Discrete Probability

Chapter 7

UWO – March 23, 2021

Plan for Chapter 7

1. Introduction to Discrete Probability

1.1 Finite Probability

1.2 The Probability of Complements and Unions of Events

2. Probability Theory

2.1 Assigning Probabilities

2.2 Probabilities of Complements and Unions of Events

2.3 Conditional Probability

2.4 Independence

2.5 Bernoulli Trials and the Binomial Distribution

3. Bayes' Theorem

3.1 Bayes' Theorem

3.2 Generalized Bayes' Theorem

3.3 Bayesian Spam Filters

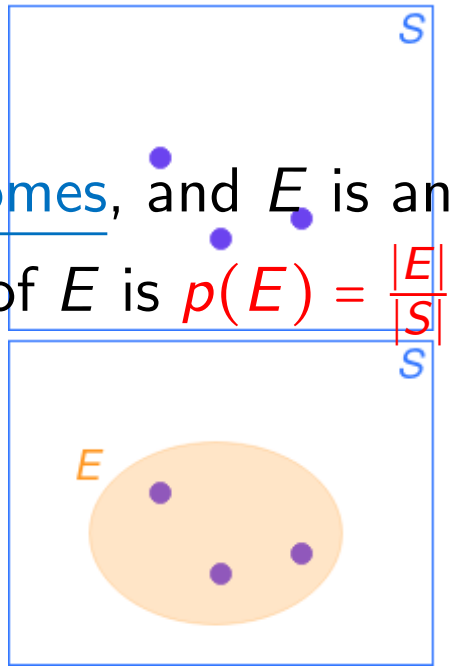
Plan for Chapter 7

Probability of an event

Pierre-Simon Laplace (1749 - 1827)

We first study Pierre-Simon Laplace's classical theory of probability, which he introduced in the 18-th century, when he analyzed games of chance.

1. Let us define these key terms:
   a. An *experiment* is a procedure that yields one outcome from a given set of possible outcomes.
   b. The *sample space* of the experiment is the set of possible outcomes.
   c. An *event* is a subset of the sample space.



Definition

If S is a finite sample space of equally likely outcomes, and E is an event, that is, a subset of S, then the *probability* of E is $p(E) = \frac{|E|}{|S|}$

2. For every event E, we have $0 \leq p(E) \leq 1$. This follows directly from the definition because $0 \leq p(E) = \frac{|E|}{|S|} \leq \frac{|S|}{|S|} \leq 1$, since we have: $0 \leq |E| \leq |S|$.

Applying Laplace's definition

Example

An urn contains four blue balls and five red balls. What is the probability that a ball chosen from the urn is blue?

Solution:

1. The probability that the ball is chosen is $\frac{4}{9}$ since there are nine possible outcomes, and four of these produce a blue ball.

sample space

events.

Example

What is the probability that when two dice are rolled, the sum of the numbers on the two dice is 7?

Solution:

Sample space

1. By the product rule there are $6^2 = 36$ possible outcomes.

Events.

2. Six of these sum to 7.
3. Hence, the probability of obtaining a 7 is $36 \cdot \frac{6}{36} = \frac{1}{6}$.

Applying Laplace's definition

Example

In a lottery, a player wins a large prize when they pick four digits that match, in correct order, four digits selected by a random mechanical process. What is the probability that a player wins the prize?

Solution:

1. By the product rule there are $10^4 = 10,000$ ways to pick four digits.
2. Since there is only 1 way to pick the correct digits, the probability of winning the large prize is $\frac{1}{10,000} = 0.0001$.

Example

A smaller prize is won if only three digits are matched. What is the probability that a player wins the small prize?

Solution:

1. If exactly three digits are matched, one of the four digits must be incorrect and the other three digits must be correct.
2. For the digit that is incorrect, there are 9 possible choices (all except the correct one).
3. Hence, by the sum rule , there a total of 36 possible ways to choose four digits that match exactly three of the winning four digits.
4. The probability of winning the small prize is $\frac{36}{10,000} = \frac{9}{2500} = 0.0036$.

Applying Laplace's definition

Example

There are many lotteries that award prizes to people who correctly choose a set of six numbers out of the first n positive integers, where n is usually between 30 and 60. What is the probability that a person picks the correct six numbers out of 40?

Solution:

1. The number of ways to choose six numbers out of 40 is
$$C(40, 6) = \frac{40!}{(34!6!)} = 3,838,380$$
2. Hence, the probability of picking a winning combination is
$$\frac{1}{3,838,380} \approx 0.00000026$$

Applying Laplace's definition

Example

What is the probability that the numbers 11, 4, 17, 39, and 23 are drawn in that order from a bin with 50 balls labeled with the numbers 1,2, . . . , 50 if:

1. the ball selected is not returned to the bin.
2. the ball selected is returned to the bin before the next ball is selected.

Solution: Use the product rule in each case.

1. *Sampling without replacement*:
   a. The probability is $\frac{1}{254,251,200}$ since there are $P(50, 5) = 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 = 254,251,200$ ways to choose the five balls.
2. *Sampling with replacement*:
   a. The probability is $\frac{1}{50^5} = \frac{1}{312,500,000}$ since $50^5 = 312,500,000$.

Plan for Chapter 7

The probability of complements and unions of events

Theorem

Let E be an event in the sample space S. The probability of the event $\overline{E} = S - E$, the complementary event of E, is given by

$$p(\overline{E}) = 1 - p(E)$$

Proof.

Using the fact that $|\overline{E}| = |S| - |E|$,

$$\begin{aligned} p(\overline{E}) &= \frac{|S| - |E|}{|S|} \\ &= 1 - \frac{|E|}{|S|} \\ &= 1 - p(E) \end{aligned}$$

■

The probability of complements and unions of events

Example

A sequence of 10 bits is chosen randomly. What is the probability that at least one of these bits is 0?

Solution:

1. Let E be the event that at least one of the 10 bits is 0.
2. Then $\overline{E}$ is the event that all of the bits are 1s.
3. The size of the sample space S is 2^{10}. Hence,

$$\begin{aligned} p(E) &= 1 - p(\overline{E}) \\ &= 1 - \frac{|\overline{E}|}{|S|} \\ &= 1 - \frac{1}{2^{10}} \\ &= 1 - \frac{1}{1024} \\ &= \frac{1023}{1024} \end{aligned}$$

The probability of complements and unions of events

Theorem

Let E_1 and E_2 be events in the sample space S. Then

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

Proof.

Given the inclusion-exclusion formula from Chapter 3, we have

$$|A \cup B| = |A| + |B| - |A \cap B|,$$

it follows that:

$$\begin{aligned} p(E_1 \cup E_2) &= \frac{|E_1 \cup E_2|}{|S|} \\ &= \frac{|E_1| + |E_2| - |E_1 \cap E_2|}{|S|} \\ &= \frac{|E_1|}{|S|} + \frac{|E_2|}{|S|} - \frac{|E_1 \cap E_2|}{|S|} \\ &= p(E_1) + p(E_2) - p(E_1 \cap E_2) \end{aligned}$$

The probability of complements and unions of events

Example

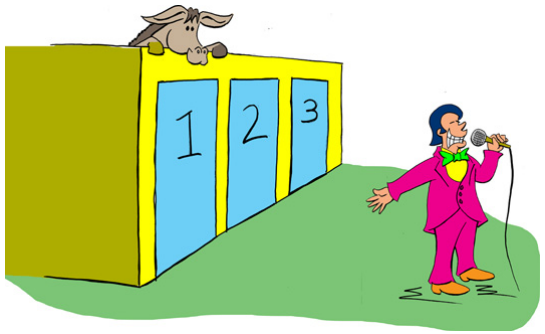
What is the probability that a positive integer selected at random from the set of positive integers not exceeding 100 is divisible by either 2 or 5?

Solution:

1. Let E_2 be the event that the integer is divisible by 2 and E_5 be the event that it is divisible 5?
2. Then the event that the integer is divisible by 2 or 5 is $E_2 \cup E_5$ and $E_2 \cap E_5$ is the event that it is divisible by 2 and 5.
3. It follows that:

$$\begin{aligned} p(E_2 \cup E_5) &= p(E_2) + p(E_5) - p(E_2 \cap E_5) \\ &= \frac{50}{100} + \frac{20}{100} - \frac{10}{100} \\ &= \frac{3}{5} \end{aligned}$$

Monty Hall puzzle



1. You are asked to select one of the three doors to open. There is a large prize behind one of the doors and if you select that door, you win the prize.
2. After you select a door, the game show host opens one of the other doors (which he knows is not the winning door). The prize is not behind the door and he gives you the opportunity to switch your selection. Should you switch?
3. This is a notoriously confusing problem that has been the subject of much discussion. Here's why you should switch:
   a. The probability that your initial pick is correct is $\frac{1}{3}$. This is the same whether or not you switch doors...
   b. ...but since the game show host always opens a door that does not have the prize, if you switch the probability of winning will be $\frac{2}{3}$, because you win if your initial pick was not the correct door and the probability your initial pick was wrong is $\frac{2}{3}$.
   c. More details on wikipedia: Monty Hall puzzle

Plan for Chapter 7

1. Introduction to Discrete Probability

1.1 Finite Probability

1.2 The Probability of Complements and Unions of Events

2. Probability Theory

2.1 Assigning Probabilities

2.2 Probabilities of Complements and Unions of Events

2.3 Conditional Probability

2.4 Independence

2.5 Bernoulli Trials and the Binomial Distribution

3. Bayes' Theorem

3.1 Bayes' Theorem

3.2 Generalized Bayes' Theorem

3.3 Bayesian Spam Filters

Plan for Chapter 7

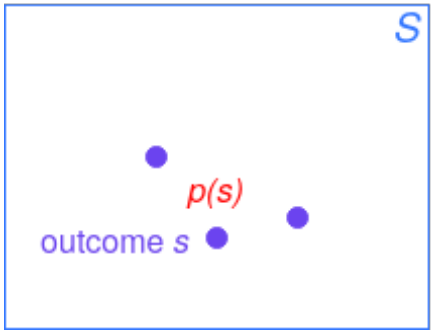
Assigning probabilities

Laplace's definition from the previous section, assumes that all outcomes are equally likely. Now we introduce a more general definition of probabilities that avoids this restriction.

Definition

Let S be a sample space of an experiment with a finite number of outcomes. We assign a probability $p(s)$ to each outcome s, so that:

1. $0 \leq p(s) \leq 1$ for each $s \in S$
2. $\sum_{s \in S} p(s) = 1$



The function p from the set of all outcomes of the sample space S to the interval $[0, 1]$ is called a *probability distribution*.

Assigning probabilities

Example

What probabilities should we assign to the outcomes H (heads) and T (tails) when a fair coin is flipped?

Solution:

1. $p(H) + p(T) = 1$
2. $p(H) = p(T)$ since it's a fair coin.
3. Hence, $P(H) = \frac{1}{2}$ and $P(T) = \frac{1}{2}$

Example

What probabilities should be assigned to these outcomes when the coin is biased so that heads comes up twice as often as tails?

1. We have $p(H) = 2p(T)$.
2. Because $p(H) + p(T) = 1$, it follows that
3. $2p(T) + p(T) = 3p(T) = 1$.
4. Hence, $p(T) = \frac{1}{3}$ and $p(H) = \frac{2}{3}$.

Uniform distribution

Definition (uniform distribution)

Suppose that S is a set with n elements. The *uniform distribution* assigns the probability $\frac{1}{n}$ to each element of S. (Note that we could have used Laplace's definition here.)

Example

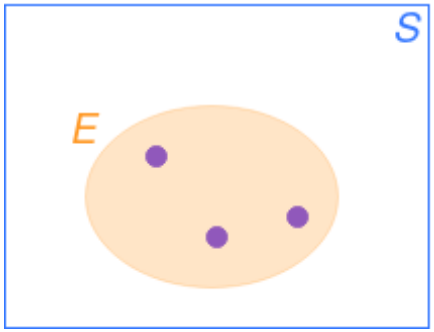
Consider again the coin flipping example, but with a fair coin. Now $p(H) = p(T) = \frac{1}{2}$.

Probability of an event

Definition

The probability of the event E is the sum of the probabilities of the outcomes in E.

$$p(E) = \sum_{s \in E} p(s)$$



Note that no assumption is being made about the distribution.

Example

Example

Suppose that a dice is biased so that 3 appears twice as often as each other number, but that the other five outcomes are equally likely. What is the probability that an odd number appears when we roll this dice?

Solution:

1. The assumptions imply:
   a. $p(3) = 2p(1)$ and $p(1) = p(2) = p(4) = p(5) = p(6)$, and
   b. $p(1) + p(2) + p(3) + p(4) + p(5) + p(6) = 1$.
2. Thus, we have:
$$p(3) = \frac{2}{7} \text{ and } p(1) = p(2) = p(4) = p(5) = p(6) = \frac{1}{7}.$$
3. We want the probability of the event $E = \{1, 3, 5\}$.
4. Hence, we have:
$$p(E) = p(1) + p(3) + p(5).$$
5. That is:
$$p(E) = \frac{1}{7} + \frac{2}{7} + \frac{1}{7} = \frac{4}{7}.$$

Plan for Chapter 7

Probabilities of complements and unions of events

1. Complements: $p(\overline{E}) = 1 - p(E)$ still holds. Since each outcome is either in E or in $\overline{E}$, but not in both, we have:

$$\sum_{s \in S} p(s) = 1 = p(E) + p(\overline{E})$$

2. Unions: $p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$ also still holds under the new definition.
3. This follows again from the inclusion-exclusion formula.

Combinations of events

Theorem

If $E_1, E_2, \ldots, E_m$ is a sequence of pairwise disjoint events in a sample space S, then we have:

$$p\left(\bigcup_{i=1}^{i=m} E_i\right) = \sum_{i=1}^{i=m} p(E_i)$$

1. Each event E_i consists of finitely many outcomes $x_{i,1}, \ldots, x_{i,n_i}$ where n_i is the cardinality of E_i, for $1 \leq i \leq m$.
2. Hence, we have

$$\begin{aligned} \bigcup_{i=1}^{i=m} E_i &= E_1 \cup \cdots \cup E_m \\ &= \{x_{1,1}, \ldots, x_{1,n_1}\} \cup \cdots \cup \{x_{m,1}, \ldots, x_{m,n_m}\} \end{aligned}$$

3. Since the events E_i are pairwise disjoint, we have:

$$\begin{aligned} p\left(\bigcup_{i=1}^{i=m} E_i\right) &= p(x_{1,1}) + \cdots + p(x_{1,n_1}) + \cdots + p(x_{m,1}) + \cdots + p(x_{m,n_m}) \\ &= \sum_{i=1}^{i=m} (p(x_{i,1}) + \cdots + p(x_{i,n_i})) \\ &= \sum_{i=1}^{i=m} p(E_i). \end{aligned}$$

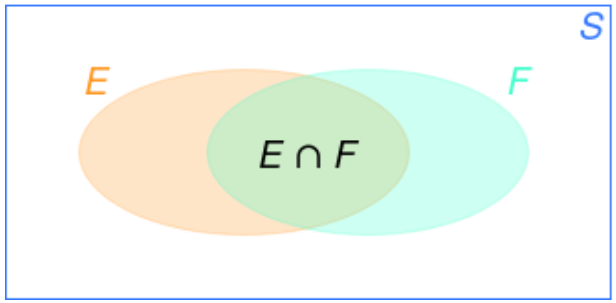
Plan for Chapter 7

Conditional probability

Definition

Let E and F be events with $p(F) > 0$. The conditional probability of E given F, denoted by $P(E \mid F)$, is defined as:

$$p(E|F) = \frac{p(E \cap F)}{p(F)}$$



Example

A bit string of length four is generated at random so that each of the 16 strings are equally likely. What is the probability that it contains at least two consecutive 0s, given that its first bit is a 0?

Solution:

1. Let E be the event that the bit string contains at least two consecutive 0s, and F be the event that the first bit is a 0.
2. Since $E \cap F = \{0000, 0001, 0010, 0011, 0100\}$, $p(E \cap F) = \frac{5}{16}$.
3. Because 8 bit strings of length 4 start with a 0, $p(F) = \frac{8}{16} = \frac{1}{2}$.
4. Hence, $p(E|F) = \frac{p(E \cap F)}{p(F)} = \frac{\frac{5}{16}}{\frac{1}{2}} = \frac{5}{8}$

Conditional probability

Example

What is the conditional probability that a family with two children has two boys, given that they have at least one boy. Assume that each of the possibilities BB, BG, GB, and GG is equally likely where B represents a boy and G represents a girl.

Solution:

1. Let E be the event that the family has two boys and let F be the event that the family has at least one boy.
2. Then $E = \{BB\}$, $F = \{BB, BG, GB\}$, and $E \cap F = \{BB\}$.
3. It follows that $p(F) = \frac{3}{4}$ and $p(E \cap F) = \frac{1}{4}$.
4. Hence, $p(E|F) = \frac{p(E \cap F)}{p(F)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$

Plan for Chapter 7

Independence

Intuition:
Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other.

Definition

Formally, the events E and F are independent if and only if

$$p(E \cap F) = p(E)p(F)$$

Note that independence of events E and F implies:

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

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

Independence

Example

Suppose E is the event that a randomly generated bit string of length four begins with a 1 and F is the event that this bit string contains an even number of 1s. Are E and F independent if the 16 bit strings of length four are equally likely?

Solution:

1. There are eight bit strings of length four that begin with a 1, and eight bit strings of length four that contain an even number of 1s.
2. Since the number of bit strings of length 4 is 16, $p(E) = p(F) = \frac{8}{16} = \frac{1}{2}$.
3. Since $E \cap F = \{1111, 1100, 1010, 1001\}$, $\quad p(E \cap F) = \frac{4}{16} = \frac{1}{4}$.
4. We conclude that E and F are independent, because $p(E \cap F) = \frac{1}{4} = (\frac{1}{2})(\frac{1}{2}) = p(E)p(F)$

Independence

Example

Assume (as in the previous example) that each of the four ways a family can have two children (BB, GG, BG, GB) is equally likely. Are the events E, that a family with two children has two boys, and F, that a family with two children has at least one boy, independent?

Solution:

1. Because $E = \{BB\}$, $p(E) = \frac{1}{4}$.
2. We saw previously that that $p(F) = \frac{3}{4}$ and $p(E \cap F) = \frac{1}{4}$.
3. The events E and F are not independent since $p(E)p(F) = \frac{3}{16} \neq \frac{1}{4} = p(E \cap F)$.

Pairwise and mutual independence

Definition (Pairwise Independence)

The events $E_1, E_2, \ldots, E_n$ are *pairwise independent* if and only if $p(E_i \cap E_j) = p(E_i)p(E_j)$ for all pairs i and j with $i \leq j \leq n$.

Definition (Mutual Independence)

The events are *mutually independent* if

$$p(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_m}) = p(E_{i_1})p(E_{i_2}) \cdots p(E_{i_m})$$

whenever $i_j, j = 1, 2, \ldots, m$, are integers with $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and $m \geq 2$.

NOTE: mutually independent events are pairwise independent, but some pairwise independent events are not mutually independent.

Plan for Chapter 7

Bernoulli trials

James Bernoulli
(1654 - 1705)

Definition

Suppose an experiment can have only two possible outcomes, *e.g.*, the flipping of a coin or the random generation of a bit.

1. Each performance of the experiment is called a *Bernoulli trial*.
2. Often, one outcome is called a *success* and the other a *failure*.
3. If p is the probability of success and q the probability of failure, then $p + q = 1$.

Many problems involve determining the probability of k successes when an experiment consists of n mutually independent Bernoulli trials.

Bernoulli trials

Example

A coin is biased so that the probability of heads is $\frac{2}{3}$. What is the probability that exactly four heads occur when the coin is flipped seven times?

Solution:

1. There are $2^7 = 128$ possible outcomes.
2. The number of ways four of the seven flips can be heads is $C(7,4)$.
3. The probability of each of the outcomes is $(\frac{2}{3})^4(\frac{1}{3})^3$ since the seven flips are independent.
4. Hence, the probability that exactly four heads occur is

$$C(7,4)(\frac{2}{3})^4(\frac{1}{3})^3 = \frac{(35 \cdot 16)}{2^7} = \frac{560}{2187}$$

Probability of k successes in n independent Bernoulli trials.

Theorem
The probability of exactly k successes in n independent Bernoulli trials, with probability of success p and probability of failure $q = 1 - p$, is

$$C(n,k)p^k q^{n-k}$$

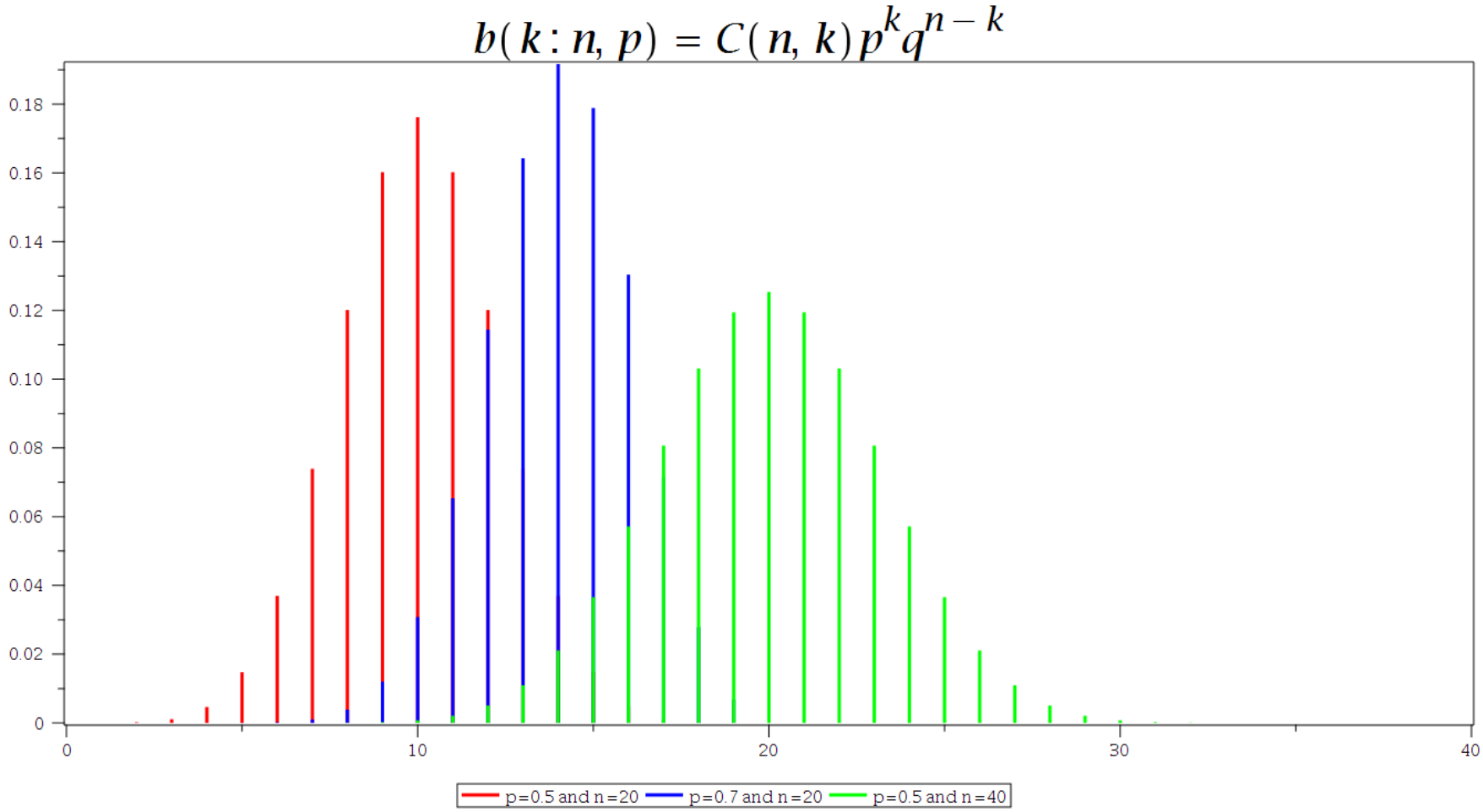
Proof.

1. The outcome of n Bernoulli trials is an n-tuple $(t_1, t_2, \ldots, t_n)$, where each is t_i either S (success) or F (failure).
2. The probability of each outcome of n trials consisting of k successes and $n - k$ failures (in any given order) is $p^k q^{n-k}$.
3. Because there are $C(n,k)$ n-tuples of S's and F's that contain exactly k S's, the probability of k successes is $C(n,k)p^k q^{n-k}$.

■

We denote by $b(k : n, p)$ the probability of k successes in n independent Bernoulli trials with p the probability of success. Viewed as a function of k, $b(k : n, p)$ is the *binomial distribution*. Hence we have, $b(k : n, p) = C(n,k)p^k q^{n-k}$.

Probability of k successes in n independent Bernoulli trials.



(Graph made in Maple!)

Plan for Chapter 7

1. Introduction to Discrete Probability

1.1 Finite Probability

1.2 The Probability of Complements and Unions of Events

2. Probability Theory

2.1 Assigning Probabilities

2.2 Probabilities of Complements and Unions of Events

2.3 Conditional Probability

2.4 Independence

2.5 Bernoulli Trials and the Binomial Distribution

3. Bayes' Theorem

3.1 Bayes' Theorem

3.2 Generalized Bayes' Theorem

3.3 Bayesian Spam Filters

Plan for Chapter 7

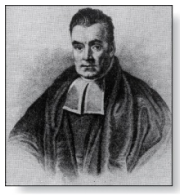
Motivation for Bayes' Theorem

Bayes' theorem allows us to use probability theory to answer questions such as the following:

1. Given that someone tests positive for having a particular disease, what is the probability that they actually do have the disease?
2. Given that someone tests negative for the disease, what is the probability, that in fact they do have the disease?

Bayes' theorem has applications to medicine, law, artificial intelligence, engineering, and many diverse other areas.

Bayes' Theorem



Thomas Bayes
(1702 - 1761)

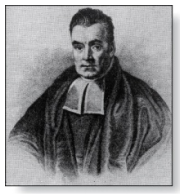
Theorem (Bayes' Theorem)

Suppose that E and F are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then, we have:

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

This helps when the conditional probability $p(E \mid F)$ is easier to estimate compared to $p(F \mid E)$.

Bayes' Theorem



Thomas Bayes
(1702 - 1761)

Example

We have two boxes. The first box contains two green balls and seven red balls. The second contains four green balls and three red balls. Bob selects one of the boxes at random. Then he selects a ball from that box at random. If he has a red ball, what is the probability that he selected a ball from the first box $p(F \mid E)$?

Solution:

1. Let E be the event that Bob has chosen a red ball and F be the event that Bob has chosen the first box.
   $$p(E \mid F) = \frac{7}{2+7} \qquad p(E \mid \overline{F}) = \frac{3}{4+3} \qquad p(F \mid E) = ?$$
2. By Bayes' theorem the probability that Bob has picked the first box is:
   $$p(F \mid E) = \frac{p(E|F)p(F)}{p(E|F)p(F)+p(E|\overline{F})p(\overline{F})}$$

$$p(F \mid E) = \frac{(\frac{7}{9})(\frac{1}{2})}{(\frac{7}{9})(\frac{1}{2})+(\frac{3}{7})(\frac{1}{2})} = \frac{\frac{7}{18}}{\frac{38}{63}} = \frac{49}{76} \approx 0.645$$

Derivation of Bayes' Theorem

1. Recall the definition of the conditional probability $p(E \mid F)$:

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

2. From this definition, it follows that:

$$p(E \mid F) = \frac{p(E \cap F)}{p(F)} \quad \text{and} \quad p(F \mid E) = \frac{p(E \cap F)}{p(E)}$$

continued →

Derivation of Bayes' Theorem

1. Rearranging the equations from the previous slide we get:

$$p(E \mid F)p(F) = p(E \cap F) \quad \text{and} \quad p(F \mid E)p(E) = p(E \cap F)$$

2. Substituting for $p(E \cap F)$ we get:

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

3. Solving for $p(E \mid F)$ and $p(F \mid E)$ we get:

$$p(E \mid F) = \frac{p(F \mid E)p(E)}{p(F)} \quad \text{and} \quad p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E)}$$

continued →

Derivation of Bayes' Theorem

1. Recall $p(F \mid E) = \frac{p(E|F)p(F)}{p(E)}$ from the previous slide.
2. Note also that $p(E) = p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})$
3. Indeed:
   a. Since $E = E \cap S = E \cap (F \cup \overline{F}) = (E \cap F) \cup (E \cap \overline{F})$
   b. and $(E \cap F) \cap (E \cap \overline{F}) = \varnothing$,
   c. we deduce: $p(E) = p(E \cap F) + p(E \cap \overline{F})$.
   d. Then, by the definition of conditional probability, we have:

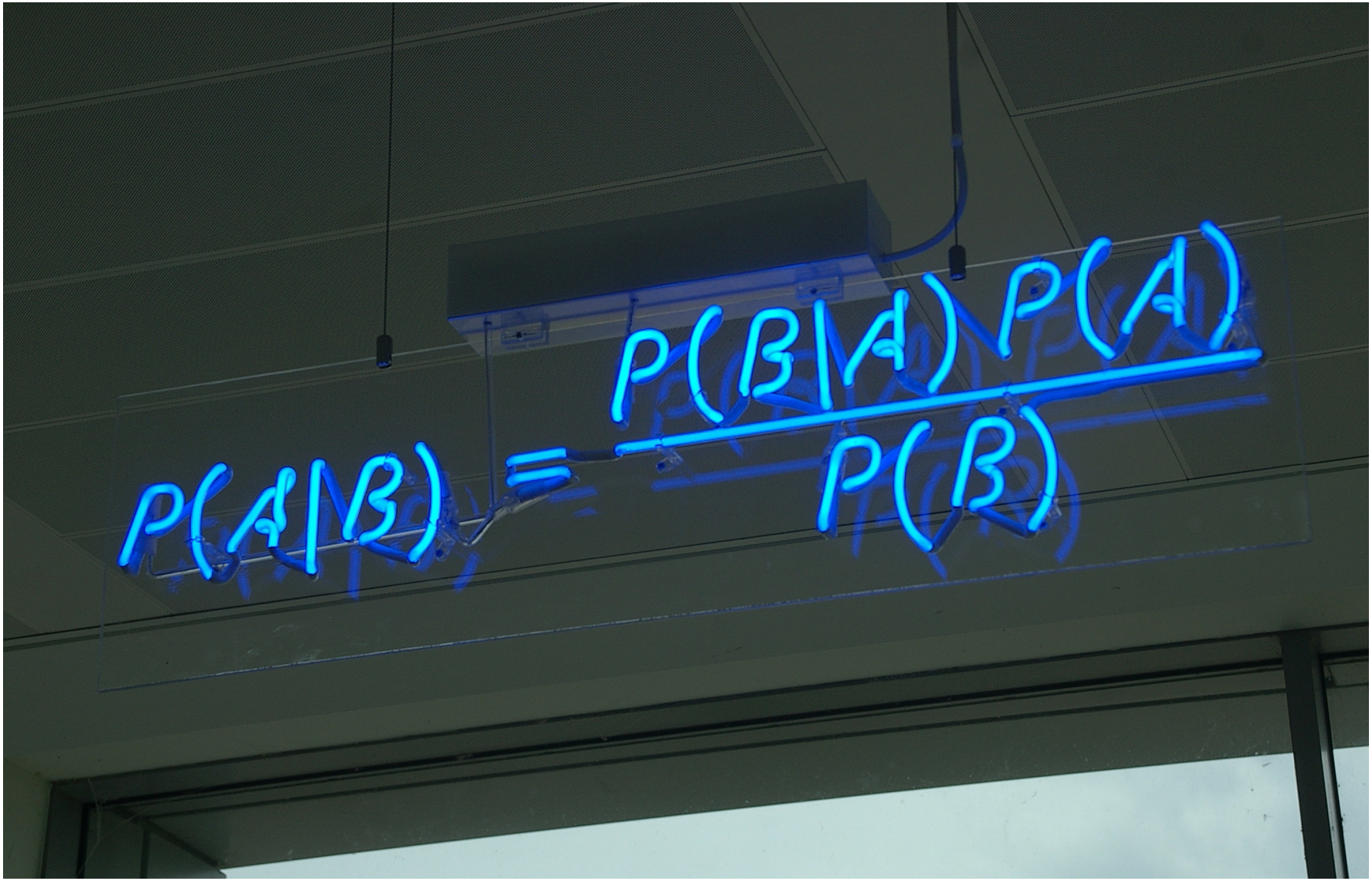
$$\begin{aligned} p(E) &= p(E \cap F) + p(E \cap \overline{F}) \\ &= p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F}) \end{aligned}$$

4. Hence, we have:

$$\begin{aligned} p(F \mid E) &= \frac{p(E|F)p(F)}{p(E)} \\ &= \frac{p(E|F)p(F)}{p(E|F)p(F)+p(E|\overline{F})p(\overline{F})}. \end{aligned}$$

■

Simple form of Bayes' Theorem



A blue neon sign at the Autonomy Corporation, Cambridge, showing the simple statement of Bayes' Theorem.

Interpretation of the simple form of Bayes' Theorem

1. Bayes' Theorem links the degree of belief in a proposition before and after accounting for evidence.
2. Proposition A, Evidence B
3. $p(A)$ – prior probability (initial degree of belief in A)
4. $p(A \mid B)$ – posterior probability (degree of belief in A after having accounted for B)
5. $\frac{p(B|A)}{p(B)}$ – the support provided for A by B

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} = \frac{p(B \mid A)}{p(B)} \cdot p(A)$$

Train problem

1. Suppose someone
   told you they had a conversation
   with a person on a train.
2. If you knew nothing else about
   this conversation, you would
   compute the probability that
   this person was a woman as 50%
3. Now, suppose you were also told
   that the person had long hair.
4. Bayes' theorem can be used
   to calculate the probability that
   the person is a woman, given the
   additional knowledge we have.

How to solve the train problem

1. W = event that the conversation partner is a woman
2. L = the conversation partner has long hair
3. Suppose we know that 75% of women have long hair and 15% of men have long hair. These are statistics that can be directly estimated.

$$p(L \mid W) = 0.75 \qquad p(L \mid \overline{W}) = 0.15$$

4. What about $p(W \mid L)$?

$$p(W \mid L) = \frac{p(L \mid W)p(W)}{p(L \mid W)p(W) + p(L \mid \overline{W})p(\overline{W})}$$

$$p(W \mid L) = \frac{0.75 \cdot 0.5}{0.75 \cdot 0.5 + 0.15 \cdot 0.5} = 0.83$$

Applying Bayes' Theorem

Example

1. Suppose that one person in 100,000 has a particular disease.
2. There is a test for the disease that gives a positive result 99% of the time when given to someone with the disease.
3. When given to someone without the disease, 99.5% of the time it gives a negative result.

Questions:

1. Find the probability that a person who test positive has the disease.
2. Find the probability that a person who test negative does not have the disease.
3. Should someone who tests positive be worried?

Applying Bayes' Theorem

Solution: What if the test is positive?

1. Let D be the event that the person has the disease, and E be the event that this person tests positive.
2. We need to compute $p(D \mid E)$ from $p(D), p(E \mid D), p(\overline{E} \mid \overline{D})$.

1. $p(D) = \frac{1}{100{,}000} = 0.00001$
2. $p(\overline{D}) = 1 - 0.00001 = 0.99999$
3. $p(E \mid D) = 0.99$
4. $p(\overline{E} \mid D) = 0.01$
5. $p(E \mid \overline{D}) = 0.005$
6. $p(\overline{E} \mid \overline{D}) = 0.995$

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

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

$$\approx 0.002$$

Can you use this formula to explain why the resulting probability is surprisingly small?

So, don't worry too much, if your test for this disease comes back positive.

Applying Bayes' Theorem

Solution: What if the result is negative?

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

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

$$\approx 0.9999999$$

So, the probability you have the disease if you test negative is $p(D \mid \overline{E}) \approx 1 - 0.9999999 = 0.0000001$.

So, it is extremely unlikely you have the disease if you test negative.

Plan for Chapter 7

Generalized Bayes' Theorem

Definition (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 $\bigcup_i^n F_i = S$. Assume that $p(E) \neq 0$ for $i = 1, 2, \ldots, n$. Then,

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

Tutorial 10 asks for the proof.

Plan for Chapter 7

Bayesian spam filters

1. How do we develop a tool for determining whether an email is likely to be spam?
2. If we have an initial set **B***(ad)* of spam messages and set **G***(ood)* of non-spam messages. We can use this information along with Bayes' law to predict the probability that a new email message is spam.
3. We look at a particular word w, and count the number of times that it occurs in B and in G; $n_B(w)$ and $n_G(w)$.
   a. Estimated probability that spam email contains w : $p(w) = \frac{n_B(w)}{|B|}$
   b. Estimated probability that non-spam email contains w : $q(w) = \frac{n_G(w)}{|G|}$

continued →

Bayesian spam filters

Let S be the event that the message is spam, and E be the event that the message contains the word w. Using Bayes' Rule,

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

Assuming that it is equally likely that an arbitrary message is spam and is not spam; i.e., $p(S) = \frac{1}{2}$.

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

Using our empirical estimates of

$$p(w) = p(E \mid S)$$
$$q(w) = p(E \mid \overline{S})$$

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

$r(w)$ estimates the probability that the message is spam. We can classify the message as spam if $r(w)$ is above a threshold.

Note: If we have data on the frequency of spam messages, we can obtain a better estimate for p(S). (*See Tutorial 10.*)

Bayesian spam filters

Example

We find that the word "Rolex" occurs in 250 out of 2000 spam messages and occurs in 5 out of 1000 non-spam messages. Estimate the probability that an incoming message is spam. Suppose our threshold for rejecting the email is 0.9.

Solution:

1. $p(Rolex) = \frac{250}{2000} = 0.0125$ and $q(Rolex) = \frac{5}{1000} = 0.005$.
2. $r(Rolex) = \frac{p(Rolex)}{p(Rolex)+q(Rolex)}$
3. $r(Rolex) = \frac{0.125}{0.125+0.005}$
4. $r(Rolex) \approx 0.962$

We classify the message as spam and reject the email!

Bayesian spam filters using multiple words

1. Accuracy can be improved by considering more than one word as evidence.
2. Consider the case where E_1 and E_2 denote the events that the message contains the words w_1 and w_2 respectively.
3. We make the simplifying assumption that the events E_1 and E_2 are independent given S, that is, we have:
$$p(E_1 \cap E_2|S) = p(E_1|S)p(E_2|S).$$
4. We again assume $p(S) = \frac{1}{2}$.

Then, we have:

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

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

See Tutorial 10 for a proof.

Bayesian spam filters using multiple words

Example

We have 2000 spam messages and 1000 non-spam messages. The word "stock" occurs 400 times in the spam messages and 60 times in the non-spam. The word "undervalued" occurs in 200 spam messages and 25 non-spam. What is the probability for an incoming message to be spam if both words are present.

1. $p(stock) = \frac{400}{2000} = 0.2$ and $q(stock) = \frac{60}{1000} = 0.06$
2. $p(undervalued) = \frac{200}{2000} = 0.1$ and
   $q(undervalued) = \frac{25}{1000} = 0.025$

$$r(stock, undervalued) = \frac{p(stock)p(undervalued)}{p(stock)p(undervalued) + q(stock)q(undervalued)}$$
$$= \frac{(0.2)(0.1)}{(0.2)(0.1) + (0.06)(0.025)} \approx 0.930$$

If our threshold is .9, we classify the message as spam and reject it.

Bayesian spam filters using multiple words

In general, the more words we consider, the more accurate the spam filter. With the independence assumption if we consider k words:

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

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

We can further improve the filter by considering pairs of words as a single block or certain types of strings.