Discrete Mathematics and Its Applications

Lecture 5: Discrete Probability: Random Variables

MING GAO

DaSE@ ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

Nov. 29, 2018

Outline

1. Random Variable
2. Bernoulli Trials and the Binomial Distribution
3. Bayes' Theorem
4. Applications of Bayes' Theorem
5. Take-aways

Random variables

Definition: A **random variable** (r.v.) X is a function from sample space Ω of an experiment to the set of real numbers in R, i.e.,

$$\forall \omega \in \Omega, X(\omega) = x \in R.$$

Random variables

Definition: A **random variable** (r.v.) X is a function from sample space Ω of an experiment to the set of real numbers in R, i.e.,

$$\forall \omega \in \Omega, X(\omega) = x \in R.$$

Remarks

- Note that a random variable is a function. It is not a variable, and it is not random!
- We usually use notation X, Y, etc. to represent a r.v., and x, y to represent the numerical values. For example, $X = x$ means that r.v. X has value x.
- The domain of the function can be countable and uncountable. If it is countable, the random variable is a discrete r.v., otherwise continuous r.v..

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then

$$X(H) = 1, X(T) = 0.$$

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then

$$X(H) = 1, X(T) = 0.$$

And then tossed again. We define sample space $\Omega = \{HH, HT, TH, TT\}$. If Y is the r.v. whose value is the number of heads obtained, then

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

Examples of r.v.

A coin is tossed. If X is the r.v. whose value is the number of heads obtained, then

$$X(H) = 1, X(T) = 0.$$

And then tossed again. We define sample space $\Omega = \{HH, HT, TH, TT\}$. If Y is the r.v. whose value is the number of heads obtained, then

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

When a player rolls a die, he will win \$1 if the outcome is 1,2 or 3, otherwise lose 1\$. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and define X as follows:

$$X(1) = X(2) = X(3) = 1, X(4) = X(5) = X(6) = -1.$$

Random variables VS. events

Suppose now that a sample space $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ is given, and r.v. X on Ω is defined the number of heads obtained when we toss a coin twice.

Random variables VS. events

Suppose now that a sample space $\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ is given, and r.v. X on Ω is defined the number of heads obtained when we toss a coin twice.

- Event E_1 represents only one head obtained. Hence,

$$E_1 = \{\omega : X(\omega) = 1\};$$

- Event E_2 represents even heads obtained. Hence,

$$E = \{\omega : X(\omega) \mod 2 = 0\};$$

- Event E_2 represents at least one heads obtained. Hence,

$$E = \{\omega : X(\omega) > 0\}.$$

These indicate that we can also define probability about r.v.s.

Distribution

Definition: The **distribution** of a r.v. X on a sample space Ω is the set of pairs $(r, p(X = r))$ for all $r \in X(\Omega)$, where $P(X = r)$ is the probability that r.v. X takes value r. That is, the set of pairs in this distribution is determined by probabilities $P(X = r)$ for $r \in X(\Omega)$.

Distribution

Definition: The **distribution** of a r.v. X on a sample space Ω is the set of pairs $(r, p(X = r))$ for all $r \in X(\Omega)$, where $P(X = r)$ is the probability that r.v. X takes value r. That is, the set of pairs in this distribution is determined by probabilities $P(X = r)$ for $r \in X(\Omega)$.

Remarks

- Distribution is also a function;
- If we define event E which X has vaule x in Ω, then,

$$P(E) = P(\{\omega : X(\omega) = x\}) = P(X = x) = f(x);$$

- $f(x)$ is a probability distribution (function) if
  - $f(x) \geq 0$;
  - $\sum_x f(x) = 1$;
- $P(X \leq c) = P(\{\omega \in \Omega : X(\omega) \leq c\})$.

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$		

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$		

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

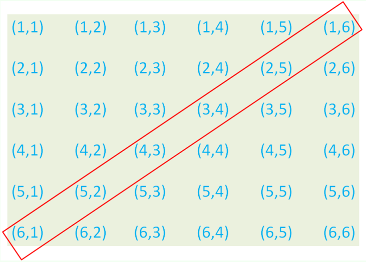
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

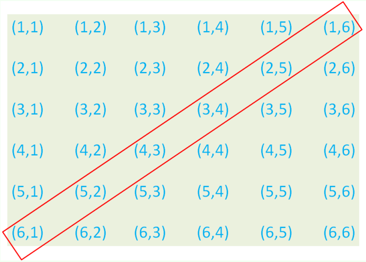


Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$		

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

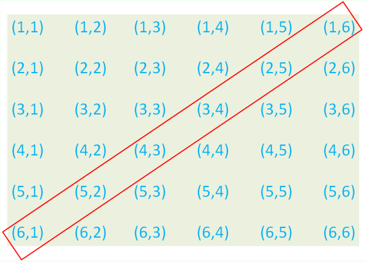


Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?



Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$		

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

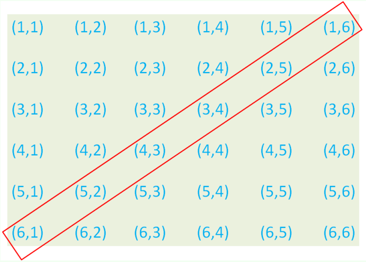
(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$		

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?



Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$	11	$\frac{1}{18}$

Examples of distribution

Question: Let X be the sum of the numbers that appear when a pair of dice is rolled. What are the values and probabilities of this random variable for 36 possible outcomes (i, j), when these two dices are rolled?

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Solution:

value	prob.	value	prob.
2	$\frac{1}{36}$	3	$\frac{1}{18}$
4	$\frac{1}{12}$	5	$\frac{1}{9}$
6	$\frac{5}{36}$	7	$\frac{1}{6}$
8	$\frac{5}{36}$	9	$\frac{1}{9}$
10	$\frac{1}{12}$	11	$\frac{1}{18}$
12	$\frac{1}{36}$		

Joint and marginal probability distributions

Definition

Let X and Y be two r.v.s,

- $f(x, y) = P(X = x \wedge Y = y)$ is the joint probability distribution;

Joint and marginal probability distributions

Definition

Let X and Y be two r.v.s,

- $f(x, y) = P(X = x \land Y = y)$ is the joint probability distribution;
- $f_1(x)$ is the marginal probability distribution for r.v. X.

Joint and marginal probability distributions

Definition

Let X and Y be two r.v.s,

- $f(x, y) = P(X = x \wedge Y = y)$ is the joint probability distribution;
- $f_1(x)$ is the marginal probability distribution for r.v. X.

Note that

$$
\begin{aligned}
f_1(x) &= P(X = x) = P(X = x \wedge \Omega) \\
&= P(X = x \wedge (Y = y_1 \vee Y = y_2 \vee \cdots)) \\
&= P((X = x \wedge Y = y_1) \vee (X = x \wedge Y = y_2) \vee \cdots) \\
&= P(X = x \wedge Y = y_1) + P(X = x \wedge Y = y_2) + \cdots \\
&= \sum_{y} P(X = x \wedge Y = y) = \sum_{y_i} f(x, y)
\end{aligned}
$$

Independence of r.v.

Definition

- Let r.v.s X and Y are **pair-wise independent** if and only if for $\forall x, y \in R$, we have

$$P(X = x \wedge Y = y) = P(X = x)P(Y = y);$$

Independence of r.v.

Definition

- Let r.v.s X and Y are **pair-wise independent** if and only if for $\forall x, y \in R$, we have

$$P(X = x \wedge Y = y) = P(X = x)P(Y = y);$$

- Let r.v.s $X_1, X_2, \cdots, X_n$ are **mutually independent** if and only if for $\forall x_{i_j} \in R$

$$P(X_{i_1} = x_{i_1} \wedge X_{i_2} = x_{i_2} \wedge \cdots \wedge X_{i_m} = x_{i_m})$$
$$= P(X_{i_1} = x_{i_1})P(X_{i_2} = x_{i_2}) \cdots P(X_{i_m} = x_{i_m}),$$

where $i_j, j = 1, 2, \cdots, m$, are integers with $1 \le i_1 < i_2 < \cdots < i_m \le n$ and $m \ge 2$.

Independence of r.v. Cont'd

Corollary

Let r.v.s X and Y are **independent** if and only if for $\forall x, y \in R$, s.t. $P(Y = y) \neq 0$, we have

$$\begin{aligned} P(X = x | Y = y) &= \frac{P(X = x \wedge Y = y)}{P(Y = y)} \\ &= \frac{P(X = x)P(Y = y)}{P(Y = y)} = P(X = x) \end{aligned}$$

Independence of r.v. Cont'd

Corollary

Let r.v.s X and Y are **independent** if and only if for $\forall x, y \in R$, s.t. $P(Y = y) \neq 0$, we have

$$\begin{aligned} P(X = x | Y = y) &= \frac{P(X = x \wedge Y = y)}{P(Y = y)} \\ &= \frac{P(X = x)P(Y = y)}{P(Y = y)} = P(X = x) \end{aligned}$$

Definition

Let X and Y be two r.v.s,

- $f(x, y) = P(X = x \wedge Y = y)$ is the joint probability function;
- $f_1(x)$ is the marginal probability function for r.v. X.

Examples of distribution

Question: A biased coin ($Pr(H) = 2/3$) is flipped twice. Let X count the number of heads. What are the values and probabilities of this random variable?

Solution:

Let X_i count the number of heads in the $i-$th flip.

$$\begin{aligned} Pr(X=0) &= Pr(X_1 = 0 \wedge X_2 = 0) = Pr(X_1 = 0)P(X_2 = 0) \\ &= (1/3)^2 = 1/9 \\ Pr(X=1) &= Pr((X_1 = 0 \wedge X_2 = 1) \vee (X_1 = 1 \wedge X_2 = 0)) \\ &= Pr(X_1 = 1)P(X_2 = 0) + Pr(X_1 = 0)P(X_2 = 1) \\ &= 2 \cdot 1/3 \cdot 2/3 = 4/9 \\ Pr(X=2) &= Pr(X_1 = 1 \wedge X_2 = 1) = Pr(X_1 = 1)P(X_2 = 1) \\ &= (2/3)^2 = 4/9 \end{aligned}$$

Bernoulli Trials

Definition

Each performance of an experiment with two possible outcomes is called a **Bernoulli trial**.

Bernoulli Trials

Definition

Each performance of an experiment with two possible outcomes is called a **Bernoulli trial**.

- In general, a possible outcome of a Bernoulli trial is called a **success** or a **failure**.
- If p is the probability of a success and q is the probability of a failure, it follows that $p + q = 1$.
- Many problems can be solved by determining the probability of k successes when an experiment consists of n **mutually independent Bernoulli trials**.

Mutually independent Bernoulli trials

Flipping coin

Question: A coin is biased so that the probability of heads is 2/3. What is the probability that exactly four heads come up when the coin is flipped seven times, assuming that the flips are independent?

Mutually independent Bernoulli trials

Flipping coin

Question: A coin is biased so that the probability of heads is 2/3. What is the probability that exactly four heads come up when the coin is flipped seven times, assuming that the flips are independent?

Solution:

Mutually independent Bernoulli trials

Flipping coin

Question: A coin is biased so that the probability of heads is $2/3$. What is the probability that exactly four heads come up when the coin is flipped seven times, assuming that the flips are independent?

Solution:

Let r.v. X_i be the $i-$th flip of the coin $(i = 1, 2, \cdots, 7)$, where X_i denote whether obtain the head or not. Hence, we have

$$X_i = \begin{cases} 1, & \text{if we obtain head;} \\ 0, & \text{otherwise.} \end{cases}$$

Mutually independent Bernoulli trials

Flipping coin

Question: A coin is biased so that the probability of heads is $2/3$. What is the probability that exactly four heads come up when the coin is flipped seven times, assuming that the flips are independent?

Solution:

Let r.v. X_i be the $i-$th flip of the coin ($i = 1, 2, \cdots, 7$), where X_i denote whether obtain the head or not. Hence, we have

$$X_i = \begin{cases} 1, & \text{if we obtain head;} \\ 0, & \text{otherwise.} \end{cases}$$

Let r.v. X be # heads when the coin is flipped seven times. We have

$$X = \sum_{i=1}^{7} X_i.$$

Flipping coin Cont'd

$X = 4$ means that there are only four 1s in seven r.v.s X_i.

Flipping coin Cont'd

$X = 4$ means that there are only four 1s in seven r.v.s X_i.
The number of ways four of the seven flips can be heads is $C(7, 4)$.

Flipping coin Cont'd

$X = 4$ means that there are only four 1s in seven r.v.s X_i.
The number of ways four of the seven flips can be heads is $C(7, 4)$.
Note that $X_1 = X_2 = X_3 = X_4 = 1$ and $X_5 = X_6 = X_7 = 0$ is one of ways. Hence, we have

$$\begin{aligned} P(X_1 = 1 \wedge X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1 \wedge \\ X_5 = 0 \wedge X_6 = 0 \wedge X_7 = 0) \\ = (2/3)^4(1/3)^3 \end{aligned}$$

Flipping coin Cont'd

$X = 4$ means that there are only four 1s in seven r.v.s X_i.
The number of ways four of the seven flips can be heads is $C(7, 4)$.
Note that $X_1 = X_2 = X_3 = X_4 = 1$ and $X_5 = X_6 = X_7 = 0$ is one of ways. Hence, we have

$$\begin{aligned} P(X_1 = 1 \wedge X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1 \wedge \\ X_5 = 0 \wedge X_6 = 0 \wedge X_7 = 0) \\ = (2/3)^4(1/3)^3 \end{aligned}$$

Therefore,

$$P(X = 4) = C(7, 4)(2/3)^4(1/3)^3.$$

Binomial distribution

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

$$P(X = k) = C(n, k)p^k q^{n-k}.$$

Binomial distribution

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

$$P(X = k) = C(n, k)p^k q^{n-k}.$$

Binomial distribution

Let $B(k; n, p)$ denote the probability of k successes in n independent Bernoulli trials with probability of success p and probability of failure $q = 1 - p$. We call this function the **binomial distribution**, i.e., $B(k; n, p) = P(X = k) = C(n, k)p^k q^{n-k}$.

Binomial distribution

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

$$P(X = k) = C(n, k)p^k q^{n-k}.$$

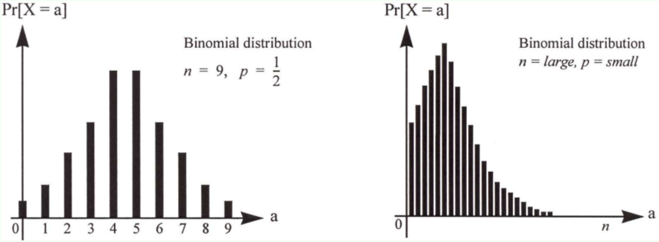
Binomial distribution

Let $B(k; n, p)$ denote the probability of k successes in n independent Bernoulli trials with probability of success p and probability of failure $q = 1 - p$. We call this function the **binomial distribution**, i.e., $B(k; n, p) = P(X = k) = C(n, k)p^k q^{n-k}$.

Note that we will say $X \sim Bin(n, p)$, and

$$\sum_{k=0}^{n} C(n, k)p^k q^{n-k} = (p + q)^n = 1.$$

Binomial distribution Cont'd



- This distribution is useful for modeling many real-world problems, such as # 3s when we roll a die n times,
- The **Bernoulli distribution** is a special case of the binomial distribution, where $n = 1$.
- Any binomial distribution, $Bin(n, p)$, is the distribution of the sum of n Bernoulli trials, $Bin(p)$, each with probability p.

Flipping coin Cont'd

Let r.v. Y be the number of coin flips until the first head obtained.

Flipping coin Cont'd

Let r.v. Y be the number of coin flips until the first head obtained.

$$
\begin{aligned}
P(Y = k) &= P(X_1 = 0 \wedge X_2 = 0 \wedge \cdots X_{k-1} = 0 \wedge X_k = 1) \\
&= \Pi_{i=1}^{k-1} P(X_i = 0) \cdot P(X_k = 1) \\
&= pq^{k-1}
\end{aligned}
$$

Flipping coin Cont'd

Let r.v. Y be the number of coin flips until the first head obtained.

$$\begin{aligned} P(Y=k) &= P(X_1 = 0 \land X_2 = 0 \land \cdots X_{k-1} = 0 \land X_k = 1) \\ &= \Pi_{i=1}^{k-1} P(X_i = 0) \cdot P(X_k = 1) \\ &= pq^{k-1} \end{aligned}$$

Let $G(k;p)$ denote the probability of failures before the $k-$th independent Bernoulli trials with probability of success p and probability of failure $q = 1 - p$.

Flipping coin Cont'd

Let r.v. Y be the number of coin flips until the first head obtained.

$$\begin{aligned} P(Y=k) &= P(X_1=0 \land X_2=0 \land \cdots X_{k-1}=0 \land X_k=1) \\ &= \Pi_{i=1}^{k-1} P(X_i=0)\cdot P(X_k=1) \\ &= pq^{k-1} \end{aligned}$$

Let $G(k;p)$ denote the probability of failures before the $k-$th independent Bernoulli trials with probability of success p and probability of failure $q=1-p$.

We call this function the **Geometric distribution**, i.e.,

$$G(k;p)=pq^{k-1}.$$

Collision in hashing

Question: Hashing functions map a large universe of keys (such as the approximately 300 million Social Security numbers in the United States) to a much smaller set of storage locations. A good hashing function yields few collisions, which are mappings of two different keys to the same memory location. What is the probability that no two keys are mapped to the same location by a hashing function, or, in other words, that there are no collisions?

Solution:

To calculate this probability, we assume that the probability that a randomly and uniformly selected key is mapped to a location is $1/m$, where m is # available locations.

Collision in hashing

Question: Hashing functions map a large universe of keys (such as the approximately 300 million Social Security numbers in the United States) to a much smaller set of storage locations. A good hashing function yields few collisions, which are mappings of two different keys to the same memory location. What is the probability that no two keys are mapped to the same location by a hashing function, or, in other words, that there are no collisions?

Solution:

To calculate this probability, we assume that the probability that a randomly and uniformly selected key is mapped to a location is $1/m$, where m is $\#$ available locations.

Suppose that the keys are $k_1, k_2, \cdots, k_n$. When we add a new record k_i, the probability that it is mapped to a location different from the locations of already hashed records, that $h(k_i) \neq h(k_j)$ for $1 \leq j < i$ is $(m - i + 1)/m$.

Collision in hashing Cont'd

Because the keys are independent, the probability that all n keys are mapped to different locations is

$$H(n,m) = \frac{m-1}{m}\frac{m-2}{m}\cdots\frac{m-n+1}{m}.$$

Collision in hashing Cont'd

Because the keys are independent, the probability that all n keys are mapped to different locations is

$$H(n,m) = \frac{m-1}{m}\frac{m-2}{m}\cdots\frac{m-n+1}{m}.$$

Recall the bounds for the same birthday problem that

$$e^{\frac{n(n-1)}{2m}} \le \frac{m^k}{m(m-1)(m-2)\cdots(m-n+1)} = \frac{1}{H(n,m)} \le e^{\frac{n(n-1)}{2(m-n+1)}},$$

Collision in hashing Cont'd

Because the keys are independent, the probability that all n keys are mapped to different locations is

$$H(n,m) = \frac{m-1}{m}\frac{m-2}{m}\cdots\frac{m-n+1}{m}.$$

Recall the bounds for the same birthday problem that

$$e^{\frac{n(n-1)}{2m}} \le \frac{m^k}{m(m-1)(m-2)\cdots(m-n+1)} = \frac{1}{H(n,m)} \le e^{\frac{n(n-1)}{2(m-n+1)}},$$

That is

$$1 - e^{-\frac{n(n-1)}{2m}} \le 1 - H(n,m) \le 1 - e^{-\frac{n(n-1)}{2(m-n+1)}}.$$

Collision in hashing Cont'd

Techniques from calculus can be used to find the smallest value of n given a value of m such that the probability of a collision is greater than a particular threshold, for example 0.5.

$$0.5 \leq 1 - e^{-\frac{n(n-1)}{2m}} \leq 1 - H(n, m).$$

Collision in hashing Cont'd

Techniques from calculus can be used to find the smallest value of n given a value of m such that the probability of a collision is greater than a particular threshold, for example 0.5.

$$0.5 \leq 1 - e^{-\frac{n(n-1)}{2m}} \leq 1 - H(n, m).$$

Hence, we have

$$n(n-1) > 2 \ln 2 \cdot m, \text{ i.e., } n > \sqrt{2 \ln 2 \cdot m} (\text{ approximately}).$$

Collision in hashing Cont'd

Techniques from calculus can be used to find the smallest value of n given a value of m such that the probability of a collision is greater than a particular threshold, for example 0.5.

$$0.5 \leq 1 - e^{-\frac{n(n-1)}{2m}} \leq 1 - H(n, m).$$

Hence, we have

$$n(n-1) > 2\ln 2 \cdot m, \text{ i.e., } n > \sqrt{2\ln 2 \cdot m}(\text{ approximately}).$$

For example, when $m = 1,000,000$, the smallest integer n such that the probability of a collision is greater than $1/2$ is 1178.

Monte Carlo algorithms

A **Monte Carlo algorithm** is a randomized or probabilistic algorithm whose output may be inaccuracy with a certain (typically small) probability.

Probabilistic algorithms make random choices at one or more steps, and result in different output even given the same input, which is different from all deterministic algorithms.

Monte Carlo algorithms

A **Monte Carlo algorithm** is a randomized or probabilistic algorithm whose output may be inaccuracy with a certain (typically small) probability.
Probabilistic algorithms make random choices at one or more steps, and result in different output even given the same input, which is different from all deterministic algorithms.

Monte Carlo algorithm for a decision problem: The probability that the algorithm answers the decision problem correctly increases as more tests are carried out.

Monte Carlo algorithms

A **Monte Carlo algorithm** is a randomized or probabilistic algorithm whose output may be inaccuracy with a certain (typically small) probability.
Probabilistic algorithms make random choices at one or more steps, and result in different output even given the same input, which is different from all deterministic algorithms.

Monte Carlo algorithm for a decision problem: The probability that the algorithm answers the decision problem correctly increases as more tests are carried out.
Step i:

$$\text{Algorithm responses} \begin{cases} \textit{true}, & \text{the answer is "true"}; \\ \textit{unknown}, & \text{either "true" or "false."} \end{cases}$$

Monte Carlo algorithm Cont'd

After running all the iterations:

Output:

$$\text{Algorithm returns} \begin{cases} \textit{true}, & \text{yield at least one "true";} \\ \textit{false}, & \text{yield "unknown" in every iteration.} \end{cases}$$

Monte Carlo algorithm Cont'd

After running all the iterations:

Output:

$$\text{Algorithm returns} \begin{cases} \textit{true}, & \text{yield at least one ``true''};\\ \textit{false}, & \text{yield ``unknown'' in every iteration.} \end{cases}$$

We will show that possibility of making mistake becomes extremely unlikely as # tests increases.

Monte Carlo algorithm Cont'd

After running all the iterations:
Output:

$$\text{Algorithm returns} \begin{cases} \textit{true}, & \text{yield at least one "true";} \\ \textit{false}, & \text{yield "unknown" in every iteration.} \end{cases}$$

We will show that possibility of making mistake becomes extremely unlikely as # tests increases.
Suppose that p is the probability that the response of a test is "true" given that the answer is "true".

Monte Carlo algorithm Cont'd

After running all the iterations:

Output:

$$\text{Algorithm returns} \begin{cases} \textit{true}, & \text{yield at least one ``true''}; \\ \textit{false}, & \text{yield ``unknown'' in every iteration.} \end{cases}$$

We will show that possibility of making mistake becomes extremely unlikely as # tests increases.

Suppose that p is the probability that the response of a test is "true" given that the answer is "true".

Because the algorithm answers "false" when all n iterations yield the answer "unknown" and the iterations perform independent tests, the probability of error is $(1-p)^n$.

Monte Carlo algorithm Cont'd

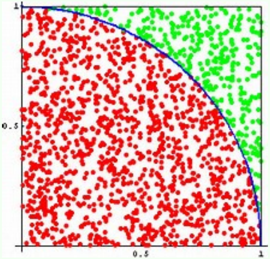
After running all the iterations:
Output:

$$\text{Algorithm returns} \begin{cases} \textit{true}, & \text{yield at least one "true"}; \\ \textit{false}, & \text{yield "unknown" in every iteration.} \end{cases}$$

We will show that possibility of making mistake becomes extremely unlikely as # tests increases.
Suppose that p is the probability that the response of a test is "true" given that the answer is "true".
Because the algorithm answers "false" when all n iterations yield the answer "unknown" and the iterations perform independent tests, the probability of error is $(1-p)^n$.
When $p \neq 0$, this probability approaches 0 as the number of tests increases. Consequently, the probability that the algorithm answers "true" when the answer is "true" approaches 1.

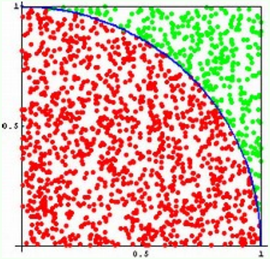
Monte Carlo II

Algorithm:



Question: How accurate of the probabilistic algorithm?
We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

Monte Carlo II



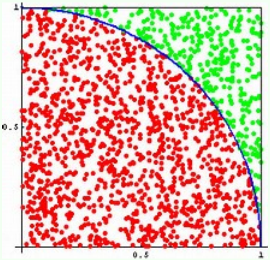
Algorithm:

Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y) | 0 \leq x, y \leq 1\}$.

Question: How accurate of the probabilistic algorithm?

We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

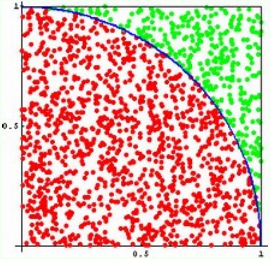
Monte Carlo II



Algorithm:
Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y)|0 \leq x, y \leq 1\}$.
Let set
$S = \{(x, y) : x^2 + y^2 \leq 1 \wedge x, y \geq 0\}$ be the circle region. And $\forall P_i \in S$, we define $I_S(P_i)$ and $I_{\Omega - S}(P_i)$;

Question: How accurate of the probabilistic algorithm?
We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

Monte Carlo II



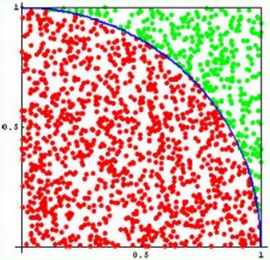
Algorithm:

Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y)|0 \leq x, y \leq 1\}$.

Let set $S = \{(x, y) : x^2 + y^2 \leq 1 \wedge x, y \geq 0\}$ be the circle region. And $\forall P_i \in S$, we define $I_S(P_i)$ and $I_{\Omega-S}(P_i)$;

$\frac{\pi}{4} \approx \frac{\sum_{i=1}^{n} I_S(P_i)}{\sum_{i=1}^{n} I_S(P_i)+\sum_{i=1}^{n} I_{\Omega-S}(P_i)}$.

Monte Carlo II



Algorithm:

Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x,y)|0 \leq x, y \leq 1\}$.

Let set $S = \{(x,y) : x^2 + y^2 \leq 1 \wedge x, y \geq 0\}$ be the circle region. And $\forall P_i \in S$, we define $I_S(P_i)$ and $I_{\Omega-S}(P_i)$;

$\frac{\pi}{4} \approx \frac{\sum_{i=1}^{n} I_S(P_i)}{\sum_{i=1}^{n} I_S(P_i) + \sum_{i=1}^{n} I_{\Omega-S}(P_i)}$.

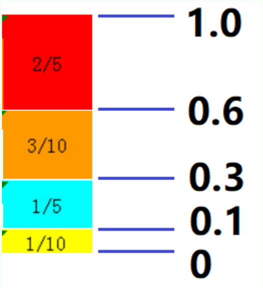
Question: How accurate of the probabilistic algorithm?
We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

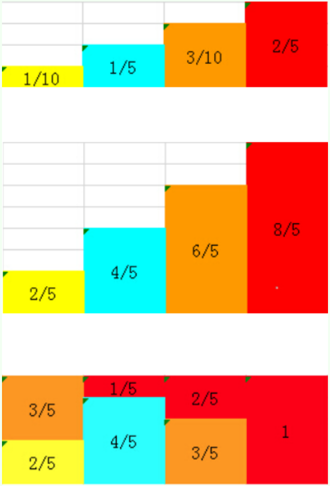
Sample with discrete distribution

How to sample from discrete distribution 0.1, 0.2, 0.3, 0.4?

Aliasing sample:

CDF sample:





$O(\log n)$ for CDF sample, and $O(1)$ for aliasing sample.

Running example

Question: 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 red ball from the first box?

Running example

Question: 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 red ball from the first box?

Solution: Let E be the event that Bob has chosen a red ball. Let F and $\overline{F}$ be the event that Bob has chosen a ball from the first box and the second box, respectively.

Running example

Question: 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 red ball from the first box?

Solution: Let E be the event that Bob has chosen a red ball. Let F and $\overline{F}$ be the event that Bob has chosen a ball from the first box and the second box, respectively.
We want to find $P(F|E)$, the probability that the ball Bob selected came from the first box, given that it is red.

Running example Cont'd

In terms of the definition of conditional probability, we have

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

Our target is to compute $P(F \cap E)$ and $P(E)$.

Running example Cont'd

In terms of the definition of conditional probability, we have

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

Our target is to compute $P(F \cap E)$ and $P(E)$. Suppose that $P(F) = P(\overline{F}) = \frac{1}{2}$. We have known that

$$P(E|F) = \frac{7}{9}, P(E|\overline{F}) = \frac{3}{7}.$$

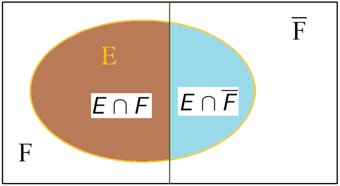
Running example Cont'd

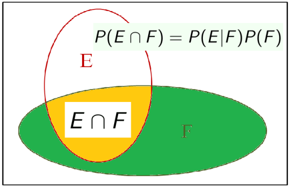
In terms of the definition of conditional probability, we have

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

Our target is to compute $P(F \cap E)$ and $P(E)$. Suppose that $P(F) = P(\overline{F}) = \frac{1}{2}$. We have known that

$$P(E|F) = \frac{7}{9}, P(E|\overline{F}) = \frac{3}{7}.$$





Running example Cont'd

Then,

$$P(E \cap F) = P(E|F)P(F) = (\frac{7}{9})(\frac{1}{2}) = \frac{7}{18},$$

$$P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}) = (\frac{3}{7})(\frac{1}{2}) = \frac{3}{14}.$$

Running example Cont'd

Then,

$$P(E \cap F) = P(E|F)P(F) = (\frac{7}{9})(\frac{1}{2}) = \frac{7}{18},$$
$$P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}) = (\frac{3}{7})(\frac{1}{2}) = \frac{3}{14}.$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.

Running example Cont'd

Then,

$$P(E \cap F) = P(E|F)P(F) = (\frac{7}{9})(\frac{1}{2}) = \frac{7}{18},$$
$$P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}) = (\frac{3}{7})(\frac{1}{2}) = \frac{3}{14}.$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.

$$P(E) = P(E \cap F) + P(E \cap \overline{F}) = \frac{7}{18} + \frac{3}{14} = \frac{38}{63}.$$

Running example Cont'd

Then,

$$P(E \cap F) = P(E|F)P(F) = (\frac{7}{9})(\frac{1}{2}) = \frac{7}{18},$$
$$P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}) = (\frac{3}{7})(\frac{1}{2}) = \frac{3}{14}.$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.

$$P(E) = P(E \cap F) + P(E \cap \overline{F}) = \frac{7}{18} + \frac{3}{14} = \frac{38}{63}.$$

We conclude that

$$P(F|E) = \frac{P(F \cap E)}{P(E)} = \frac{7/18}{38/63} = \frac{49}{76}.$$

Bayes' Theorem

Theorem

Suppose that E and F are events from a sample space Ω such that $P(E) \neq 0$ and $P(F) \neq 0$. Then

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}.$$

Bayes' Theorem

Theorem

Suppose that E and F are events from a sample space Ω such that $P(E) \neq 0$ and $P(F) \neq 0$. Then

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}.$$

Proof.

Since we have $P(F|E) = \frac{P(F \cap E)}{P(E)}$, our target is therefore to compute $P(F \cap E)$ and $P(E)$.

Bayes' Theorem

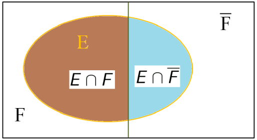
Theorem

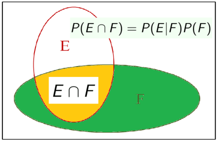
Suppose that E and F are events from a sample space Ω such that $P(E) \neq 0$ and $P(F) \neq 0$. Then

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}.$$

Proof.

Since we have $P(F|E) = \frac{P(F \cap E)}{P(E)}$, our target is therefore to compute $P(F \cap E)$ and $P(E)$.





Bayes' Theorem Cont'd

Proof

Then,

$$P(E \cap F) = P(E|F)P(F), P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}).$$

Bayes' Theorem Cont'd

Proof

Then,

$$P(E \cap F) = P(E|F)P(F), P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}).$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.

Bayes' Theorem Cont'd

Proof

Then,

$$P(E \cap F) = P(E|F)P(F), P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}).$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.
We have

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

Bayes' Theorem Cont'd

Proof

Then,

$$P(E \cap F) = P(E|F)P(F), P(E \cap \overline{F}) = P(E|\overline{F})P(\overline{F}).$$

Note that $E = (E \cap F) \cup (E \cap \overline{F})$ and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$.
We have

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

We can conclude that

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}.$$

Generalized Bayes' Theorem

Theorem

Suppose that E is an event from a sample space Ω and $F_1, F_2, \cdots, F_n$ is a partition of the sample space. Let $P(E) \neq 0$ and $P(F_i) \neq 0$ for $\forall i$. Then

$$P(F_i|E) = \frac{P(E|F_i)P(F_i)}{\sum_{k=1}^{n} P(E|F_k)P(F_k)}.$$

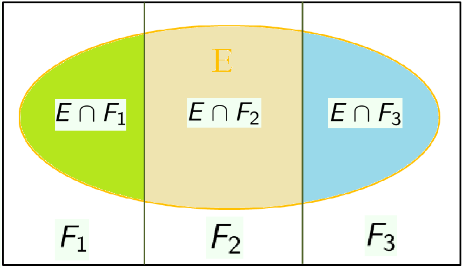
Generalized Bayes' Theorem

Theorem

Suppose that E is an event from a sample space Ω and $F_1, F_2, \cdots, F_n$ is a partition of the sample space. Let $P(E) \neq 0$ and $P(F_i) \neq 0$ for $\forall i$. Then

$$P(F_i|E) = \frac{P(E|F_i)P(F_i)}{\sum_{k=1}^{n} P(E|F_k)P(F_k)}.$$

Proof:



Diagnostic test for rare disease

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

- the probability that a person who tests positive for the disease has the disease?
- 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?

Diagnostic test for rare disease

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

- the probability that a person who tests positive for the disease has the disease?
- 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:

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. Hence, we have $p(F) = 1/100,000 = 10^{-5}$.

Diagnostic test for rare disease Cont'd

Then we also have $P(E|F) = 0.99$, $P(\overline{E}|F) = 0.01$, $P(\overline{E}|\overline{F}) = 0.995$, and $P(E|\overline{F}) = 0.005$.

Diagnostic test for rare disease Cont'd

Then we also have $P(E|F) = 0.99$, $P(\overline{E}|F) = 0.01$, $P(\overline{E}|\overline{F}) = 0.995$, and $P(E|\overline{F}) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$\begin{aligned} P(F|E) &= \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})} \\ &= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002 \end{aligned}$$

Diagnostic test for rare disease Cont'd

Then we also have $P(E|F) = 0.99$, $P(\overline{E}|F) = 0.01$, $P(\overline{E}|\overline{F}) = 0.995$, and $P(E|\overline{F}) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$\begin{aligned} P(F|E) &= \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})} \\ &= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002 \end{aligned}$$

Diagnostic test for rare disease Cont'd

Then we also have $P(E|F) = 0.99$, $P(\overline{E}|F) = 0.01$, $P(\overline{E}|\overline{F}) = 0.995$, and $P(E|\overline{F}) = 0.005$.

Case a: In terms of Bayes' theorem, we have

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}$$
$$= \frac{0.99 \cdot 10^{-5}}{0.99 \cdot 10^{-5} + 0.005 \cdot 0.99999} \approx 0.002$$

Case b: Similarly, we have

$$P(\overline{F}|\overline{E}) = \frac{P(\overline{E}|\overline{F})P(\overline{F})}{P(\overline{E}|\overline{F})P(\overline{F}) + P(\overline{E}|F)P(F)}$$
$$= \frac{0.995 \cdot 0.99999}{0.995 \cdot 0.99999 + 0.01 \cdot 10^{-5}} \approx 0.9999999$$

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.

On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.
On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?
Solution: 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 $\#$ times w appears in a message from a large set of messages known to be spam and $\#$ times it appears in a large set of messages known not to be spam.

Bayesian spam filters

Most electronic mailboxes receive a flood of unwanted and unsolicited messages, known as **spam**.
On the Internet, an **Internet Water Army** is a group of Internet ghostwriters paid to post online comments with particular content.

Question: How to detect spam email?
Solution: 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 # times w appears in a message from a large set of messages known to be spam and # times it appears in a large set of messages known not to be spam.
Step 1: Collect ground-truth Suppose we have a set B of messages known to be spam and a set G of messages known not to be spam.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be # messages containing word w in sets B and G, respectively.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be # messages containing word w in sets B and G, respectively.
Let $p(w) = n_B(w)/|B|$ and $q(w) = n_G(w)/|G|$ be the empirical probabilities that a message are not spam and spam contains word w, respectively.
Step 3: Make decision Now suppose we receive a new e-mail message containing word w. Let F be the event that the message is spam. Let E be the event that the message contains word w.

Bayesian spam filters Cont'd

Step 2: Learn parameters We next identify the words that occur in B and in G. Let $n_B(w)$ and $n_G(w)$ be $\#$ messages containing word w in sets B and G, respectively.

Let $p(w) = n_B(w)/|B|$ and $q(w) = n_G(w)/|G|$ be the empirical probabilities that a message are not spam and spam contains word w, respectively.

Step 3: Make decision Now suppose we receive a new e-mail message containing word w. Let F be the event that the message is spam. Let E be the event that the message contains word w.

By Bayes theorem, the probability that the message is spam, given that it contains word w, is

$$P(F|E) = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|\overline{F})P(\overline{F})}.$$

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(F)$, the probability that an incoming message is spam, as well as $P(\overline{F})$, the probability that the incoming message is not spam.

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(F)$, the probability that an incoming message is spam, as well as $P(\overline{F})$, 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, i.e., $P(F) = P(\overline{F}) = 1/2$.

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(F)$, the probability that an incoming message is spam, as well as $P(\overline{F})$, 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, i.e., $P(F) = P(\overline{F}) = 1/2$.

Using this assumption, we find that the probability that a message is spam, given that it contains word w, is

$$P(F|E) = \frac{P(E|F)}{P(E|F) + P(E|\overline{F})}.$$

Bayesian spam filters Cont'd

To apply the above formula, we first estimate $P(F)$, the probability that an incoming message is spam, as well as $P(\overline{F})$, 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, i.e., $P(F) = P(\overline{F}) = 1/2$.

Using this assumption, we find that the probability that a message is spam, given that it contains word w, is

$$P(F|E) = \frac{P(E|F)}{P(E|F) + P(E|\overline{F})}.$$

By estimating $P(E|F)$ and $P(E|\overline{F})$, $P(F|E)$ can be estimated by

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

Extended Bayesian spam filters

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 $P(S) = P(\overline{S})$, and that events $E_i|S$ are independent, then by Bayes theorem the probability that a message containing all words $w_1, w_2, \cdots, w_k$ is spam is

$$\begin{aligned}
P(S|\bigcap_{i=1}^{k} E_i) &= \frac{P(\bigcap_{i=1}^{k} E_i|S)P(S)}{P(\bigcap_{i=1}^{k} E_i|S)P(S) + P(\bigcap_{i=1}^{k} E_i|\overline{S})P(\overline{S})} \\
&= \frac{\Pi_{i=1}^{k} P(E_i|S)}{\Pi_{i=1}^{k} P(E_i|S) + \Pi_{i=1}^{k} P(E_i|\overline{S})} \\
&\approx \frac{\Pi_{i=1}^{k} p(w_i)}{\Pi_{i=1}^{k} p(w_i) + \Pi_{i=1}^{k} q(w_i)} = r(w_1, w_2, \cdots, w_k).
\end{aligned}$$

Naive Bayes

- Why is this called Naive Bayes?

Naive Bayes

- Why is this called Naive Bayes?
- The model employs the chain rule for repeated applications of the definition of conditional probability.

Naive Bayes

- Why is this called Naive Bayes?
- The model employs the chain rule for repeated applications of the definition of conditional probability.
- To handle underflow, we calculate $\Pi_{i=1}^{n} P(X_i|S) = exp(\sum_{i=1}^{n} \log P(X_i|S))$.

Take-aways

Conclusions

- Random variable
- Bernoulli Trials and the Binomial Distribution
- Bayes' Theorem
- Applications of Bayes' Theorem