EXERCISES 12

QUANTIFYING UNCERTAINTY

12.1 Acting under Uncertainty

Exercise 12.XXXX

The chapter proposes probability theory as the basis for reasoning under uncertainty and provides one argument based on work by de Finetti. Compare and contrast de Finetti's argument with other arguments by Ramsey, Cox, Savage, Jeffrey, and Jaynes (see the historical and bibliographical notes for references). Also consider any rebuttals you can find.

All offer arguments premised on an individual making bets proportional to their degree of belief.

Ramsey (1931) introduced the terminology of "degrees of belief" by imagining a rational individual making decisions. Such a decision must be based on the utility of an outcome to an agent, the evidence available to the agent, and the agent's degrees of belief in the outcomes available. His formulation considered "possible worlds" to ensue based on which propositions came out true.

de Finetti (1937) validates the axioms of probability as identified by Kolmogorov (1933) through the betting argument described in this text. He also considers expected utility like Ramsey, but does not go so far as Ramsey, Savage, and later thinkers to consider an agent's degrees of belief as ascertainable from their decisions.

Cox (1946) with his consistency (or rationality) theorems, proves that any system of reasoning over uncertainty which admits his assumptions is equivalent to probability theory.

Savage (1954) offers an alternative decision theoretic formulation to that of Ramsey, one encompassing states, consequences, acts, events, and preferences (relative to an agent) over acts. A few of his assumptions, regarding the relation between consequences and acts and events and acts, later researchers found to be limiting.

Jeffrey (1983) demonstrates the analytic superiority of the Bayesian, degrees of belief interpretation in light of various failures of a frequentist approach by use of an agent expressing preferences and confidences. His approach suffers (or allows) a non-uniqueness problem in which the decisions of an agent might be shown to adhere to multiple formulations of expected utility.

Jaynes (2003) provides "a deeper logical foundation" of Kolmogorov's axioms, avoiding the infinite sets which paradoxically plague the formulation of de Finetti. Much of his text allows for diverse interpretation and he contrasts the Bayesian approach as expressing a prior on an agent's belief state (such as a model of the world) which can alternatively be ignored by assuming maximum entropy over the space of consideration.

12.2 Basic Probability Notation

Exercise 12.PAGB

Show from first principles that $P(a\,|\,b\wedge a)=1$.

The “first principles” needed here are the definition of conditional probability, $P(X|Y) = P(X\wedge Y)/P(Y)$, and the definitions of the logical connectives. It is not enough to say that if $B\wedge A$ is “given” then A must be true! From the definition of conditional probability, and the fact that conjunction is commutative, associative, and idempotent, we have

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

Exercise 12.SUMO

a. Suppose a sample space is defined by the Cartesian product of the ranges of a set of Boolean variables $X_1,\ldots,X_n$, and let ϕ be any logical proposition expressed in terms of these variables. Prove from first principles (including the basic axioms of probability, Equation (12.1)) that Equation (12.2), i.e., $P(\phi)=\sum_{\omega\in\phi}P(\omega)$.

b. Now show that $\sum_{x_i}P(X_i=x_i)=1$ for any variable X_i, i.e., the distribution for each random variable must sum to 1.

a. Assuming the propositions are independent, that there is a unique proposition which denotes the assignment to the random variables,

$$P(\phi) = P(\bigcup_{\omega\in\phi}\omega) = \sum_{\omega\in\phi}P(\omega)$$

.

b. As is the case for a random variable, each assignment is mutually exclusive and exhaustive; therefore each assignment is a possible world, ω, of the total sample space, Ω, of the random variable.

$$\sum_{x_i}P(X=x_i) = \sum_{\omega}P(X=\omega) = \sum_{\omega\in\Omega}P(\omega) = 1$$

.

Exercise 12.INEX

Prove Equation (12.5) from Equations (12.1) and (12.2).

Equation (12.5) states that $P(a \lor b) = P(a) + P(b) - P(a \land b)$. This can be proved directly from Equation (12.2), using obvious abbreviations for the possible-world probabilities:

$$\begin{aligned} P(a \lor b) &= p_{a,b} + p_{a,\neg b} + p_{\neg a,b} \\ P(a) &= p_{a,b} + p_{a,\neg b} \\ P(b) &= p_{a,b} + p_{\neg a,b} \\ P(a \land b) &= p_{a,b} \,. \end{aligned}$$

Exercise 12.TFPG

For each of the following statements, either prove it is true or give a counterexample.

a. If $P(a \mid b, c) = P(b \mid a, c)$, then $P(a \mid c) = P(b \mid c)$

b. If $P(a \mid b, c) = P(a)$, then $P(b \mid c) = P(b)$

c. If $P(a \mid b) = P(a)$, then $P(a \mid b, c) = P(a \mid c)$

a. True. By the product rule we know $P(b, c)P(a|b, c) = P(a, c)P(b|a, c)$, which by assumption reduces to $P(b, c) = P(a, c)$. Dividing through by $P(c)$ gives the result.

b. False. The statement $P(a|b, c) = P(a)$ merely states that a is independent of b and c, it makes no claim regarding the dependence of b and c. A counter-example: a and b record the results of two independent coin flips, and $c = b$.

c. False. While the statement $P(a|b) = P(a)$ implies that a is independent of b, it does not imply that a is conditionally independent of b given c. A counter-example: a and b record the results of two independent coin flips, and c equals the xor of a and b.

Exercise 12.RATB

Would it be rational for an agent to hold the three beliefs $P(A) = 0.4$, $P(B) = 0.3$, and $P(A \lor B) = 0.5$? If so, what range of probabilities would be rational for the agent to hold for $A \land B$? Make up a table like the one in Figure 12.2, and show how it supports your argument about rationality. Then draw another version of the table where $P(A \lor B) = 0.7$. Explain why it is rational to have this probability, even though the table shows one case that is a loss and three that just break even. (*Hint:* what is Agent 1 committed to about the probability of each of the four cases, especially the case that is a loss?)

Probably the easiest way to keep track of what's going on is to look at the probabilities of the atomic events. A probability assignment to a set of propositions is consistent with the axioms of probability if the probabilities are consistent with an assignment to the atomic events that sums to 1 and has all probabilities between 0 and 1 inclusive. We call the probabilities of the atomic events w, x, y, and z, as follows:

	B	$\neg B$
A	w	x
$\neg A$	y	z

We then have the following equations:

$$P(A) = w + x = 0.4$$
$$P(B) = w + y = 0.3$$
$$P(A \vee B) = w + x + Y = 0.5$$
$$P(True) = w + x + y + z = 1$$

From these, it is straightforward to infer that $w = 0.2$, $x = 0.2$, $y = 0.1$, and $z = 0.5$. Therefore, $P(A \wedge B) = w = 0.2$. Thus the probabilities given are consistent with a rational assignment, and the probability $P(A \wedge B)$ is exactly determined.

If $P(A \vee B) = 0.7$, then $P(A \wedge B) = w = 0$. Thus, even though the bet outlined in Figure 12.2 loses if A and B are both true, the agent believes this to be impossible so the bet is still rational.

Exercise 12.EXEX

This question deals with the properties of possible worlds, defined on page 410 as assignments to all random variables. We will work with propositions that correspond to exactly one possible world because they pin down the assignments of all the variables. In probability theory, such propositions are called **atomic events**. For example, with Boolean variables X_1, X_2, X_3, the proposition $x_1 \wedge \neg x_2 \wedge \neg x_3$ is an atomic event because it fixes the assignment of the variables; in the language of propositional logic, we would say it has exactly one model.

a. Prove, for the case of n Boolean variables, that any two distinct atomic events are mutually exclusive; that is, their conjunction is equivalent to *false*.

b. Prove that the disjunction of all possible atomic events is logically equivalent to $true$.

c. Prove that any proposition is logically equivalent to the disjunction of the atomic events that entail its truth.

a. Each atomic event is a conjunction of n literals, one per variable, with each literal either positive or negative. For the events to be distinct, at least one pair of corresponding literals must be nonidentical; hence, the conjunction of the two events contains the literals X_i and $\neg X_i$ for some i, so the conjunction reduces to $False$.

b. Proof by induction on n. For $n = 0$, the only event is the empty conjunction $True$, and the disjunction containing only this event is also $True$. Inductive step: assume the claim holds for n variables. The disjunction for $n + 1$ variables consists of pairs of disjuncts of the form $(T_n \wedge X_{n+1}) \vee (T_n \wedge \neg X_{n+1})$ for all possible atomic event conjunctions T_n. Each pair logically reduces to T_n, so the entire disjunction reduces to the disjunction for n variables, which by hypothesis is equivalent to $True$.

c. Let α be the sentence in question and $\mu_1, \ldots, \mu_k$ be the atomic event sentences that entail its truth. Let M_i be the model corresponding to μ_i (its *only* model). To prove that $\mu_1 \vee \ldots \vee \mu_k \equiv \alpha$, simply observe the following:

- Because $\mu_i \models \alpha$, α is true in all the models of μ_i, so α is true in M_i.
- The models of $\mu_1 \vee \ldots \vee \mu_k$ are exactly $M_1, \ldots, M_k$ because any two atomic events are mutually exclusive, so any given model can satisfy at most one disjunct, and

a model that satisfies a disjunct must be the model corresponding to that atomic event.

- If any model M satisfies α, then the corresponding atomic-event sentence μ entails α, so the models of α are exactly $M_1, \ldots, M_k$.

Hence, α and $\mu_1 \vee \ldots \vee \mu_k$ have the same models, so are logically equivalent.

Exercise 12.POKF

Consider the set of all possible five-card poker hands dealt fairly from a standard deck of fifty-two cards.

a. How many atomic events are there in the joint probability distribution (i.e., how many five-card hands are there)?

b. What is the probability of each atomic event?

c. What is the probability of being dealt a royal straight flush? Four of a kind?

This is a classic combinatorics question that could appear in a basic text on discrete mathematics. The point here is to refer to the relevant axioms of probability. The question also helps students to grasp the concept of the joint probability distribution as the distribution over all possible states of the world.

a. There are $\binom{52}{5}$ = $(52 \times 51 \times 50 \times 49 \times 48)/(1 \times 2 \times 3 \times 4 \times 5)$ = 2,598,960 possible five-card hands. Note that the order of the cards does not matter.

b. By the fair-dealing assumption, each of these is equally likely. By Equation (12.1), each hand therefore occurs with probability 1/2,598,960.

c. There are four hands that are royal straight flushes (one in each suit). By Equation (12.2), since the events are mutually exclusive, the probability of a royal straight flush is just the sum of the probabilities of the atomic events, i.e., 4/2,598,960 = 1/649,740. For "four of a kind" events, there are 13 possible "kinds" and for each, the fifth card can be one of 48 possible other cards. The total probability is therefore $(13 \times 48)/2,598,960 = 1/4,165$.

These questions can easily be augmented by more complicated ones, e.g., what is the probability of getting a full house given that you already have two pairs? What is the probability of getting a flush given that you have three cards of the same suit? Or you could assign a project of producing a poker-playing agent, and have a tournament among them. Note that poker play (mostly betting) is complicated by the game-theoretic nature of the problem. For example, a player who bets a large amount may be bluffing. See Chapter 17.

Exercise 12.PASC

In his letter of August 24, 1654, Pascal was trying to show how a pot of money should be allocated when a gambling game must end prematurely. Imagine a game where each turn consists of the roll of a die, player *E* gets a point when the die is even, and player *O* gets a point when the die is odd. The first player to get 7 points wins the pot. Suppose the game is interrupted with E leading 4–2. How should the money be fairly split in this case? What is the general formula? (Fermat and Pascal made several errors before solving the problem, but you should be able to get it right the first time.)

Let e and o be the initial scores, m be the score required to win, and p be the probability that E wins each round. One can easily write down a recursive formula for the probability that E wins from the given initial state:

$$w_E(p,e,o,m) = \begin{cases} 1 & \text{if } e=m \\ 0 & \text{if } o=m \\ p \cdot w_E(p,e+1,o,m) + (1-p)\cdot w_E(p,e,o+1,m) & \text{otherwise} \end{cases}$$

This translates directly into code that can be used to compute the answer,

$$w_E(0.5,4,2,7) = 0.7734375 \,.$$

With a bit more work, we can derive a nonrecursive formula:

$$w_E(p,e,o,m) = p^{m-e} \sum_{i=0}^{m-o-1} \binom{i+m-e-1}{i} (1-p)^i \,.$$

Each term in the sum corresponds to the probability of winning by exactly a particular score; e.g., starting from 4–2, one can win by 7–2, 7–3, 7–4, 7–5, or 7–6. Each final score requires E to win exactly $m-e$ rounds while the opponent wins exactly i rounds, where $i=0,1,\ldots,m-o-1$; and the combinatorial term counts the number of ways this can happen without E winning first by a larger margin. One can check the nonrecursive formula by showing that it satisfies the recursive formula. (It may be helpful to suggest to students that they start by building the lattice of states implied by the above recursive formula and calculating (bottom-up) the symbolic win probabilities in terms of p rather than 0.5, so that they can see the general shape emerging.)

Exercise 12.XXXX

In this question we consider conditional distributions for binary random variables, expressed as tables.

a. A, B, C, and D are binary random variables. How many entries are in the following conditional probability tables and what is the sum of the values in each table?

(i) $\mathbf{P}(A \mid C)$

(ii) $\mathbf{P}(A, D \mid B = \mathit{true}, C = \mathit{true})$

(iii) $\mathbf{P}(B \mid A = \mathit{true}, C, D)$

b. Consider the conditional distribution $\mathbf{P}(X_1,\ldots X_\ell \mid Y_1,\ldots,Y_m, Z_1 = \mathbf{z}_1,\ldots Z_n = z_n$, represented as a complete table. Assuming that all variables are binary, derive expressions for the *number* of the probabilities in the table and their *sum*.

a. (i) $\mathbf{P}(A \mid C)$: 4 and 2.
(ii) $\mathbf{P}(A, D \mid B = \mathit{true}, C = \mathit{true})$: 4 and 1.
(iii) $\mathbf{P}(B \mid A = \mathit{true}, C, D)$: 8 and 4.

b. Because the Z-variables are all instantiated, they play no role. There are 2^m "rows" of the table, each corresponding to an assignment to the Y-variables, and each row has 2^ℓ entries for the possible assignments to the X-variables, so there are $2^{\ell+m}$ entries in all. Each row must sum to one, being a distribution over the X-variables, so the sum is 2^m.

12.3 Inference Using Full Joint Distributions

Exercise 12.XXXX

Given the full joint distribution shown in Figure 12.3, calculate the following:

a. $\mathbf{P}(\mathit{toothache})$.
b. $\mathbf{P}(\mathit{Cavity})$.
c. $\mathbf{P}(\mathit{Toothache} \mid \mathit{cavity})$.
d. $\mathbf{P}(\mathit{Cavity} \mid \mathit{toothache} \lor \mathit{catch})$.

The main point of this exercise is to understand the various notations of bold versus non-bold P, and uppercase versus lowercase variable names. The rest is easy, involving a small matter of addition.

a. This asks for the probability that *Toothache* is true.

$$P(\mathit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

b. This asks for the vector of probability values for the random variable *Cavity*. It has two values, which we list in the order $\langle \mathit{true}, \mathit{false} \rangle$. First add up $0.108 + 0.012 + 0.072 + 0.008 = 0.2$. Then we have

$$\mathbf{P}(\mathit{Cavity}) = \langle 0.2, 0.8 \rangle \ .$$

c. This asks for the vector of probability values for *Toothache*, given that *Cavity* is true.

$$\mathbf{P}(\mathit{Toothache}|\mathit{cavity}) = \langle (.108 + .012)/0.2, (0.072 + 0.008)/0.2 \rangle = \langle 0.6, 0.4 \rangle$$

d. This asks for the vector of probability values for *Cavity*, given that either *Toothache* or *Catch* is true. First compute $P(\mathit{toothache} \lor \mathit{catch}) = 0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.144 = 0.416$. Then

$$\begin{aligned}&\mathbf{P}(\mathit{Cavity}|\mathit{toothache} \lor \mathit{catch}) = \\ &\quad \langle (0.108 + 0.012 + 0.072)/0.416, (0.016 + 0.064 + 0.144)/0.416 \rangle = \\ &\quad \langle 0.4615, 0.5384 \rangle\end{aligned}$$

12.4 Independence

Exercise 12.XXXX

Find values for the probabilities a and b in joint probability table below so that the binary variables X and Y are independent.

X	Y	$P(X,Y)$
t	t	$3/5$
t	f	$1/5$
f	t	a
f	f	b

First, we know that the distribution sums to 1, so $a+b=1/5$. Second, independence requires that $\mathbf{P}(X,Y) = \mathbf{P}(X)\mathbf{P}(Y)$. From the first two rows, where the value of X is the same, we have that $P(Y=t)/P(Y=f)=3$. In the third and fourth row, the value of X is the same, so $a/b=3$ also. Hence, $a=3/20$ and $b=1/20$.

Exercise 12.XXXX

Suppose X and Y are *independent* random variables over the domain $\{1,2,3\}$ with $P(X=3) = 1/6$. Given the following partially specified joint distribution, what are the remaining values? Write your answers as simplified fractions in the blanks.

$X \setminus Y$	1	2	3
1	1/4	1/16	
2	1/6	1/24	
3			

Since X and Y are independent, we have that $P(X=x, Y=y) = P(X=x)P(Y=y)$ for all x and y, so it suffices to determine the marginal distributions $P(X)$ and $P(Y)$.

We begin with $P(X)$. First observe that

$$\frac{P(X=1, Y=1)}{P(X=2, Y=1)} = \frac{P(X=1)P(Y=1)}{P(X=2)P(Y=1)} = \frac{P(X=1)}{P(X=2)} = \frac{1/4}{1/6} = \frac{3}{2}.$$

Combining this with the fact that any probability distribution sums to 1, we find that

$$P(X=1)+P(X=2)+P(X=3) = \frac{3}{2}\cdot P(X=2)+P(X=2)+\frac{1}{6} = \frac{5}{2}\cdot P(X=2)+\frac{1}{6} = 1,$$

which implies $P(X=2) = 1/3$. It follows that $P(X=1) = (3/2)\cdot P(X=2) = 1/2$.

To recover the marginal distribution of Y, we note that

$$\begin{aligned} P(X=1,Y=1) &= P(X=1)P(Y=1) = (1/2)\cdot P(Y=1) = 1/4, \\ P(X=1,Y=2) &= P(X=1)P(Y=2) = (1/2)\cdot P(Y=2) = 1/16, \end{aligned}$$

so $P(Y=1) = 1/2$ and $P(Y=2) = 1/8$. It follows that $P(Y=3) = 1 - P(Y=1) - P(Y=2) = 3/8$.

Exercise 12.XXXX

a. Suppose $A \perp\!\!\!\perp B$. Determine the missing entries x and y of the joint distribution $P(A,B)$, where A and B take values in $\{0,1\}$.

$$\begin{aligned} P(A=0,B=0) &= 0.1 \\ P(A=0,B=1) &= 0.3 \\ P(A=1,B=0) &= x \\ P(A=1,B=1) &= y \end{aligned}$$

b. Suppose $B \perp\!\!\!\perp C \mid A$. Determine the missing entries x,y,z of the joint distribution $P(A,B,C)$.

$$\begin{aligned} P(A=0,B=0,C=0) &= 0.01 \\ P(A=0,B=0,C=1) &= 0.02 \\ P(A=0,B=1,C=0) &= 0.03 \\ P(A=0,B=1,C=1) &= x \\ P(A=1,B=0,C=0) &= 0.01 \\ P(A=1,B=0,C=1) &= 0.1 \\ P(A=1,B=1,C=0) &= y \\ P(A=1,B=1,C=1) &= z \end{aligned}$$

a. To solve this we use the two constraints: A and B are independent, and the distribution sums to 1. From independence, we have

$$y/x = \frac{P(A=1,B=1)}{P(A=1,B=0)} = \frac{P(A=0,B=1)}{P(A=0,B=0)} = \frac{P(B=1)}{P(B=0) = 3}\,.$$

So $y = 3x$. From sum-to-1 we have $x + y = 0.6$. Solving, we obtain $x = 0.15$ and $y = 0.45$.

b. From conditional independence, as in (a), we obtain $x = 0.03 \cdot \frac{0.02}{0.01} = 0.06$ and $z/y = 10$. From sum-to-1, we have $0.01 + 0.02 + 0.03 + 0.06 + 0.01 + 0.1 + y + z = 1$ so $y + z = 0.77$. Solving, we get $y = 0.07, z = 0.7$.

Exercise 12.XXXX

Deciding to put probability theory to good use, we encounter a slot machine with three independent wheels, each producing one of the four symbols BAR, BELL, LEMON, or CHERRY with equal probability. The slot machine has the following payout scheme for a bet of 1 coin (where "?" denotes that we don't care what comes up for that wheel):

BAR/BAR/BAR pays 20 coins
BELL/BELL/BELL pays 15 coins
LEMON/LEMON/LEMON pays 5 coins
CHERRY/CHERRY/CHERRY pays 3 coins
CHERRY/CHERRY/? pays 2 coins
CHERRY/?/? pays 1 coin

a. Compute the expected "payback" percentage of the machine. In other words, for each coin played, what is the expected coin return?

b. Compute the probability that playing the slot machine once will result in a win.

c. Estimate the mean and median number of plays you can expect to make until you go broke, if you start with 10 coins. You can run a simulation to estimate this, rather than trying to compute an exact answer.

a. To compute the expected payback for the machine, we determine the probability for each winning outcome, multiply it by the amount that would be won in that instance, and sum all possible winning combinations. Since each symbol is equally likely, the first four cases have probability $(1/4)^3 = 1/64$.

However, in the case of computing winning probabilities for cherries, we must only consider the highest paying case, so we must subtract the probability for dominating winning cases from each subsequent case (e.g., in the case of two cherries, we subtract off the probability of getting three cherries):

CHERRY/CHERRY/? $3/64 = (1/4)^2 - 1/64$
CHERRY/?/? $12/64 = (1/4)^1 - 3/64 - 1/64$

The expectation is therefore

$$20 \cdot 1/64 + 15 \cdot 1/64 + 5 \cdot 1/64 + 3 \cdot 1/64 + 2 \cdot 3/64 + 1 \cdot 12/64 = 61/64.$$

Thus, the expected payback percentage is 61/64 (which is less than 1 as we would expect of a slot machine that was actually generating revenue for its owner).

b. We can tally up the probabilities we computed in the previous section, to get

$$1/64 + 1/64 + 1/64 + 1/64 + 3/64 + 12/64 = 19/64.$$

Alternatively, we can observe that we win if either all symbols are the same (denote this event S), or if the first symbol is cherry (denote this event C). Then applying the inclusion-exclusion identity for disjunction:

$$P(S \vee C) = P(S) + P(C) - P(S \wedge C) = (1/4)^2 + 1/4 - 1/64 = 19/64.$$

c. Using a simple Python simulation, we find a mean of about 210, and a median of 21. This shows the distribution of number of plays is heavy tailed: most of the time you run out of money relatively quickly, but occasionally you last for thousands of plays.

```
import random

def trial():
        funds = 10
        plays = 0
        while funds >= 1:
                funds -=1
                plays += 1
                slots = [random.choice(
                          ["bar", "bell", "lemon", "cherry"])
                         for i in range(3)]
                if slots[0] == slots[1]:
                        if slots[1] == slots[2]:
                                num_equal = 3
                        else:
                                num_equal = 2
                else:
                        num_equal = 1
                if slots[0] == "cherry":
                        funds += num_equal
                elif num_equal == 3:
                        if slots[0] == "bar":
                                funds += 20
                        elif slots[0] == "bell":
                                funds += 15
                        else:
                                funds += 5
        return plays

def test(trials):
        results = [trial() for i in xrange(trials)]
        mean = sum(results) / float(trials)
        median = sorted(results)[trials/2]
        print "%s trials: mean=%s, median=%s" % (trials, mean, median)

test(10000)
```

Exercise 12.XXXX

We wish to transmit an n-bit message to a receiving agent. The bits in the message are independently corrupted (flipped) during transmission with ϵ probability each. With an extra parity bit sent along with the original information, a message can be corrected by the receiver if at most one bit in the entire message (including the parity bit) has been corrupted. Suppose we want to ensure that the correct message is received with probability at least $1 - \delta$. What is the maximum feasible value of n? Calculate this value for the case $\epsilon = 0.001$, $\delta = 0.01$.

The correct message is received if either zero or one of the $n+1$ bits are corrupted. Since corruption occurs independently with probability ϵ, the probability that zero bits are corrupted is $(1-\epsilon)^{n+1}$. There are $n+1$ mutually exclusive ways that exactly one bit can be corrupted, one for each bit in the message. Each has probability $\epsilon(1-\epsilon)^n$, so the overall probability that exactly one bit is corrupted is $n\epsilon(1-\epsilon)^n$. Thus, the probability that the correct message is received is $(1-\epsilon)^{n+1} + n\epsilon(1-\epsilon)^n$.

The maximum feasible value of n, therefore, is the largest n satisfying the inequality

$$(1-\epsilon)^{n+1} + n\epsilon(1-\epsilon)^n \geq 1-\delta.$$

Numerically solving this for $\epsilon = 0.001$, $\delta = 0.01$, we find $n = 147$.

Exercise 12.INDI

Show that the three forms of independence in Equation (12.11) are equivalent.

Independence is symmetric (that is, a and b are independent iff b and a are independent) so $P(a|b) = P(a)$ is the same as $P(b|a) = P(b)$. So we need only prove that $P(a|b) = P(a)$ is equivalent to $P(a \wedge b) = P(a)P(b)$. The product rule, $P(a \wedge b) = P(a|b)P(b)$, can be used to rewrite $P(a \wedge b) = P(a)P(b)$ as $P(a|b)P(b) = P(a)P(b)$, which simplifies to $P(a|b) = P(a)$.

Exercise 12.XXXX

Consider the following probability distributions:

A	$P(A)$
t	0.8
f	0.2

A	B	$P(B\|A)$
t	t	0.9
t	f	0.1
f	t	0.6
f	f	0.4

B	C	$P(C\|B)$
t	t	0.8
t	f	0.2
f	t	0.8
f	f	0.2

C	D	$P(D\|C)$
t	t	0.25
t	f	0.75
f	t	0.5
f	f	0.5

Given just these tables and no independence assumptions, calculate the following probabilities, showing your working. If it is impossible to calculate without more independence assumptions, specify instead a minimal set of independence assumptions that would allow you to answer the question.

a. $P(a, \neg b)$.

b. $P(b)$.

c. $P(\neg a, \neg b, c)$.

d. Now assume C is independent of A given B and D is independent of A and B given C. Calculate $P(a, \neg b, c, d)$.

a. $P(a, \neg b) = P(a)P(\neg b \mid a) = 0.8 \times 0.1 = 0.08$.

b.

$$\begin{aligned} P(b) &= \sum_a P(b \,|\, a)P(a) = P(b \,|\, a)P(a) + P(b \,|\, \neg a)P(\neg a) \\ &= (0.9 \times 0.8) + (0.6 \times 0.2) = 0.84\ . \end{aligned}$$

c. Not possible. Could do it given that C is independent of A given B.

d. This is a preview of some ideas in Chapter 13, specifically the chain rule. It might be a good idea to provide the chain rule as part of the question, but it can be done using material from this chapter by repeated application of the product rule and use of the given conditional independence assertions:

$$\begin{aligned} P(a, \neg b, c, d) &= P(\neg b, c, d \,|\, a)P(a) = P(c, d \,|\, a, \neg b)P(\neg b \,|\, a)P(a) \\ &= P(d \,|\, a, \neg b, c)P(c \,|\, a, \neg b)P(\neg b \,|\, a)P(a) \\ &= P(d \,|\, c)P(c \,|\, \neg b)P(\neg b \,|\, a)P(a) = 0.25 \times 0.8 \times 0.1 \times 0.8 = 0.016\ . \end{aligned}$$

Exercise 12.XXXX

For each of the following assertions, say whether it is true or false and support your answer with arguments or counterexamples where appropriate.

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

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

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

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

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

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

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

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

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

j. $\mathbf{P}(A, B) = \sum_c \mathbf{P}(A \,|\, B, C = c)\mathbf{P}(B \,|\, C = c)\mathbf{P}(C = c)$.

k. If $\mathbf{P}(X, Y \,|\, Z) = \mathbf{P}(X \,|\, Z)\mathbf{P}(Y \,|\, Z)$ then X is independent of Y given Z.

l. If $\mathbf{P}(X, Y, Z) = \mathbf{P}(X, Z)\mathbf{P}(Y)$ then X is independent of Y given Z.

m. If $\mathbf{P}(X, Y, Z) = \mathbf{P}(X)\mathbf{P}(Z)\mathbf{P}(Y)$ then X is independent of Y given Z.

a. False, $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$ iff. $A \perp\!\!\!\perp B$.

b. False, $\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A,B)}{\mathbf{P}(B)}$.

c. False, eg. if $\mathbf{P}(A) \perp\!\!\!\perp \mathbf{P}(B)$ then $\mathbf{P}(A, B) \neq \mathbf{P}(A)\mathbf{P}(B) - \mathbf{P}(A \,|\, B) = \mathbf{P}(A)\mathbf{P}(B) - \mathbf{P}(A)$.

d. True by application of the chain rule .

e. False, $\mathbf{P}(A, B, C) = \mathbf{P}(C \,|\, A, B)\mathbf{P}(A \,|\, B)\mathbf{P}(B)$.

f. False, $\mathbf{P}(A, B, C) = \mathbf{P}(A \,|\, B, C)\mathbf{P}(B \,|\, C)\mathbf{P}(C)$.

g. False, $\mathbf{P}(A \,|\, B, C)\mathbf{P}(B \,|\, A, C)\mathbf{P}(C \,|\, A, B) = \frac{3\mathbf{P}(A,B,C)}{\mathbf{P}(A,B)\mathbf{P}(A,C)\mathbf{P}(B,C)}$.

h. True by application of the product rule where $A' = B, C$ and $B' = A$.

i. True by application of the product rule where $A' = C$ and $B' = A, C$.

j. True by definition of a probability distribution, $\mathbf{P}(C)$, over all assignments—the marginalization of a conditioned variable.

k. True, $X \perp\!\!\!\perp Y \mid Z$ by the definition of conditional independence.

l. False, for $X \perp\!\!\!\perp Y \mid Z$, $\mathbf{P}(X, Y, Z) = \mathbf{P}(X \mid Z)\ \mathbf{P}(Y \mid Z)\ \mathbf{P}(Z)$.

m. False, for $X \perp\!\!\!\perp Y \mid Z$, $\mathbf{P}(X, Y, Z) = \mathbf{P}(X \mid Z)\ \mathbf{P}(Y \mid Z)\ \mathbf{P}(Z)$.

Exercise 12.XXXX

a. Assuming that A is independent of B given C and that A and C are absolutely independent, what is the most factored representation of $\mathbf{P}(A, B, C)$?

b. Assuming that A is independent of B given C, what is the most factored representation of $\mathbf{P}(A, B \mid C)$?

c. Given no independence assumptions, what is the most factored representation of $\mathbf{P}(A \mid B, C)$?

d. Assuming that A is independent of B given C, what is the most factored representation of $\mathbf{P}(A \mid B, C)$?

e. Assuming that A is absolutely independent of B, what is the most factor representation of $\mathbf{P}(A \mid B, C)$?

f. Assuming that A is independent of B given C, write an expression equivalent to $\mathbf{P}(A \mid B)$ using A, B, and C.

g. Which of the following expressions are equal to 1, given no independence assumptions?

(i) $\sum_a \mathbf{P}(A = a \mid B)$.

(ii) $\sum_b \mathbf{P}(A \mid B = b)$.

(iii) $\sum_a \sum_b P(A = a, B = b)$.

(iv) $\sum_a \sum_b P(A = a \mid B = b)$.

(v) $\sum_a \sum_b P(A = a)\ P(B = b)$.

h. Which of the following expressions hold for any distribution over four random variables A, B, C and D?

(i) $\mathbf{P}(A, B \mid C, D) = \mathbf{P}(A \mid C, D)\mathbf{P}(B \mid A, C, D)$.

(ii) $\mathbf{P}(A, B) = \mathbf{P}(A, B \mid C, D)\mathbf{P}(C, D)$.

(iii) $\mathbf{P}(A, B \mid C, D) = \mathbf{P}(A, B)\mathbf{P}(C, D)\mathbf{P}(C, D \mid A, B)$.

(iv) $\mathbf{P}(A, B \mid C, D) = \mathbf{P}(A, B)\mathbf{P}(D)\mathbf{P}(C, D \mid A, B)$.

a. $\mathbf{P}(A, B \mid C)\ \mathbf{P}(C) = \mathbf{P}(A \mid C)\ \mathbf{P}(B \mid C)\ \mathbf{P}(C) = \mathbf{P}(A)\ \mathbf{P}(B \mid C)\ \mathbf{P}(C)$.

b. $\mathbf{P}(A \mid B, C) = \mathbf{P}(A \mid C)\ \mathbf{P}(B \mid C)$.

c. The same or $\frac{\mathbf{P}(B,C \mid A)\ \mathbf{P}(A)}{\mathbf{P}(B,C)}$.

d. $\mathbf{P}(A \mid B, C) = \frac{\mathbf{P}(A,B,C)}{\mathbf{P}(B,C)} = \frac{\mathbf{P}(A,B \mid C)\ \mathbf{P}(C)}{\mathbf{P}(B \mid C)\ \mathbf{P}(C)} = \frac{\mathbf{P}(A \mid C)\ \mathbf{P}(B \mid C)\ \mathbf{P}(C)}{\mathbf{P}(B \mid C)\ \mathbf{P}(C)} = \mathbf{P}(A \mid C)$.

e. The same, or: $\mathbf{P}(A \mid B, C) = \frac{\mathbf{P}(A,B,C)}{\mathbf{P}(B,C)} = \frac{\mathbf{P}(C \mid A,B)\ \mathbf{P}(A)\ \mathbf{P}(B)}{\mathbf{P}(B,C)}$.

f. $\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A,B)}{\mathbf{P}(B)} = \frac{\sum_c P(A,B \mid C=c)\ P(C=c)}{\mathbf{P}(B)} = \frac{\sum_c P(A \mid C=c) P(B \mid C=c)\ P(C=c)}{\mathbf{P}(B)}$.

g. Most steps follow the rules for marginalization or conditioning.

(i) $\sum_a \mathbf{P}(A = a \mid B) = 1$.

(ii) $\sum_b \mathbf{P}(A \mid B = b) = \alpha \mathbf{P}(A)$.

(iii) $\sum_a \sum_b P(A = a, B = b) = \sum_a P(A = a) = 1$.

(iv) $\sum_a \sum_b P(A = a \mid B = b) = \sum_a \alpha P(A = a) = \alpha$.

(v) $\sum_a \sum_b P(A = a)\ P(B = b) = \sum_a P(A = a) \sum_b P(B = b) = \sum_a P(A = a) = 1$.

h. (i) True by the product rule; $\mathbf{P}(A, B \mid C, D) = \mathbf{P}(B \mid A, C, D)\ \mathbf{P}(A \mid C, D)$.

(ii) False; $\mathbf{P}(A, B) = \sum_c \sum_d P(A, B \mid C = c, D = d)\ P(C = c, D = d)$.

(iii) False; $\mathbf{P}(A, B \mid C, D) = \mathbf{P}(A, B) \frac{1}{\mathbf{P}(C,D)} \mathbf{P}(C, D \mid A, B)$.

(iv) False; see above.

Exercise 12.XXXX

For each of the following expression, derive an equivalent expression in terms of *only* the given conditional distributions, using the given conditional independence assumptions. If it is not possible, say so.

a. Express $\mathbf{P}(A, B \mid C)$ in terms of $\mathbf{P}(A), \mathbf{P}(A \mid C), \mathbf{P}(B \mid C), \mathbf{P}(C \mid A, B)$ given no conditional independence assumptions.

b. Express $\mathbf{P}(B \mid A, C)$ in terms of $\mathbf{P}(A), \mathbf{P}(A \mid C), \mathbf{P}(B \mid A), \mathbf{P}(C \mid A, B)$ and no conditional independence assumptions.

c. Express $\mathbf{P}(C)$ in terms of $\mathbf{P}(A \mid B), \mathbf{P}(B), \mathbf{P}(B \mid A, C), \mathbf{P}(C \mid A)$ given that A and B are absolutely independent.

d. Express $\mathbf{P}(A, B, C)$ in terms of $\mathbf{P}(A \mid B, C), \mathbf{P}(B), \mathbf{P}(B \mid A, C), \mathbf{P}(C \mid B, A)$ given that A is conditionally independent of B given C.

a. Not possible.

b. $\frac{\mathbf{P}(A)\ \mathbf{P}(B \mid A)\ \mathbf{P}(C \mid A,B)}{\sum_b \mathbf{P}(A)\ \mathbf{P}(B \mid A)\ \mathbf{P}(C \mid A,B)}$

c. $\sum_a \mathbf{P}(A \mid B)\ \mathbf{P}(C \mid A)$

d. Not possible.

Exercise 12.XXXX

a. Consider the following two assertions, where **U**, **V**, **W**, **X**, **Y**, and **Z** are *sets* of random variables:

(i) **U** is independent of **V** given **W**.

(ii) **X** is independent of **Y** given **Z**.

Under what conditions, in terms of subset/superset relationships among **U**, **V**, **W**, **X**, **Y**, and **Z**, can one say that (i) entails (ii)?

b. We will say that a conditional independence assertion C_1 is **strictly weaker** than an assertion C_2 if C_2 entails C_1 and C_1 does not entail C_2. For each of the following equations, give the weakest conditional independence assertion required to make it true (if any).

(i) $\mathbf{P}(A,C) = \mathbf{P}(A \mid B)\, \mathbf{P}(C)$.

(ii) $\mathbf{P}(A \mid B,C) = \frac{\mathbf{P}(A)\, \mathbf{P}(B \mid A)\, \mathbf{P}(C \mid A)}{\mathbf{P}(B \mid C)\, \mathbf{P}(C)}$

(iii) $\mathbf{P}(A,B) = \sum_c \mathbf{P}(A \mid B,c)\, \mathbf{P}(B \mid c)\, \mathbf{P}(c)$

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

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

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

a. $\mathbf{X} \subseteq \mathbf{U}$, $\mathbf{Y} \subseteq \mathbf{V}$, and $\mathbf{Z} \supseteq \mathbf{W}$.

b. (i) $A \perp\!\!\!\perp C$ and $A \perp\!\!\!\perp B$.

(ii) $B \perp\!\!\!\perp C \mid A$.

(iii) None required by conditioning and the product rule.

(iv) None required by the product rule.

(v) $A \perp\!\!\!\perp D \mid B,C$.

(vi) $B \perp\!\!\!\perp C$.

Exercise 12.XXXX

Simplify each of the following expressions down a single probability term, without making any independence assumptions:

a. $\sum_{a'} \mathbf{P}(a' \mid D)\mathbf{P}(b \mid a', D)$.

b. $\sum_{b',c',d'} \mathbf{P}(A)\mathbf{P}(b' \mid A)\mathbf{P}(c' \mid A, b')\mathbf{P}(d' \mid A, b', c')$.

c. $\frac{\sum_{b'} \mathbf{P}(A \mid b', D)\mathbf{P}(b' \mid D)\mathbf{P}(D)}{\sum_{a',b'} \mathbf{P}(a' \mid b', D)\mathbf{P}(b' \mid D)\mathbf{P}(D)}$.

a. $\mathbf{P}(b \mid D)$

b. $\mathbf{P}(A)$

c. $\mathbf{P}(A \mid D)$

12.5 Bayes' Rule and Its Use

Exercise 12.XXXX

Formulate each of the following as probability models, stating all your assumptions, and use the probability model to answer the questions.

a. Aliens can be friendly or not; 75% are friendly. Friendly aliens arrive during the day 90% of the time, while unfriendly ones always arrive at night. If an alien arrives at night, how likely is it to be friendly?

b. Half of all monsters live in attics, while the rest live in basements. While 80 % of all monsters are fuzzy, *all* attic-living monsters are fuzzy. What is the probability that a basement-living monsters is fuzzy?

a. Let $\langle f, \neg f\rangle$ represent whether or not an alien is friendly and $< d, \neg d\rangle$ whether an alien arrives during the day. Then, $P(f) = .75$, $P(\neg f) = .25$, $P(d \,|\, f) = .9$, $P(\neg d \,|\, f) = .1$, $P(d \,|\, \neg f) = 0$, and $P(\neg d \,|\, \neg f) = 1$. Thus,

$$\begin{aligned} P(f \,|\, \neg d) &= \frac{P(\neg d \,|\, f)\; P(f)}{P(\neg d)} \\ &= \frac{P(\neg d \,|\, f)\; P(f)}{P(\neg d \,|\, f)\; P(f) + P(\neg d \,|\, \neg f)\; P(\neg f)} \\ &= \frac{.1 \times .75}{.1 \times .75 + 1 \times .25} \\ &\approx .23\;. \end{aligned}$$

b. Let $\langle a, b\rangle$ represent whether a monster lives in the attic or basement and $\langle f, \neg f\rangle$ whether it is fuzzy. Then we have $P(a) = .5$, $P(b) = .5$, $P(f) = .8$, and $P(f \,|\, a) = 1$, $P(\neg f \,|\, a) = 0$. Thus,

$$\begin{aligned} P(f) &= P(f \,|\, b)\; P(b) + P(f \,|\, a)\; P(a) \\ P(f \,|\, b) &= \frac{P(f) - P(f \,|\, a)\; P(a)}{P(b)} \\ &= \frac{.8 - 1 \times .5}{.5} \\ &= .6\;. \end{aligned}$$

Exercise 12.XXXX

Let A and B be Boolean random variables. You are given the following quantities:

$$\begin{aligned} P(A = Jtrue) &= \tfrac{1}{2} \\ P(B = true \,|\, A = \text{true}) &= 1 \\ P(B = true) &= \tfrac{3}{4} \end{aligned}$$

What is $P(B = true \,|\, A = false)$?

The simplest way to solve this is to realize that

$$\mathbf{P}(B = true) = \mathbf{P}(B = true \,|\, A = true)\mathbf{P}(A = true) + \mathbf{P}(B = true \,|\, A = false)\mathbf{P}(A = false).$$

Using this fact, you can solve for $\mathbf{P}(B = \mathit{true} \mid A = \mathit{false})$:

$$\begin{aligned}(1)\left(\frac{1}{2}\right) + \mathbf{P}(B = \mathit{true} \mid A = \mathit{false})\left(\frac{1}{2}\right) &= \frac{3}{4}\\ \Rightarrow\ \mathbf{P}(B = \mathit{true} \mid A = \mathit{false}) &= \frac{1}{2}\end{aligned}$$

Exercise 12.MDAB

Consider two medical tests, A and B, for a virus. Test A is 95% effective at recognizing the virus when it is present, but has a 10% false positive rate (indicating that the virus is present, when it is not). Test B is 90% effective at recognizing the virus, but has a 5% false positive rate. The two tests use independent methods of identifying the virus. The virus is carried by 1% of all people. Say that a person is tested for the virus using only one of the tests, and that test comes back positive for carrying the virus. Which test returning positive is more indicative of someone really carrying the virus? Justify your answer mathematically.

Let V be the statement that the patient has the virus, and A and B the statements that the medical tests A and B returned positive, respectively. The problem statement gives:

$$\begin{aligned}P(V) &= 0.01\\ P(A|V) &= 0.95\\ P(A|\neg V) &= 0.10\\ P(B|V) &= 0.90\\ P(B|\neg V) &= 0.05\end{aligned}$$

The test whose positive result is more indicative of the virus being present is the one whose posterior probability, $P(V|A)$ or $P(V|B)$ is largest. One can compute these probabilities directly from the information given, finding that $P(V|A) = 0.0876$ and $P(V|B) = 0.1538$, so B is more indicative.

Equivalently, the questions is asking which test has the highest posterior odds ratio $P(V|A)/P(\neg V|A)$. From the odd form of Bayes theorem:

$$\frac{P(V|A)}{P(\neg V|A)} = \frac{P(A|V)}{P(A|\neg V)}\frac{P(V)}{P(\neg V)}$$

we see that the ordering is independent of the probability of V, and that we just need to compare the likelihood ratios $P(A|V)/P(A|\neg V) = 9.5$ and $P(B|V)/P(V|\neg V) = 18$ to find the answer.

Exercise 12.HEDP

Suppose you are given a coin that lands *heads* with probability x and *tails* with probability $1 - x$. Are the outcomes of successive flips of the coin independent of each other given that you know the value of x? Are the outcomes of successive flips of the coin independent of each other if you do *not* know the value of x? Justify your answer.

If the probability x is known, then successive flips of the coin are independent of each other, since we know that each flip of the coin will land $heads$ with probability x. Formally, if $F1$ and $F2$ represent the results of two successive flips, we have

$$P(F1 = heads, F2 = heads|x) = x * x = P(F1 = heads|x)P(F2 = heads|x)$$

Thus, the events $F1 = heads$ and $F2 = heads$ are independent.

If we do not know the value of x, however, the probability of each successive flip is dependent on the result of all previous flips. The reason for this is that each successive flip gives us information to better estimate the probability x (i.e., determining the posterior estimate for x given our prior probability and the evidence we see in the most recent coin flip). This new estimate of x would then be used as our "best guess" of the probability of the coin coming up $heads$ on the next flip. Since this estimate for x is based on all the previous flips we have seen, the probability of the next flip coming up $heads$ depends on how many $heads$ we saw in all previous flips, making them dependent.

For example, if we had a uniform prior over the probability x, then one can show that after n flips if m of them come up heads then the probability that the next one comes up heads is $(m + 1)/(n + 2)$, showing dependence on previous flips.

Exercise 12.XXXX

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

We are given the following information:

$$\begin{aligned} P(test|disease) &= 0.99 \\ P(\neg test|\neg disease) &= 0.99 \\ P(disease) &= 0.0001 \end{aligned}$$

and the observation $test$. What the patient is concerned about is $P(disease|test)$. Roughly speaking, the reason it is a good thing that the disease is rare is that $P(disease|test)$ is proportional to $P(disease)$, so a lower prior for $disease$ will mean a lower value for $P(disease|test)$. Roughly speaking, if 10,000 people take the test, we expect 1 to actually have the disease, and most likely test positive, while the rest do not have the disease, but 1% of them (about 100 people) will test positive anyway, so $P(disease|test)$ will be about 1 in 100. More precisely, using the normalization equation from page 498:

$$\begin{aligned}
&P(disease|test)\\
&\quad = \frac{P(test|disease)P(disease)}{P(test|disease)P(disease)+P(test|\neg disease)P(\neg disease)}\\
&\quad = \frac{0.99\times 0.0001}{0.99\times 0.0001+0.01\times 0.9999}\\
&\quad = .009804
\end{aligned}$$

The moral is that when the disease is much rarer than the test accuracy, a positive test result does not mean the disease is likely. A false positive reading remains much more likely.

Here is an alternative exercise along the same lines: A doctor says that an infant who predominantly turns the head to the right while lying on the back will be right-handed, and one who turns to the left will be left-handed. Isabella predominantly turned her head to the left. Given that 90% of the population is right-handed, what is Isabella's probability of being right-handed if the test is 90% accurate? If it is 80% accurate?

The reasoning is the same, and the answer is 50% right-handed if the test is 90% accurate, 69% right-handed if the test is 80% accurate.

Exercise 12.CONB

It is quite often useful to consider the effect of some specific propositions in the context of some general background evidence that remains fixed, rather than in the complete absence of information. The following questions ask you to prove more general versions of the product rule and Bayes' rule, with respect to some background evidence **e**:

a. Prove the conditionalized version of the general product rule:

$$\mathbf{P}(X, Y \mid \mathbf{e}) = \mathbf{P}(X \mid Y, \mathbf{e})\mathbf{P}(Y \mid \mathbf{e})\ .$$

b. Prove the conditionalized version of Bayes' rule in Equation (12.13).

The basic axiom to use here is the definition of conditional probability:

a. We have

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

and

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

hence

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

b. The derivation here is the same as the derivation of the simple version of Bayes' Rule on page 444. First we write down the dual form of the conditionalized product rule, simply by switching A and B in the above derivation:

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

Therefore the two right-hand sides are equal:

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

Dividing through by $\mathbf{P}(B|E)$ we get

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

Exercise 12.PXYZ

Show that the statement of conditional independence

$$\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\mathbf{P}(Y \mid Z)$$

is equivalent to each of the statements

$$\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z) \quad \text{and} \quad \mathbf{P}(B \mid X, Z) = \mathbf{P}(Y \mid Z)\,.$$

The key to this exercise is rigorous and frequent application of the definition of conditional probability, $\mathbf{P}(X|Y) = \mathbf{P}(X, Y)/\mathbf{P}(Y)$. The original statement that we are given is:

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

We start by applying the definition of conditional probability to two of the terms in this statement:

$$\mathbf{P}(A, B|C) = \frac{\mathbf{P}(A, B, C)}{\mathbf{P}(C)} \text{ and } \mathbf{P}(B|C) = \frac{\mathbf{P}(B, C)}{\mathbf{P}(C)}$$

Now we substitute the right hand side of these definitions for the left hand sides in the original statement to get:

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

Now we need the definition once more:

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

We substitute this right hand side for $\mathbf{P}(A, B, C)$ to get:

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

Finally, we cancel the $\mathbf{P}(B, C)$ and $\mathbf{P}(C)$s to get:

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

The second part of the exercise follows from by a similar derivation, or by noticing that A and B are interchangeable in the original statement (because multiplication is commutative and A, B means the same as B, A).

In Chapter 13, we will see that in terms of Bayesian networks, the original statement means that C is the lone parent of A and also the lone parent of B. The conclusion is that knowing the values of B and C is the same as knowing just the value of C in terms of telling you something about the value of A.

Exercise 12.XXXX

You have three coins in your pocket:

- Coin 1 is a fair coin that comes up heads with probability $1/2$.
- Coin 2 is a biased coin that comes up heads with probability $1/4$.
- Coin 3 is a biased coin that comes up heads with probability $3/4$.

Suppose you pick one of the coins uniformly at random and flip it three times. If you observe the sequence HHT (where H stands for heads and T stands for tails), what is the probability that you chose Coin 3?

Let C_i denote the event that coin i was chosen. Using Bayes' rule, we have that

$$P(C_3 \,|\, HHT) = \frac{P(HHT \,|\, C_3)P(C_3)}{P(HHT)} = \frac{P(HHT \,|\, C_3)P(C_3)}{\sum_{i=1}^{3} P(HHT \,|\, C_i)P(C_i)} = \frac{P(HHT \,|\, C_3)}{\sum_{i=1}^{3} P(HHT \,|\, C_i)},$$

where the final equality follows from the fact that $P(C_i) = 1/3$ for each i. Substituting in the values from the problem, we obtain

$$P(C_3 \,|\, HHT) = \frac{\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}+\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}} = \frac{\frac{9}{64}}{\frac{1}{8}+\frac{3}{64}+\frac{9}{64}} = \frac{9}{20}.$$

Exercise 12.XXXX

Suppose you are given a bag containing n unbiased coins. You are told that $n-1$ of these coins are normal, with heads on one side and tails on the other, whereas one coin is a fake, with heads on both sides.

a. Suppose you reach into the bag, pick out a coin at random, flip it, and get a head. What is the (conditional) probability that the coin you chose is the fake coin?

b. Suppose you continue flipping the coin for a total of k times after picking it and see k heads. Now what is the conditional probability that you picked the fake coin?

c. Suppose you wanted to decide whether the chosen coin was fake by flipping it k times. The decision procedure returns *fake* if all k flips come up heads; otherwise it returns $normal$. What is the (unconditional) probability that this procedure makes an error?

a. A typical "counting" argument goes like this: There are n ways to pick a coin, and 2 outcomes for each flip (although with the fake coin, the results of the flip are indistinguishable), so there are $2n$ total atomic events, each equally likely. Of those, only 2 pick the fake coin, and $2 + (n - 1)$ result in heads. So the probability of a fake coin given heads, $P(\mathit{fake}|\mathit{heads})$, is $2/(2 + n - 1) = 2/(n + 1)$.

Often such counting arguments go astray when the situation gets complex. It may be better to do it more formally:

$$\begin{aligned}\mathbf{P}(Fake|heads) &= \alpha\mathbf{P}(heads|Fake)\mathbf{P}(Fake)\\ &= \alpha\langle 1.0, 0.5\rangle\langle 1/n, (n-1)/n\rangle\\ &= \alpha\langle 1/n, (n-1)/2n\rangle\\ &= \langle 2/(n+1), (n-1)/(n+1)\rangle\end{aligned}$$

b. Now there are $2^k n$ atomic events, of which 2^k pick the fake coin, and $2^k + (n-1)$ result in heads. So the probability of a fake coin given a run of k heads, $P(\mathit{fake}|\mathit{heads}^k)$, is $2^k/(2^k + (n-1))$. Note this approaches 1 as k increases, as expected. If $k = n = 12$, for example, than $P(\mathit{fake}|\mathit{heads}^{10}) = 0.9973$.

Doing it the formal way:

$$\begin{aligned}\mathbf{P}(Fake|heads^k) &= \alpha\mathbf{P}(heads^k|Fake)\mathbf{P}(Fake)\\ &= \alpha\langle 1.0, 0.5^k\rangle\langle 1/n, (n-1)/n\rangle\\ &= \alpha\langle 1/n, (n-1)/2^k n\rangle\\ &= \langle 2^k/(2^k+n-1), (n-1)/(2^k+n-1)\rangle\end{aligned}$$

c. The procedure makes an error if and only if a fair coin is chosen and turns up heads k times in a row. The probability of this

$$P(heads^k|\neg fake)P(\neg fake) = (n-1)/2^k n\,.$$

Exercise 12.NOEM

In this exercise, you will complete the normalization calculation for the meningitis example. First, make up a suitable value for $P(s\,|\,\neg m)$, and use it to calculate unnormalized values for $P(m\,|\,s)$ and $P(\neg m\,|\,s)$ (i.e., ignoring the $P(s)$ term in the Bayes' rule expression, Equation (12.14)). Now normalize these values so that they add to 1.

The important point here is that although there are often many possible routes by which answers can be calculated in such problems, it is usually better to stick to systematic "standard" routes such as Bayes' Rule plus normalization. Chapter 14 describes general-purpose, systematic algorithms that make heavy use of normalization. We could guess that $P(S|\neg M) \approx 0.05$, or we could calculate it from the information already given (although the idea here is to assume that $P(S)$ is *not* known):

$$P(S|\neg M) = \frac{P(\neg M|S)P(S)}{P(\neg M)} = \frac{(1 - P(M|S))P(S)}{1 - P(\neg M)} = \frac{0.9998 \times 0.05}{0.99998} = 0.049991$$

Normalization proceeds as follows:

$$\begin{array}{l}
P(M|S) \propto P(S|M)P(M) = 0.5/50,000 = 0.00001 \\
P(\neg M|S) \propto P(S|\neg M)P(\neg M) = 0.049991 \times 0.99998 = 0.04999 \\
P(M|S) = \frac{0.00001}{0.00001+0.04999} = 0.0002 \\
P(\neg M|S) = \frac{0.00001}{0.00001+0.04999} = 0.9998
\end{array}$$

Exercise 12.XXXX

This exercise investigates the way in which conditional independence relationships affect the amount of information needed for probabilistic calculations.

a. Suppose we wish to calculate $P(h \,|\, e_1, e_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation?

(i) $\mathbf{P}(E_1, E_2), \mathbf{P}(H), \mathbf{P}(E_1 \,|\, H), \mathbf{P}(E_2 \,|\, H)$

(ii) $\mathbf{P}(E_1, E_2), \mathbf{P}(H), \mathbf{P}(E_1, E_2 \,|\, H)$

(iii) $\mathbf{P}(H), \mathbf{P}(E_1 \,|\, H), \mathbf{P}(E_2 \,|\, H)$

b. Suppose we know that $\mathbf{P}(E_1 \,|\, H, E_2) = \mathbf{P}(E_1 \,|\, H)$ for all values of H, E_1, E_2. Now which of the three sets are sufficient?

The question would have been slightly more consistent if we had asked about the calculation of $\mathbf{P}(H|E_1, E_2)$ instead of $P(H|E_1, E_2)$. Showing that a given set of information is *sufficient* is relatively easy: find an expression for $\mathbf{P}(H|E_1, E_2)$ in terms of the given information. Showing *insufficiency* can be done by showing that the information provided does not contain enough independent numbers.

a. Bayes' Rule gives

$$\mathbf{P}(H|E_1, E_2) = \frac{\mathbf{P}(E_1, E_2|H)\mathbf{P}(H)}{\mathbf{P}(E_1, E_2)}$$

Hence the information in (ii) is sufficient—in fact, we don't need $\mathbf{P}(E_1, E_2)$ because we can use normalization. Intuitively, the information in (iii) is insufficient because $\mathbf{P}(E_1|H)$ and $\mathbf{P}(E_2|H)$ provide no information about correlations between E_1 and E_2 that might be induced by H. Mathematically, suppose H has m possible values and E_1 and E_2 have n_1 and n_2 possible respectively. $\mathbf{P}(H|E_1, E_2)$ contains $(m - 1)n_1n_2$

independent numbers, whereas the information in (iii) contains $(m-1)+m(n_1-1)+m(n_2-1)$ numbers—clearly insufficient for large m, n_1, and n_2. Similarly, the information in (i) contains $(n_1 n_2 - 1) + m + m(n_1-1) + m(n_2-1)$ numbers—again insufficient.

b. If E_1 and E_2 are conditionally independent given H, then

$$\mathbf{P}(E_1, E_2|H) = \mathbf{P}(E_1|H)\mathbf{P}(E_2|H).$$

Using normalization, (i), (ii), and (iii) are each sufficient for the calculation.

Exercise 12.BLRV

Let X, Y, Z be Boolean random variables. Label the eight entries in the joint distribution $\mathbf{P}(X, Y, Z)$ as a through h. Express the statement that X and Y are conditionally independent given Z, as a set of equations relating a through h. How many *nonredundant* equations are there?

Let the probabilities be as follows:

x	y	z	$P(x, y, z)$
F	F	F	a
F	F	T	b
F	T	F	c
F	T	T	d
T	F	F	e
T	F	T	f
T	T	F	g
T	T	T	h

Conditional independence asserts that

$$\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z)\mathbf{P}(Y \mid Z)$$

which we can rewrite in terms of the joint distribution using the definition of conditional probability and marginals:

$$\begin{aligned}
\frac{\mathbf{P}(X, Y, Z)}{\mathbf{P}(Z)} &= \frac{\mathbf{P}(X, Z)}{\mathbf{P}(Z)} \cdot \frac{\mathbf{P}(Y, Z)}{\mathbf{P}(Z)} \\
\mathbf{P}(X, Y, Z) &= \frac{\mathbf{P}(X, Z)\mathbf{P}(Y, Z)}{\mathbf{P}(Z)} = \frac{\left(\sum_y \mathbf{P}(X, y, Z)\right)\left(\sum_x \mathbf{P}(x, Y, Z)\right)}{\sum_{x,y} \mathbf{P}(x, y, Z)} .
\end{aligned}$$

Now we instantiate X, Y, Z in all 8 ways to obtain the following 8 equations:

$$\begin{aligned}
a &= (a+c)(a+e)/(a+c+e+g) \text{ or } ag = ce \\
b &= (b+d)(b+f)/(b+d+f+h) \text{ or } bh = df \\
c &= (a+c)(c+g)/(a+c+e+g) \text{ or } ce = ag \\
d &= (b+d)(d+h)/(b+d+f+h) \text{ or } df = bh \\
e &= (e+g)(a+e)/(a+c+e+g) \text{ or } ce = ag \\
f &= (f+h)(b+f)/(b+d+f+h) \text{ or } df = bh \\
g &= (e+g)(c+g)/(a+c+e+g) \text{ or } ag = ce \\
h &= (f+h)(d+h)/(b+d+f+h) \text{ or } bh = df \, .
\end{aligned}$$

Thus, there are only 2 nonredundant equations, $ag = ce$ and $bh = df$. This is what we would expect: the general distribution requires $8-1=7$ parameters, whereas the Bayes net with Z as root and X and Y as conditionally indepednent children requires 1 parameter for Z and 2 each for X and Y, or 5 in all. Hence the conditional independence assertion removes two degrees of freedom.

Exercise 12.XXXX

Suppose you are a witness to a nighttime hit-and-run accident involving a taxi in Athens. All taxis in Athens are blue or green. You swear, under oath, that the taxi was blue. Extensive testing shows that, under the dim lighting conditions, discrimination between blue and green is 75% reliable. (Adapted from Pearl (1988).)

a. Is it possible to calculate the most likely color for the taxi? (*Hint:* distinguish carefully between the proposition that the taxi *is* blue and the proposition that it *appears* blue.)

b. What if you know that 9 out of 10 Athenian taxis are green?

The relevant aspect of the world can be described by two random variables: B means the taxi *was* blue, and LB means the taxi *looked blue*. The information on the reliability of color identification can be written as

$$P(LB|B) = 0.75 \qquad P(\neg LB|\neg B) = 0.75$$

We need to know the probability that the taxi was blue, given that it looked blue:

$$\begin{aligned}
&P(B|LB) \propto P(LB|B)P(B) \propto 0.75P(B) \\
&P(\neg B|LB) \propto P(LB|\neg B)P(\neg B) \propto 0.25(1-P(B))
\end{aligned}$$

Thus we cannot decide the probability without some information about the prior probability of blue taxis, $P(B)$. For example, if we knew that all taxis were blue, i.e., $P(B) = 1$, then obviously $P(B|LB) = 1$. On the other hand, if we adopt Laplace's *Principle of Indifference*, which states that propositions can be deemed equally likely in the absence of any differentiating information, then we have $P(B) = 0.5$ and $P(B|LB) = 0.75$. Usually we will have *some* differentiating information, so this principle does not apply.

Given that 9 out of 10 taxis are green, and *assuming the taxi in question is drawn ran-*

domly from the taxi population, we have $P(B) = 0.1$. Hence

$$\begin{array}{l} P(B|LB) \propto 0.75 \times 0.1 \propto 0.075 \\ P(\neg B|LB) \propto 0.25 \times 0.9 \propto 0.225 \\ P(B|LB) = \frac{0.075}{0.075+0.225} = 0.25 \\ P(\neg B|LB) = \frac{0.225}{0.075+0.225} = 0.75 \end{array}$$

12.6 Naive Bayes Models

Exercise 12.XXXX

Consider a model with one binary class variable Y and n k-ary features $F_1, F_2, \ldots, F_n$.

a. How many parameters are required to represent the full joint distribution $\mathbf{P}(Y, F_1, F_2, \ldots, F_n)$ in tabular form, given no conditional independence properties?

b. How many are needed when the naive Bayes model is applicable?

a. The table size is $|Y| \cdot |F_i|^n = 2k^n$, minus 1 for the sum-to-1 constraint, so $2k^n - 1$ parameters are required.

b. In the naive Bayes model, we need 1 for the prior on Y and, for each F_i, $k-1$ parameters for each of the 2 values of Y, so $2n(k-1) + 1$ parameters are needed.

Exercise 12.XXXX

Write out a general algorithm for answering queries of the form $\mathbf{P}(Cause \mid \mathbf{e})$, using a naive Bayes distribution. Assume that the evidence $\mathbf{e}$ may assign values to *any subset* of the effect variables.

We can apply the definition of conditional independence as follows:

$$\mathbf{P}(Cause|\mathbf{e}) = \mathbf{P}(\mathbf{e}, Cause)/\mathbf{P}(\mathbf{e}) = \alpha \mathbf{P}(\mathbf{e}, Cause) \, .$$

Now, divide the effect variables into those with evidence, $\mathbf{E}$, and those without evidence, $\mathbf{Y}$.

We have

$$\begin{aligned}
\mathbf{P}(Cause|\mathbf{e}) &= \alpha \sum_{\mathbf{y}} \mathbf{P}(\mathbf{y}, \mathbf{e}, Cause) \\
&= \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause)\mathbf{P}(\mathbf{y}|Cause) \left(\prod_j \mathbf{P}(e_j|Cause) \right) \\
&= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j|Cause) \right) \sum_{\mathbf{y}} \mathbf{P}(Cause)\mathbf{P}(\mathbf{y}|Cause) \\
&= \alpha \mathbf{P}(Cause) \left(\prod_j \mathbf{P}(e_j|Cause) \right)
\end{aligned}$$

where the last line follows because the summation over $\mathbf{y}$ is 1. Therefore, the algorithm computes the product of the conditional probabilities of the evidence variables given each value of the cause, multiplies each by the prior probability of the cause, and normalizes the result.

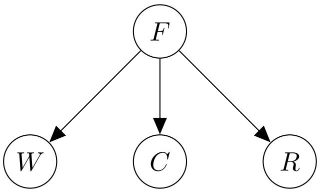

Figure S12.1 A naive Bayes model for fishing.

Exercise 12.FISH

A non-angler, curious to know what counts as a good day of fishing (F) at the lake and puzzled by the phenomenon that it can sometimes be a good day even when no fish are caught, decides to create a naive Bayes model with F as the Boolean class variable and three features: whether it rained (R), how many fish were caught (C) with values $\{none, some, lots\}$, and whether it was windy (W). A naive Bayes net is shown in Figure S12.1.

a. Here are some plausible probability tables estimated from (entirely made up) data:

$P(F=false)$	$P(F=true)$
0.9	0.1

F	$P(W=false \mid F)$	$P(W=true \mid F)$
false	0.5	0.5
true	0.8	0.2

F	$P(C=none \mid F)$	$P(C=some \mid F)$	$P(C=lots \mid F)$
false	0.7	0.2	0.1
true	0.3	0.4	0.3

F	$P(R = \mathit{false} \mid F)$	$P(R = \mathit{true} \mid F)$
false	0.6	0.4
true	0.9	0.1

Given this model, calculate the following probabilities:

(i) $P(F = \mathit{true} \mid R = \mathit{true}, C = \mathit{none}, W = \mathit{true})$.

(ii) $P(F = \mathit{true} \mid R = \mathit{false}, C = \mathit{lots}, W = \mathit{false})$.

(iii) $P(F = \mathit{true} \mid R = \mathit{true}, C = \mathit{some}, W = \mathit{false})$.

b. Derive symbolic expressions for the following probabilities in terms of $\mathbf{P}(R \mid C)$, $\mathbf{P}(C \mid F)$, $\mathbf{P}(W \mid F)$, and $\mathbf{P}(F)$:

(i) $\mathbf{P}(R)$.

(ii) $\mathbf{P}(R \mid C, W)$.

(iii) $\mathbf{P}(R, C, W \mid F)$.

c. Comment on the plausibility of the conditional independence assumptions made by the naive Bayes model.

a. (i)

$$\begin{aligned}
& P(F = \mathit{true} \mid R = \mathit{true}, C = \mathit{none}, W = \mathit{true}) \\
= & \frac{P(R = \mathit{true}, C = \mathit{none}, W = \mathit{true} \mid F = \mathit{true})\ P(F = \mathit{True})}{P(R = \mathit{true}, C = \mathit{none}, W = \mathit{true})} \\
= & \frac{P(R = \mathit{true} \mid F = \mathit{true})\ P(C = \mathit{none} \mid F = \mathit{true})\ P(W = \mathit{true} \mid F = \mathit{true})\ P(F = \mathit{True})}{P(R = \mathit{true})\ P(C = \mathit{none})\ P(W = \mathit{true})} \\
= & \frac{.1 \times .3 \times .2 \times .1}{.1 \times .3 \times .2 \times .1 + .4 \times .7 \times .5 \times .9} \\
\approx & .005\,.
\end{aligned}$$

(ii)

$$\begin{aligned}
& P(F = \mathit{true} \mid R = \mathit{false}, C = \mathit{lots}, W = \mathit{false}) \\
= & \frac{P(R = \mathit{false} \mid F = \mathit{true})\ P(C = \mathit{lots} \mid F = \mathit{true})\ P(W = \mathit{false} \mid F = \mathit{true})\ P(F = \mathit{True})}{P(R = \mathit{false})\ P(C = \mathit{lots})\ P(W = \mathit{false})} \\
= & \frac{.9 \times .3 \times .8 \times .1}{.9 \times .3 \times .8 \times .1 + .6 \times .1 \times .5 \times .9} \\
\approx & .4\,.
\end{aligned}$$

(iii)

$$\begin{aligned}
& P(F = \mathit{true} \mid R = \mathit{true}, C = \mathit{some}, W = \mathit{false}) \\
= & \frac{P(R = \mathit{true} \mid F = \mathit{true})\ P(C = \mathit{some} \mid F = \mathit{true})\ P(W = \mathit{false} \mid F = \mathit{true})\ P(F = \mathit{True})}{P(R = \mathit{true})\ P(C = \mathit{some})\ P(W = \mathit{false})} \\
= & \frac{.1 \times .4 \times .8 \times .1}{.1 \times .4 \times .8 \times .1 + .9 \times .2 \times .5 \times .9} \\
\approx & .04\,.
\end{aligned}$$

b. (i) $\mathbf{P}(R) = \sum_f \mathbf{P}(R \mid f)P(f)$.
(ii) $\frac{\sum_f P(f)P(R \mid f)P(C \mid f)P(W \mid f)}{\sum_f P(f)P(W \mid f)P(C \mid f)}$.
(iii) $\mathbf{P}(R \mid F)\mathbf{P}(C \mid F)\mathbf{P}(W \mid F)$.

c. The assumptions of the model, that W, C, R are independent given F, are unrealistic; wind and rain are not independent and might each reduce the likelihood of catching fish (for only the seasoned anglers would be likely to stay out in such conditions). Also, having a good day fishing is an effect and not a cause of whether it is windy, one catches any fish, and whether it is rainy.

Exercise 12.XXXX

You are given points from 2 classes, shown as rectangles and dots in Figure S12.2. For each of the following sets of points, decide whether the set satisfies or fails to satisfy all the Naïve Bayes modelling assumptions. Note that in (c), 4 rectangles overlap with 4 dots.

The conditional independence assumptions made by the Naïve Bayes model are that features are conditionally independent when given the class. Features being independent once the class label is known means that for a fixed class the distribution for f_1 cannot depend on f_2, and the other way around. Concretely, for discrete-valued features as shown below, this means each class needs to have a distribution that corresponds to an axis-aligned rectangle. No other assumption is made by the Naïve Bayes model. Note that linear separability is not an assumption of the Naïve Bayes model—what is true is that for a Naïve Bayes model with all binary variables the decision boundary between the two classes is a hyperplane (i.e., it's a linear classifier). That, however, wasn't relevant to the question as the question examined which probability distribution a Naïve Bayes model can represent, not which decision boundaries.

A note about feature independence: The Naïve Bayes model assumes features are conditionally independent given the class. Why does this result in axis-aligned rectangles for discrete feature distributions? Intuitively, this is because fixing one value is uninformative about the other: within a class, the values of one feature are constant across the other. For instance, the dark square class in (b) has $f_1 \in [-0.5, 0.5]$ and $f_2 \in [-1, 0]$ and fixing one has no impact on the domain of the other. However, when the features of a class are not axis-aligned then fixing one limits the domain of the other, inducing dependence. In (e), fixing $f_2 = 1.5$ restricts f_1 to the two points at the top, whereas fixing $f_2 = 0$ gives a larger domain.

Exercise 12.BAYS

Text categorization is the task of assigning a given document to one of a fixed set of categories on the basis of the text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence or absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by the document category.

a. Explain precisely how such a model can be constructed, given as "training data" a set

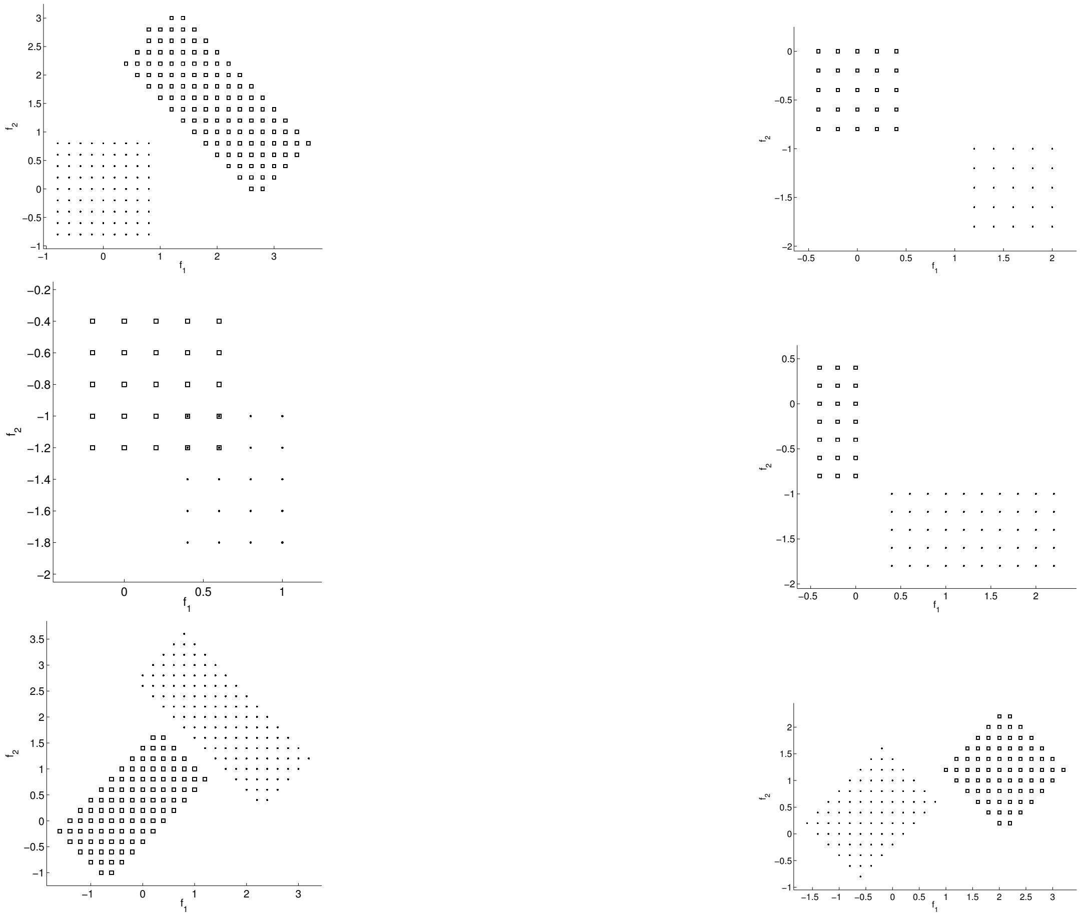

Figure S12.2 Six data sets that may or may not satisfy the naive Bayes assumption.

of documents that have been assigned to categories.

b. Explain precisely how to categorize a new document.

c. Is the conditional independence assumption reasonable? Discuss.

This question is essentially previewing material in Chapter 24 (page 860), but students should have little difficulty in figuring out how to estimate a conditional probability from complete data.

a. The model consists of the prior probability $\mathbf{P}(Category)$ and the conditional probabilities $\mathbf{P}(Word_i|Category)$. For each category c, $\mathbf{P}(Category = c)$ is estimated as the fraction of all documents that are of category c. Similarly, $\mathbf{P}(Word_i = true|Category = c)$ is estimated as the fraction of documents of category c that contain word i.

b. See the answer for 13.17. Here, every evidence variable is observed, since we can tell if any given word appears in a given document or not.

c. The independence assumption is clearly violated in practice. For example, the word pair "artificial intelligence" occurs more frequently in any given document category than would be suggested by multiplying the probabilities of "artificial" and "intelligence".

12.7 The Wumpus World Revisited

Exercise 12.PITP

In our analysis of the wumpus world, we used the fact that each square contains a pit with probability 0.2, independently of the contents of the other squares. Suppose instead that exactly $N/5$ pits are scattered at random among the N squares other than [1,1]. Are the variables $P_{i,j}$ and $P_{k,l}$ still independent? What is the joint distribution $\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$ now? Redo the calculation for the probabilities of pits in [1,3] and [2,2].

This probability model is also appropriate for Minesweeper (Ex. 7.11). If the total number of pits is fixed, then the variables $P_{i,j}$ and $P_{k,l}$ are no longer independent. In general,

$$P(P_{i,j} = true | P_{k,l} = true) < P(P_{i,j} = true | P_{k,l} = false)$$

because learning that $P_{k,l} = true$ makes it less likely that there is a mine at $[i, j]$ (as there are now fewer to spread around). The joint distribution places equal probability on all assignments to $P_{1,2} \ldots P_{4,4}$ that have exactly 3 pits, and zero on all other assignments. Since there are 15 squares, the probability of each 3-pit assignment is $1/\binom{15}{3} = 1/455$.

To calculate the probabilities of pits in $[1, 3]$ and $[2, 2]$, we start from Figure 13.7. We have to consider the probabilities of complete assignments, since the probability of the "other" region assignment does not cancel out. We can count the total number of 3-pit assignments that are consistent with each partial assignment in 13.7(a) and 13.7(b).

In 13.7(a), there are three partial assignments with $P_{1,3} = true$:

- The first fixes all three pits, so corresponds to 1 complete assignment.
- The second leaves 1 pit in the remaining 10 squares, so corresponds to 10 complete assignments.
- The third also corresponds to 10 complete assignments.

Hence, there are 21 complete assignments with $P_{1,3} = true$.

In 13.7(b), there are two partial assignments with $P_{1,3} = false$:

- The first leaves 1 pit in the remaining 10 squares, so corresponds to 10 complete assignments.
- The second leaves 2 pits in the remaining 10 squares, so corresponds to $\binom{10}{2} = 45$ complete assignments.

Hence, there are 55 complete assignments with $P_{1,3} = false$. Normalizing, we obtain

$$\mathbf{P}(P_{1,3}) = \alpha\langle 21, 55\rangle = \langle 0.276, 0.724\rangle \ .$$

With $P_{2,2} = true$, there are four partial assignments with a total of $\binom{10}{2} + 2 \cdot \binom{10}{1} + \binom{10}{0} = 66$ complete assignments. With $P_{2,2} = false$, there is only one partial assignment with $\binom{10}{1} = 10$ complete assignments. Hence

$$\mathbf{P}(P_{2,2}) = \alpha\langle 66, 10\rangle = \langle 0.868, 0.132\rangle \ .$$

Exercise 12.PITQ

Redo the probability calculation for pits in [1,3] and [2,2], assuming that each square contains a pit with probability 0.01, independent of the other squares. What can you say about the relative performance of a logical versus a probabilistic agent in this case?

First we redo the calculations of $P(\mathit{frontier})$ for each model in Figure 12.6. The three models with $P_{1,3} = \mathit{true}$ have probabilities 0.0001, 0.0099, 0.0099; the two models with $P_{1,3} = \mathit{false}$ have probabilities 0.0001, 0.0099. Then

$$\begin{aligned}\mathbf{P}(P_{1,3} \mid \mathit{known}, b) &= \alpha' \langle 0.01(0.0001 + 0.0099 + 0.0099),\ 0.99(0.0001 + 0.0099)\rangle \\ &\approx \langle 0.1674, 0.8326\rangle \ .\end{aligned}$$

The four models with $P_{2,2} = \mathit{true}$ have probabilities 0.0001, 0.0099, 0.0099, 0.9801; the one model with $P_{2,2} = \mathit{false}$ has probability 0.0001. Then

$$\begin{aligned}\mathbf{P}(P_{2,2} \mid \mathit{known}, b) &= \alpha' \langle 0.01(0.0001 + 0.0099 + 0.0099 + 0.9801),\ 0.99 \times 0.0001\rangle \\ &\approx \langle 0.9902, 0.0098\rangle \ .\end{aligned}$$

This means that [2,2] is almost certain death; a probabilistic agent can figure this out and choose [1,3] or [3,1] instead. Its chance of death at this stage will be 0.1674, while a logical agent choosing at random among the three squares will die with probability $(0.1674 + 0.9902 + 0.1674)/3 = 0.4416$. The reason that [2,2] is so much more likely to be a pit in this case is that, for it *not* to be a pit, *both* of [1,3] and [3,1] must contain pits, which is very unlikely. Indeed, as the prior probability of pits tends to 0, the posterior probability of [2,2] tends to 1.

Exercise 12.WUMA

Implement a hybrid probabilistic agent for the wumpus world, based on the hybrid agent in Figure 7.20 and the probabilistic inference procedure outlined in this chapter.

The solution for this exercise is omitted. The main modification to the agent in Figure 7.20 is to calculate, after each move, the safety probability for each square that is not provably safe or fatal, and choose the safest if there is no unvisited safe square.