31. Find the probability that a family with five children does not have a boy, if the sexes of children are independent and if

   a) a boy and a girl are equally likely.

   b) the probability of a boy is 0.51.

   c) the probability that the ith child is a boy is $0.51 - (i/100)$.

32. Find the probability that a randomly generated bit string of length 10 begins with a 1 or ends with a 00 for the same conditions as in parts (a), (b), and (c) of Exercise 30, if bits are generated independently.

33. Find the probability that the first child of a family with five children is a boy or that the last two children of the family are girls, for the same conditions as in parts (a), (b), and (c) of Exercise 31.

34. Find each of the following probabilities when n independent Bernoulli trials are carried out with probability of success p.

   a) the probability of no successes

   b) the probability of at least one success

   c) the probability of at most one success

   d) the probability of at least two successes

35. Find each of the following probabilities when n independent Bernoulli trials are carried out with probability of success p.

   a) the probability of no failures

   b) the probability of at least one failure

   c) the probability of at most one failure

   d) the probability of at least two failures

36. Use mathematical induction to prove that if $E_1, E_2, \ldots, E_n$ is a sequence of n pairwise disjoint events in a sample space S, where n is a positive integer, then $p(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} p(E_i)$.

∗37. (*Requires calculus*) Show that if $E_1, E_2, \ldots$ is an infinite sequence of pairwise disjoint events in a sample space S, then $p(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} p(E_i)$. [*Hint:* Use Exercise 36 and take limits.]

38. A pair of dice is rolled in a remote location and when you ask an honest observer whether at least one die came up six, this honest observer answers in the affirmative.

   a) What is the probability that the sum of the numbers that came up on the two dice is seven, given the information provided by the honest observer?

   b) Suppose that the honest observer tells us that at least one die came up five. What is the probability the sum of the numbers that came up on the dice is seven, given this information?

∗∗39. This exercise employs the probabilistic method to prove a result about round-robin tournaments. In a **round-robin tournament** with m players, every two players play one game in which one player wins and the other loses.

   We want to find conditions on positive integers m and k with $k < m$ such that it is possible for the outcomes of the tournament to have the property that for every set of k players, there is a player who beats every member in this set. So that we can use probabilistic reasoning to draw conclusions about round-robin tournaments, we assume that when two players compete it is equally likely that either player wins the game and we assume that the outcomes of different games are independent. Let E be the event that for every set S with k players, where k is a positive integer less than m, there is a player who has beaten all k players in S.

   a) Show that $p(\overline{E}) \leq \sum_{j=1}^{\binom{m}{k}} p(F_j)$, where F_j is the event that there is no player who beats all k players from the jth set in a list of the $\binom{m}{k}$ sets of k players.

   b) Show that the probability of F_j is $(1 - 2^{-k})^{m-k}$.

   c) Conclude from parts (a) and (b) that $p(\overline{E}) \leq \binom{m}{k}(1 - 2^{-k})^{m-k}$ and, therefore, that there must be a tournament with the described property if $\binom{m}{k}(1 - 2^{-k})^{m-k} < 1$.

   d) Use part (c) to find values of m such that there is a tournament with m players such that for every set S of two players, there is a player who has beaten both players in S. Repeat for sets of three players.

∗40. Devise a Monte Carlo algorithm that determines whether a permutation of the integers 1 through n has already been sorted (that is, it is in increasing order), or instead, is a random permutation. A step of the algorithm should answer "true" if it determines the list is not sorted and "unknown" otherwise. After k steps, the algorithm decides that the integers are sorted if the answer is "unknown" in each step. Show that as the number of steps increases, the probability that the algorithm produces an incorrect answer is extremely small. [*Hint:* For each step, test whether certain elements are in the correct order. Make sure these tests are independent.]

41. Use pseudocode to write out the probabilistic primality test described in Example 16.

7.3   Bayes' Theorem

Introduction

There are many times when we want to assess the probability that a particular event occurs on the basis of partial evidence. For example, suppose we know the percentage of people who have a particular disease for which there is a very accurate diagnostic test. People who test positive for

this disease would like to know the likelihood that they actually have the disease. In this section we introduce a result that can be used to determine this probability, namely, the probability that a person has the disease given that this person tests positive for it. To use this result, we will need to know the percentage of people who do not have the disease but test positive for it and the percentage of people who have the disease but test negative for it.

Similarly, suppose we know the percentage of incoming e-mail messages that are spam. We will see that we can determine the likelihood that an incoming e-mail message is spam using the occurrence of words in the message. To determine this likelihood, we need to know the percentage of incoming messages that are spam, the percentage of spam messages in which each of these words occurs, and the percentage of messages that are not spam in which each of these words occurs.

The result that we can use to answer questions such as these is called Bayes' theorem and dates back to the eighteenth century. In the past two decades, Bayes' theorem has been extensively applied to estimate probabilities based on partial evidence in areas as diverse as medicine, law, machine learning, engineering, and software development.

Bayes' Theorem

We illustrate the idea behind Bayes' theorem with an example that shows that when extra information is available, we can derive a more realistic estimate that a particular event occurs. That is, suppose we know $p(F)$, the probability that an event F occurs, but we have knowledge that an event E occurs. Then the conditional probability that F occurs given that E occurs, $p(F \mid E)$, is a more realistic estimate than $p(F)$ that F occurs. In Example 1 we will see that we can find $p(F \mid E)$ when we know $p(F)$, $p(E \mid F)$, and $p(E \mid \overline{F})$.

EXAMPLE 1

Extra
Examples

We have two boxes. The first contains two green balls and seven red balls; the second contains four green balls and three red balls. Bob selects a ball by first choosing one of the two boxes at random. He then selects one of the balls in this box at random. If Bob has selected a red ball, what is the probability that he selected a ball from the first box?

Solution: Let E be the event that Bob has chosen a red ball; $\overline{E}$ is the event that Bob has chosen a green ball. Let F be the event that Bob has chosen a ball from the first box; $\overline{F}$ is the event that Bob has chosen a ball from the second box. We want to find $p(F \mid E)$, the probability that the ball Bob selected came from the first box, given that it is red. By the definition of conditional probability, we have $p(F \mid E) = p(F \cap E)/p(E)$. Can we use the information provided to determine both $p(F \cap E)$ and $p(E)$ so that we can find $p(F \mid E)$?

First, note that because the first box contains seven red balls out of a total of nine balls, we know that $p(E \mid F) = 7/9$. Similarly, because the second box contains three red balls out of a total of seven balls, we know that $p(E \mid \overline{F}) = 3/7$. We assumed that Bob selects a box at random, so $p(F) = p(\overline{F}) = 1/2$. Because $p(E \mid F) = p(E \cap F)/p(F)$, it follows that $p(E \cap F) = p(E \mid F)p(F) = \frac{7}{9} \cdot \frac{1}{2} = \frac{7}{18}$ [as we remarked earlier, this is one of the quantities we need to find to determine $p(F \mid E)$]. Similarly, because $p(E \mid \overline{F}) = p(E \cap \overline{F})/p(\overline{F})$, it follows that $p(E \cap \overline{F}) = p(E \mid \overline{F})p(\overline{F}) = \frac{3}{7} \cdot \frac{1}{2} = \frac{3}{14}$.

We can now find $p(E)$. Note that $E = (E \cap F) \cup (E \cap \overline{F})$, where $E \cap F$ and $E \cap \overline{F}$ are disjoint sets. (If x belongs to both $E \cap F$ and $E \cap \overline{F}$, then x belongs to both F and $\overline{F}$, which is impossible.) It follows that

$$p(E) = p(E \cap F) + p(E \cap \overline{F}) = \frac{7}{18} + \frac{3}{14} = \frac{49}{126} + \frac{27}{126} = \frac{76}{126} = \frac{38}{63}.$$

We have now found both $p(F \cap E) = 7/18$ and $p(E) = 38/63$. We conclude that

$$p(F \mid E) = \frac{p(F \cap E)}{p(E)} = \frac{7/18}{38/63} = \frac{49}{76} \approx 0.645.$$

Before we had any extra information, we assumed that the probability that Bob selected the first box was $1/2$. However, with the extra information that the ball selected at random is red, this probability has increased to approximately 0.645. That is, the probability that Bob selected a ball from the first box increased from $1/2$, when no extra information was available, to 0.645 once we knew that the ball selected was red. ◄

Using the same type of reasoning as in Example 1, we can find the conditional probability that an event F occurs, given that an event E has occurred, when we know $p(E \mid F)$, $p(E \mid \overline{F})$, and $p(F)$. The result we can obtain is called **Bayes' theorem**; it is named after Thomas Bayes, an eighteenth-century British mathematician and minister who introduced this result.

THEOREM 1   **BAYES' THEOREM**   Suppose that E and F are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})}.$$

Proof: The definition of conditional probability tells us that $p(F \mid E) = p(E \cap F)/p(E)$ and $p(E \mid F) = p(E \cap F)/p(F)$. Therefore, $p(E \cap F) = p(F \mid E)p(E)$ and $p(E \cap F) = p(E \mid F)p(F)$. Equating these two expressions for $p(E \cap F)$ shows that

$$p(F \mid E)p(E) = p(E \mid F)p(F).$$

Dividing both sides by $p(E)$, we find that

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E)}.$$

Next, we show that $p(E) = p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})$. To see this, first note that $E = E \cap S = E \cap (F \cup \overline{F}) = (E \cap F) \cup (E \cap \overline{F})$. Furthermore, $E \cap F$ and $E \cap \overline{F}$ are disjoint, because if $x \in E \cap F$ and $x \in E \cap \overline{F}$, then $x \in F \cap \overline{F} = \emptyset$. Consequently, $p(E) = p(E \cap F) + p(E \cap \overline{F})$. We have already shown that $p(E \cap F) = p(E \mid F)p(F)$. Moreover, we have $p(E \mid \overline{F}) = p(E \cap \overline{F})/p(\overline{F})$, which shows that $p(E \cap \overline{F}) = p(E \mid \overline{F})p(\overline{F})$. It now follows that

$$p(E) = p(E \cap F) + p(E \cap \overline{F}) = p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F}).$$

To complete the proof we insert this expression for $p(E)$ into the equation $p(F \mid E) = p(E \mid F)p(F)/p(E)$. We have proved that

Links

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})}.$$

◁

APPLYING BAYES' THEOREM   Bayes' theorem can be used to solve problems that arise in many disciplines. Next, we will discuss an application of Bayes' theorem to medicine. In particular, we will illustrate how Bayes' theorem can be used to assess the probability that someone testing positive for a disease actually has this disease. The results obtained from Bayes' theorem are often somewhat surprising, as Example 2 shows.

EXAMPLE 2   Suppose that one person in 100,000 has a particular rare disease for which there is a fairly accurate diagnostic test. This test is correct 99.0% of the time when given to a person selected at random who has the disease; it is correct 99.5% of the time when given to a person selected at random who does not have the disease. Given this information can we find

(a) the probability that a person who tests positive for the disease has the disease?
(b) the probability that a person who tests negative for the disease does not have the disease?

Should a person who tests positive be very concerned that he or she has the disease?

Solution: (a) Let F be the event that a person selected at random has the disease, and let E be the event that a person selected at random tests positive for the disease. We want to compute $p(F \mid E)$. To use Bayes' theorem to compute $p(F \mid E)$ we need to find $p(E \mid F)$, $p(E \mid \overline{F})$, $p(F)$, and $p(\overline{F})$.

We know that one person in 100,000 has this disease, so $p(F) = 1/100{,}000 = 0.00001$ and $p(\overline{F}) = 1 - 0.00001 = 0.99999$. Because a person who has the disease tests positive 99% of the time, we know that $p(E \mid F) = 0.99$; this is the probability of a true positive, that a person with the disease tests positive. It follows that $p(\overline{E} \mid F) = 1 - p(E \mid F) = 1 - 0.99 = 0.01$; this is the probability of a false negative, that a person who has the disease tests negative.

Furthermore, because a person who does not have the disease tests negative 99.5% of the time, we know that $p(\overline{E} \mid \overline{F}) = 0.995$. This is the probability of a true negative, that a person without the disease tests negative. Finally, we see that $p(E \mid \overline{F}) = 1 - p(\overline{E} \mid \overline{F}) = 1 - 0.995 = 0.005$; this is the probability of a false positive, that a person without the disease tests positive.

The probability that a person who tests positive for the disease actually has the disease is $p(F \mid E)$. By Bayes' theorem, we know that

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})}$$

$$= \frac{(0.99)(0.00001)}{(0.99)(0.00001) + (0.005)(0.99999)} \approx 0.002.$$

(b) The probability that someone who tests negative for the disease does not have the disease is $p(\overline{F} \mid \overline{E})$. By Bayes' theorem, we know that

$$p(\overline{F} \mid \overline{E}) = \frac{p(\overline{E} \mid \overline{F})p(\overline{F})}{p(\overline{E} \mid \overline{F})p(\overline{F}) + p(\overline{E} \mid F)p(F)}$$

$$= \frac{(0.995)(0.99999)}{(0.995)(0.99999) + (0.01)(0.00001)} \approx 0.9999999.$$

Consequently, 99.99999% of the people who test negative really do not have the disease.

In part (a) we showed that only 0.2% of people who test positive for the disease actually have the disease. Because the disease is extremely rare, the number of false positives on the diagnostic test is far greater than the number of true positives, making the percentage of people who test positive who actually have the disease extremely small. People who test positive for the diseases should not be overly concerned that they actually have the disease.   ◄

GENERALIZING BAYES' THEOREM   Note that in the statement of Bayes' theorem, the events F and $\overline{F}$ are mutually exclusive and cover the entire sample space S (that is, $F \cup \overline{F} = S$). We can extend Bayes' theorem to any collection of mutually exclusive events that cover the entire sample space S, in the following way.

THEOREM 2   **GENERALIZED BAYES' THEOREM**   Suppose that E is an event from a sample space S and that $F_1, F_2, \ldots, F_n$ are mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$. Assume that $p(E) \neq 0$ and $p(F_i) \neq 0$ for $i = 1, 2, \ldots, n$. Then

$$p(F_j \mid E) = \frac{p(E \mid F_j)p(F_j)}{\sum_{i=1}^{n} p(E \mid F_i)p(F_i)}.$$

We leave the proof of this generalized version of Bayes' theorem as Exercise 17.

Bayesian Spam Filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**. Because spam threatens to overwhelm electronic mail systems, a tremendous amount of work has been devoted to filtering it out. Some of the first tools developed for eliminating spam were based on Bayes' theorem, such as **Bayesian spam filters**.

A Bayesian spam filter uses information about previously seen e-mail messages to guess whether an incoming e-mail message is spam. Bayesian spam filters look for occurrences of particular words in messages. For a particular word w, the probability that w appears in a spam e-mail message is estimated by determining the number of times w appears in a message from a large set of messages known to be spam and the number of times it appears in a large set of messages known not to be spam. When we examine e-mail messages to determine whether they might be spam, we look at words that might be indicators of spam, such as "offer," "special," or "opportunity," as well as words that might indicate that a message is not spam, such as "mom," "lunch," or "Jan" (where Jan is one of your friends). Unfortunately, spam filters sometimes fail to identify a spam message as spam; this is called a false negative. And they sometimes identify a message that is not spam as spam; this is called a false positive. When testing for spam, it is important to minimize false positives, because filtering out wanted e-mail is much worse than letting some spam through.

Links

The use of the word *spam* for unsolicited e-mail comes from a Monty Python comedy sketch about a cafe where the food product Spam comes with everything regardless of whether customers want it.

Links

THOMAS BAYES (1702–1761)   Thomas Bayes was the son a minister in a religious sect known as the Nonconformists. This sect was considered heretical in eighteenth-century Great Britain. Because of the secrecy of the Nonconformists, little is known of Thomas Bayes' life. When Thomas was young, his family moved to London. Thomas was likely educated privately; Nonconformist children generally did not attend school. In 1719 Bayes entered the University of Edinburgh, where he studied logic and theology. He was ordained as a Nonconformist minister like his father and began his work as a minister assisting his father. In 1733 he became minister of the Presbyterian Chapel in Tunbridge Wells, southeast of London, where he remained minister until 1752.

Bayes is best known for his essay on probability published in 1764, three years after his death. This essay was sent to the Royal Society by a friend who found it in the papers left behind when Bayes died. In the introduction to this essay, Bayes stated that his goal was to find a method that could measure the probability that an event happens, assuming that we know nothing about it, but that, under the same circumstances, it has happened a certain proportion of times. Bayes' conclusions were accepted by the great French mathematician Laplace but were later challenged by Boole, who questioned them in his book *Laws of Thought*. Since then Bayes' techniques have been subject to controversy.

Bayes also wrote an article that was published posthumously: "An Introduction to the Doctrine of Fluxions, and a Defense of the Mathematicians Against the Objections of the Author of The Analyst," which supported the logical foundations of calculus. Bayes was elected a Fellow of the Royal Society in 1742, with the support of important members of the Society, even though at that time he had no published mathematical works. Bayes' sole known publication during his lifetime was allegedly a mystical book entitled *Divine Benevolence*, discussing the original causation and ultimate purpose of the universe. Although the book is commonly attributed to Bayes, no author's name appeared on the title page, and the entire work is thought to be of dubious provenance. Evidence for Bayes' mathematical talents comes from a notebook that was almost certainly written by Bayes, which contains much mathematical work, including discussions of probability, trigonometry, geometry, solutions of equations, series, and differential calculus. There are also sections on natural philosophy, in which Bayes looks at topics that include electricity, optics, and celestial mechanics. Bayes is also the author of a mathematical publication on asymptotic series, which appeared after his death.

We will develop some basic Bayesian spam filters. First, suppose we have a set B of messages known to be spam and a set G of messages known not to be spam. (For example, users could classify messages as spam when they examine them in their inboxes.) We next identify the words that occur in B and in G. We count the number of messages in the set containing each word to find $n_B(w)$ and $n_G(w)$, the number of messages containing the word w in the sets B and G, respectively. Then, the empirical probability that a spam message contains the word w is $p(w) = n_B(w)/|B|$, and the empirical probability that a message that is not spam contains the word w is $q(w) = n_G(w)/|G|$. We note that $p(w)$ and $q(w)$ estimate the probabilities that an incoming spam message, and an incoming message that is not spam, contain the word w, respectively.

Now suppose we receive a new e-mail message containing the word w. Let S be the event that the message is spam. Let E be the event that the message contains the word w. The events S, that the message is spam, and $\overline{S}$, that the message is not spam, partition the set of all messages. Hence, by Bayes' theorem, the probability that the message is spam, given that it contains the word w, is

$$p(S \mid E) = \frac{p(E \mid S)p(S)}{p(E \mid S)p(S) + p(E \mid \overline{S})p(\overline{S})}.$$

To apply this formula, we first estimate $p(S)$, the probability that an incoming message is spam, as well as $p(\overline{S})$, the probability that the incoming message is not spam. Without prior knowledge about the likelihood that an incoming message is spam, for simplicity we assume that the message is equally likely to be spam as it is not to be spam. That is, we assume that $p(S) = p(\overline{S}) = 1/2$. Using this assumption, we find that the probability that a message is spam, given that it contains the word w, is

$$p(S \mid E) = \frac{p(E \mid S)}{p(E \mid S) + p(E \mid \overline{S})}.$$

(Note that if we have some empirical data about the ratio of spam messages to messages that are not spam, we can change this assumption to produce a better estimate for $p(S)$ and for $p(\overline{S})$; see Exercise 22.)

Next, we estimate $p(E \mid S)$, the conditional probability that the message contains the word w given that the message is spam, by $p(w)$. Similarly, we estimate $p(E \mid \overline{S})$, the conditional probability that the message contains the word w, given that the message is not spam, by $q(w)$. Inserting these estimates for $p(E \mid S)$ and $p(E \mid \overline{S})$ tells us that $p(S \mid E)$ can be estimated by

$$r(w) = \frac{p(w)}{p(w) + q(w)};$$

that is, $r(w)$ estimates the probability that the message is spam, given that it contains the word w. If $r(w)$ is greater than a threshold that we set, such as 0.9, then we classify the message as spam.

EXAMPLE 3   Suppose that we have found that the word "Rolex" occurs in 250 of 2000 messages known to be spam and in 5 of 1000 messages known not to be spam. Estimate the probability that an incoming message containing the word "Rolex" is spam, assuming that it is equally likely that an incoming message is spam or not spam. If our threshold for rejecting a message as spam is 0.9, will we reject such messages?

Solution: We use the counts that the word "Rolex" appears in spam messages and messages that are not spam to find that $p(\text{Rolex}) = 250/2000 = 0.125$ and $q(\text{Rolex}) = 5/1000 = 0.005$.

Because we are assuming that it is equally likely for an incoming message to be spam as it is not to be spam, we can estimate the probability that an incoming message containing the word "Rolex" is spam by

$$r(\text{Rolex}) = \frac{p(\text{Rolex})}{p(\text{Rolex}) + q(\text{Rolex})} = \frac{0.125}{0.125 + 0.005} = \frac{0.125}{0.130} \approx 0.962.$$

Because $r(\text{Rolex})$ is greater than the threshold 0.9, we reject such messages as spam. ◀

Detecting spam based on the presence of a single word can lead to excessive false positives and false negatives. Consequently, spam filters look at the presence of multiple words. For example, suppose that the message contains the words w_1 and w_2. Let E_1 and E_2 denote the events that the message contains the words w_1 and w_2, respectively. To make our computations simpler, we assume that E_1 and E_2 are independent events and that $E_1 \mid S$ and $E_2 \mid S$ are independent events and that we have no prior knowledge regarding whether or not the message is spam. (The assumptions that E_1 and E_2 are independent and that $E_1 \mid S$ and $E_2 \mid S$ are independent may introduce some error into our computations; we assume that this error is small.) Using Bayes' theorem and our assumptions, we can show (see Exercise 23) that $p(S \mid E_1 \cap E_2)$, the probability that the message is spam given that it contains both w_1 and w_2, is

$$p(S \mid E_1 \cap E_2) = \frac{p(E_1 \mid S)p(E_2 \mid S)}{p(E_1 \mid S)p(E_2 \mid S) + p(E_1 \mid \overline{S})p(E_2 \mid \overline{S})}.$$

We estimate the probability $p(S \mid E_1 \cap E_2)$ by

$$r(w_1, w_2) = \frac{p(w_1)p(w_2)}{p(w_1)p(w_2) + q(w_1)q(w_2)}.$$

That is, $r(w_1, w_2)$ estimates the probability that the message is spam, given that it contains the words w_1 and w_2. When $r(w_1, w_2)$ is greater than a preset threshold, such as 0.9, we determine that the message is likely spam.

EXAMPLE 4 Suppose that we train a Bayesian spam filter on a set of 2000 spam messages and 1000 messages that are not spam. The word "stock" appears in 400 spam messages and 60 messages that are not spam, and the word "undervalued" appears in 200 spam messages and 25 messages that are not spam. Estimate the probability that an incoming message containing both the words "stock" and "undervalued" is spam, assuming that we have no prior knowledge about whether it is spam. Will we reject such messages as spam when we set the threshold at 0.9?

Solution: Using the counts of each of these two words in messages known to be spam or known not to be spam, we obtain the following estimates: $p(\text{stock}) = 400/2000 = 0.2$, $q(\text{stock}) = 60/1000 = 0.06$, $p(\text{undervalued}) = 200/2000 = 0.1$, and $q(\text{undervalued}) = 25/1000 = 0.025$. Using these probabilities, we can estimate the probability that the message is spam by

$$r(\text{stock, undervalued}) = \frac{p(\text{stock})p(\text{undervalued})}{p(\text{stock})p(\text{undervalued}) + q(\text{stock})q(\text{undervalued})}$$

$$= \frac{(0.2)(0.1)}{(0.2)(0.1) + (0.06)(0.025)} \approx 0.930.$$

Because we have set the threshold for rejecting messages at 0.9, such messages will be rejected by the filter. ◀

The more words we use to estimate the probability that an incoming mail message is spam, the better is our chance that we correctly determine whether it is spam. In general, if E_i is the

event that the message contains word w_i, assuming that the number of incoming spam messages is approximately the same as the number of incoming messages that are not spam, and that the events $E_i \mid S$ are independent, then by Bayes' theorem the probability that a message containing all the words $w_1, w_2, \ldots, w_k$ is spam is

$$p(S \mid \bigcap_{i=1}^{k} E_i) = \frac{\prod_{i=1}^{k} p(E_i \mid S)}{\prod_{i=1}^{k} p(E_i \mid S) + \prod_{i=1}^{k} p(E_i \mid \overline{S})}.$$

We can estimate this probability by

$$r(w_1, w_2, \ldots, w_k) = \frac{\prod_{i=1}^{k} p(w_i)}{\prod_{i=1}^{k} p(w_i) + \prod_{i=1}^{k} q(w_i)}.$$

Bayesian poisoning, the insertion of extra words to defeat spam filters, can use random or purposefully selected words.

For the most effective spam filter, we choose words for which the probability that each of these words appears in spam is either very high or very low. When we compute this value for a particular message, we reject the message as spam if $r(w_1, w_2, \ldots, w_k)$ exceeds a preset threshold, such as 0.9.

Another way to improve the performance of a Bayesian spam filter is to look at the probabilities that particular pairs of words appear in spam and in messages that are not spam. We then treat appearances of these pairs of words as appearance of a single block, rather than as the appearance of two separate words. For example, the pair of words "enhance performance" most likely indicates spam, while "operatic performance" indicates a message that is not spam. Similarly, we can assess the likelihood that a message is spam by examining the structure of a message to determine where words appear in it. Also, spam filters look at appearances of certain types of strings of characters rather than just words. For example, a message with the valid e-mail address of one of your friends is less likely to be spam (if not sent by a worm) than one containing an e-mail address that came from a country known to originate a lot of spam. There is an ongoing war between people who create spam and those trying to filter their messages out. This leads to the introduction of many new techniques to defeat spam filters, including inserting into spam messages long strings of words that appear in messages that are not spam, as well as including words inside pictures. The techniques we have discussed here are only the first steps in fighting this war on spam.

Exercises

1. Suppose that E and F are events in a sample space and $p(E) = 1/3$, $p(F) = 1/2$, and $p(E \mid F) = 2/5$. Find $p(F \mid E)$.

2. Suppose that E and F are events in a sample space and $p(E) = 2/3$, $p(F) = 3/4$, and $p(F \mid E) = 5/8$. Find $p(E \mid F)$.

3. Suppose that Frida selects a ball by first picking one of two boxes at random and then selecting a ball from this box at random. The first box contains two white balls and three blue balls, and the second box contains four white balls and one blue ball. What is the probability that Frida picked a ball from the first box if she has selected a blue ball?

4. Suppose that Ann selects a ball by first picking one of two boxes at random and then selecting a ball from this box. The first box contains three orange balls and four black balls, and the second box contains five orange balls and

six black balls. What is the probability that Ann picked a ball from the second box if she has selected an orange ball?

5. Suppose that 8% of all bicycle racers use steroids, that a bicyclist who uses steroids tests positive for steroids 96% of the time, and that a bicyclist who does not use steroids tests positive for steroids 9% of the time. What is the probability that a randomly selected bicyclist who tests positive for steroids actually uses steroids?

6. When a test for steroids is given to soccer players, 98% of the players taking steroids test positive and 12% of the players not taking steroids test positive. Suppose that 5% of soccer players take steroids. What is the probability that a soccer player who tests positive takes steroids?

7. Suppose that a test for opium use has a 2% false positive rate and a 5% false negative rate. That is, 2% of people who do not use opium test positive for opium, and