvement remains an active area of research in AI today. The biggest obstacles to Bayesian AI having a broad and deep impact outside of the research community are the difficulties in developing applications, difficulties with eliciting knowledge from experts, and integrating and validating the results. One issue is that there is no clear methodology for developing, testing and deploying Bayesian network technology in industry and government — there is no recognized discipline of "software engineering" for Bayesian networks. We make a preliminary effort at describing one — Knowledge Engineering with Bayesian Networks (KEBN) in Part III, including its illustration in case studies of Bayesian network development in Chapter 11.

Another important response to the difficulty of building Bayesian networks by hand is the development of methods for their automated learning — the machine learning of Bayesian networks (aka "data mining"). In Part II we introduce and develop the main methods for learning Bayesian networks with reference to the theory of causality underlying them. These techniques logically come before the knowledge engineering methodology, since that draws upon and integrates machine learning with expert elicitation.

1.8 Are Bayesian networks Bayesian?

Many AI researchers like to point out that Bayesian networks are not inherently Bayesian at all; some have even claimed that the label is a misnomer. At the 2002 Australasian Data Mining Workshop, for example, Geoff Webb made the former

claim. Under questioning it turned out he had two points in mind: (1) Bayesian networks are frequently "data mined" (i.e., learned by some computer program) via non-Bayesian methods. (2) Bayesian networks at bottom represent probabilities; but probabilities can be interpreted in any number of ways, including as some form of frequency; hence, the networks are not intrinsically either Bayesian or non-Bayesian, they simply represent values needing further interpretation.

These two points are entirely correct. We shall ourselves present non-Bayesian methods for automating the learning of Bayesian networks from statistical data. We shall also present Bayesian methods for the same, together with some evidence of their superiority. The interpretation of the probabilities represented by Bayesian networks is open so long as the philosophy of probability is considered an open question. Indeed, much of the work presented here ultimately depends upon the probabilities being understood as *physical probabilities*, and in particular as propensities or probabilities determined by propensities. Nevertheless, we happily invoke the Principal Principle: where we are convinced that the probabilities at issue reflect the true propensities in a physical system we are certainly going to use them in assessing our own degrees of belief.

The advantages of the Bayesian network representations are largely in simplifying conditionalization, planning decisions under uncertainty and explaining the outcome of stochastic processes. These purposes all come within the purview of a clearly Bayesian interpretation of what the probabilities mean, and so, we claim, the Bayesian network technology which we here introduce is aptly named: it provides the technical foundation for a truly Bayesian artificial intelligence.

1.9 Summary

How best to reason about uncertain situations has always been of concern. From the 17th century we have had available the basic formalism of probability calculus, which is far and away the most promising formalism for coping with uncertainty. Probability theory has been used widely, but not deeply, since then. That is, the elementary ideas have been applied to a great variety of problems — e.g., actuarial calculations for life insurance, coping with noise in measurement, business decision making, testing scientific theories, gambling — but the problems have typically been of highly constrained size, because of the computational infeasibility of conditionalization when dealing with large problems. Even in dealing with simplified problems, humans have had difficulty handling the probability computations. The development of Bayesian network technology automates the process and so promises to free us from such difficulties. At the same time, improvements in computer capacity, together with the ability of Bayesian networks to take computational advantage of any available independencies between variables, promise to both widen and deepen the domain of probabilistic reasoning.

1.10 Bibliographic notes

An excellent source of information about different attempts to formalize reasoning about uncertainty — including certainty factors, non-monotonic logics, Dempster-Shafer calculus, as well as probability — is the anthology *Readings in Uncertain Reasoning* edited by Shafer and Pearl (1990). Three polemics against non-Bayesian approaches to uncertainty are those by Drew McDermott (1987), Peter Cheeseman (1988) and Kevin Korb (1995). For understanding Bayesian philosophy, Ramsey's original paper "Truth and Probability" is beautifully written, original and compelling (1931); for a more comprehensive and recent presentation of Bayesianism see Howson and Urbach's *Scientific Reasoning* (2007). For Bayesian decision analysis see Richard Jeffrey's *The Logic of Decision* (1983). DeGroot and Schervish (2002) provide an accessible introduction to both the probability calculus and statistics.

Karl Popper's original presentation of the propensity interpretation of probability is (Popper, 1959). This view is related to the elaboration of a probabilistic account of causality in recent decades. Wesley Salmon (1984) provides an overview of probabilistic causality.

Naive Bayes models, despite their simplicity, have done surprisingly well as predictive classifiers for data mining problems; see Chapter 7.

1.11 Technical notes

A Dutch book

Here is a simple Dutch book. Suppose someone assigns $P(A) = -0.1$, violating probability Axiom 2. Then $O(A) = -0.1/(1-(-0.1)) = -0.1/1.1$. The reward for a bet on A with a \$1 stake is $\$(1-P(U)) = \$1.1$ if A comes true and $\$-P(U) = \$0.1$ if A is false. That's everywhere positive and so is a "Good Book." The Dutch book simply requires this agent to take the fair bet *against* A, which has the payoffs $-\$1.1$ if A is true and $-\$0.1$ otherwise.

The rehabilitated Dutch book

Following Hájek, we can show that incoherence (violating the probability axioms) leads to being "dominated" by someone who is coherent — that is, the coherent bettor can take advantage of offered bets that the incoherent bettor cannot and otherwise will do as well.

Suppose Ms. Incoherent assigns $P_I(U) < 1$ (where U is the universal event that *must* occur), for example. Then Ms. Incoherent will take any bet for U at odds of $P_I(U)/(1-P_I(U))$ or greater. But Ms. Coherent has assigned $P_C(U) = 1$, of course, and so can take any bet for U at any odds offered greater than zero. So for the odds within the range $[0, \frac{P_I(U)}{1-P_I(U)}]$ Ms. Coherent is guaranteed a profit whereas Ms. Inco-

herent is sitting on her hands.

NP hardness
A problem is Non-deterministic Polynomial-time (NP) if it is solvable in polynomial time on a non-deterministic Turing machine. A problem is Non-deterministic Polynomial time hard (NP hard) if every problem that is NP can be translated into this NP hard problem in polynomial time. If there is a polynomial time solution to any NP hard problem, then because of polynomial time translatability for all other NP problems, there must be a polynomial time solution to all NP problems. No one knows of a polynomial time solution to any NP hard problem; the best known solutions are exponentially explosive. Thus, "NP hard" problems are generally regarded as computationally intractable. (The classic introduction to computational complexity is Garey and Johnson (1979).)

1.12 Problems

Probability Theory

Problem 1

Prove that the conditional probability function $P(\cdot|e)$, if well defined, is a probability function (i.e., satisfies the three axioms of Kolmogorov).

Problem 2

Given that two pieces of evidence e_1 and e_2 are conditionally independent given the hypothesis — i.e., $P(e_1|e_2,h) = P(e_1|h)$ — prove the "product rule": $P(e_1,e_2|h) = P(e_1|h) \times P(e_2|h)$.

Problem 3

Prove the theorems of §1.3.1, namely the Total Probability theorem and the Chain Rule.

Problem 4

There are five containers of milk on a shelf; unbeknownst to you, two of them have passed their use-by date. You grab two at random. What's the probability that neither have passed their use-by date? Suppose someone else has got in just ahead of you, taking one container, after examining the dates. What's the probability that the two you take at random after that are ahead of their use-by dates?

Problem 5

The probability of a child being a boy (or a girl) is 0.5 (let us suppose). Consider all the families with exactly two children. What is the probability that such a family has two girls given that it has at least one girl?

Problem 6

The frequency of male births at the Royal Women's Hospital is about 51 in 100. On a particular day, the last eight births have been female. The probability that the next birth will be male is:

1. About 51%
2. Clearly greater than 51%
3. Clearly less than 51%
4. Almost certain
5. Nearly zero

Bayes' Theorem

Problem 7

After winning a race, an Olympic runner is tested for the presence of steroids. The test comes up positive, and the athlete is accused of doping. Suppose it is known that 5% of all victorious Olympic runners do use performance-enhancing drugs. For this particular test, the probability of a positive finding given that drugs are used is 95%. The probability of a false positive is 2%. What is the (posterior) probability that the athlete did in fact use steroids, given the positive outcome of the test?

Problem 8

You consider the probability that a coin is double-headed to be 0.01 (call this option h'); if it isn't double-headed, then it's a fair coin (call this option h). For whatever reason, you can only test the coin by flipping it and examining the coin (i.e., you can't simply examine both sides of the coin). In the worst case, how many tosses do you need before having a posterior probability for either h or h' that is greater than 0.99, i.e., what's the maximum number of tosses until that happens?

Problem 9

(Adapted from Fischoff and Bar-Hillel (1984).) Two cab companies, the Blue and the Green, operate in a given city. Eighty-five percent of the cabs in the city are Blue; the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. A witness identified the cab as a Green cab. The court tested the witness' ability to distinguish between Blue and Green cabs under night-time visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time.

What are the chances that the errant cab was indeed Green, as the witness claimed?

Odds and Expected Value

Problem 10

Construct a Dutch book against someone who violates the Axiom of Additivity. That is, suppose a Mr. Fuzzy declares about the weather tomorrow that $P(Sunny) = 0.5, P(Inclement) = 0.5,$ and $P(Sunny\ or\ inclement) = 0.5$. Mr. Fuzzy and you agree about what will count as sunny and as inclement weather and you both agree that they are incompatible states. How can you construct a Dutch book against Fuzzy, using only fair bets?

Problem 11

A bookie offers you a ticket for \$5.00 which pays \$6.00 if Manchester United beats Arsenal and nothing otherwise. What are the odds being offered? To what probability of Manchester United winning does that correspond?

Problem 12

You are offered a Keno ticket in a casino which will pay you \$1 million if you win! It only costs you \$1 to buy the ticket. You choose 4 numbers out of a 9x9 grid of distinct numbers. You win if all of your 4 numbers come up in a random draw of four from the 81 numbers. What is the expected dollar value of this gamble?

Applications

Problem 13

(Note: this is the case of Sally Clark, convicted in the UK in 1999, found innocent on appeal in 2003, and tragically died in 2007 of alcohol poisoning. See Innocent, 2002.) A mother was arrested after her second baby died a few months old, apparently of sudden infant death syndrome (SIDS), exactly as her first child had died a year earlier. According to prosecution testimony, about 2 in 17200 babies die of SIDS. So, according to their argument, there is only a probability of $(2/17200)^2 \approx 1/72000000$ that two such deaths would happen in the same family by chance alone. In other words, according to the prosecution, the woman was guilty beyond a reasonable doubt. The jury returned a guilty verdict, even though there was no significant evidence of guilt presented beyond this argument. Which of the following is the truth of the matter? Why?

1. Given the facts presented, the probability that the woman is guilty is greater than 99%, so the jury decided correctly.
2. The argument presented by the prosecution is irrelevant to the mother's guilt or innocence.
3. The prosecution argument is relevant but inconclusive.

4. The prosecution argument only establishes a probability of guilt of about 16%.
5. Given the facts presented, guilt and innocence are equally likely.

Problem 14

A DNA match between the defendant and a crime scene blood sample has a probability of 1/100000 if the defendant is innocent. There is no other significant evidence.

1. What is the probability of guilt?
2. Suppose we agree that the prior probability of guilt under the (unspecified) circumstances is 10%. What then is the probability of guilt?
3. The suspect has been picked up through a universal screening program applied to all Australians seeking a Medicare card. So far, 10 million people have been screened. What then is the probability of guilt?