Probability and Information Theory

Debdeep Mukhopadhyay

Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
INDIA -721302

Objectives

- **Importance of Probability**
- **Computational Security**
- **Binomial Distribution**
- **The Birthday Paradox**
- **Concept of Entropy and Information**

Importance of Probability

- We often need to answer : “how probable is the insecure event”?
  - like in our example on Coin flipping over telephone, what is the probability of Alice to create a $x \neq y$, st $f(x)=f(y)$?
  - What is the probability that Bob can guess the parity of x from f(x)?
- So, theory of probability is central to the development of cryptography.

Uncertainty of ciphers

- A good crypto scheme should produce a ciphertext, which has a random distribution
  - in the entire space of its ciphertext message
  - If it is “perfectly random”, then there is no information.
  - Like the output of the magic function, f(x) has no information about the parity of x.
  - This information or lack of information was called “uncertainty of ciphers”

Semantic Security

- **Semantically Secured:**
  - **Alice encrypts, either 0 or 1 with equal probability, and sends the resultant cipher, c to Bob as a challenge:**
  - **if Bob cannot guess without the decryption key, whether 0 or 1 was encrypted better than a random guess, then the encryption algorithm is said to be "semantically secured".**

- **That is Bob or any eves-dropper does not have an advantage over a random guess.**

Notions of security we have seen

- **Message Indistinguishability**

- **Semantic Security**

  - **But we have not talked about the computational power of the adversary…**
  - **Bounded or Unbounded**

Computational Security

- We define a crypto-system to be computationally secure if the best algorithm for breaking it requires at least N operations, where N is a very large number.
- Another approach is to reduce the problem of breaking a cryptosystem to a known problem, like “factoring a large number to its prime factors”.
- There is no absolute proof of security: *everything is relative*

Probability is a good tool

- **Definition:**
  - Probability Space: Arbitrary, but fixed set of points. Denote by S.
  - An experiment is an action of taking a point from S.
  - Sample Point: Commonly called outcome of an experiment.

Tossing an unbiased Coin

- **Two possibilities of an experiment are Head or Tail**
- **An experiment is "toss the coin for 10 times"**
- **Event is 5 times head, 5 times tail.**
- **Probability of the event is:** $\dfrac{\binom{10}{5}}{2^{10}}$

Classical Definition

- **Suppose that an experiment can yield one of n=#S equally probable points and that every experiment must yield a point. Let m be the number of points which form event E. Then the probability of an event E is:**

$$\Pr[E]=m/n$$

Statistical Definition

- Suppose that n experiments are carried out under the same condition, in which event E has occurred μ times. For a large value of n, then the event E is said to have the probability which is denoted by: $\Pr[E] \approx \mu / n$

Some Probability Rules

- **Addition Rules:**
  - Pr[AUB]=Pr[A]+Pr[B]-Pr[A∩B]
  - Mutually Exclusive: Pr[A∩B]=0
- **Conditional Probability**
  - Pr[A|B]=Pr[A∩B]/Pr[B]
- **Independent Events**
  - Pr[A∩B]=Pr[A]Pr[B]

Law of Total Probability

If $\bigcup_{i=1}^{n} E_i = S$ and $E_i \cap E_j = \Phi \ (i \neq j)$,

for any event A

$$\Pr[A] = \sum_{i=1}^{n} \Pr[A|E_i]\Pr[E_i]$$

Random Variables and their Probability Distribution

- **In cryptography, we discuss functions defined on discrete spaces.**
- **Let a discrete space, S have a countable number of points, $x_1, x_2, \ldots, x_{\#S}$**
- **A discrete variable is a numerical result of an experiment. It is a function defined on a discrete sample space.**

Random Variables and their Probability Distribution

- **Let S be a discrete probability space and X be a random variable (r.v).**
- **A discrete probability function of X is of type, S→R (set of reals), provided by a list of probability values:**

  Pr[X=x_i]=p_i (i=1,2,…,#S), st

$$
\begin{aligned}
&i) \quad p_i \geq 0; \\
&ii) \sum_{i=1}^{\#S} p_i = 1
\end{aligned}
$$

Uniform Distribution

- **Most frequently used distribution is:**

  Pr[X=x_i]=1/(#S), i=1,2,…,#S

  Then X is said to follow a uniform distribution.
- **Notation: p $\in_U$S**
  - **Choose p uniformly from S**

Binomial Distribution

- **Suppose an experiment has two possible outcomes, HEAD (success) or TAIL (failure)**
- **Repeated independent such experiments are called Bernoulli Trials**
- **Pr[H]=p, pr[T]=1-p**

$$\text{Pr[k "success" in n trials]}=\binom{n}{k} p^k(1-p)^{n-k}$$

No of ways of choosing k points out of n

Binomial Distribution

- **If a random variable Y, takes values, 0, 1, …, n and for values 0<p<1, and**

$$\Pr[Y=k]=\binom{n}{k} p^k(1-p)^{n-k}$$

then Y follows Binomial Distribution.

A useful result

Let ε be an event in a probability space X, with Pr[ε]=p>0. Repeatedly, we perform the random experiment X independently. Let, G be the expected number of experiments of X, until ε occurs the first time. Prove that: E(G)=$\frac{1}{\text{p}}$

$$\Pr[G=t]=(1-p)^{t-1}p \Rightarrow E(G)=\sum_{t=1}^{\infty} tp(1-p)^{t-1}=-p\frac{d}{dp}\sum_{t=1}^{\infty}(1-p)^{t}=\text{-p}\frac{d}{dp}(\frac{1}{p}-1)=\frac{1}{p}.$$

Law of large Numbers

- **Repeat a trial for a large number of time (n→infinity) and note the number of success.**
- **After a point the number of success will remain constant and equal to np (often referred to as the Expected number of success) or the <u>Expectation</u> of the r.v.**

$$\lim_{n\to\infty}\Pr[|\frac{\xi_n}{n}-p|<\alpha]=1$$

α: small fixed number

The Birthday Paradox

- **Consider a function, f: X→Y, where Y is a set of n elements.**
  - **eg, consider this class of students form X. Let Y denote the birthday, say 15th September is the birthday of a person X.**
  - **thus, Y is the 365 days of a year (let us consider that no-body in the class was born on 29th February)**

The Problem

- **Choose k pair-wise distinct points from X uniformly.**
- **Define, collision to be the event when for $i \neq j$, $f(x_i)=f(x_j)$**
- **Also, check from the corresponding $f(x_i)$'s, when a collision occurs.**
- **Clearly, the probability of a collision increases if k is increased.**

- **Question: What is the least value of k, so that the probability of a collision is more than say, ϵ?**

Let us compute for the class

- **Probability of no collision in k persons in the class is:** $(1-\frac{1}{365})(1-\frac{2}{365})...(1-\frac{k-1}{365})=\prod_{i=1}^{k-1}(1-\frac{i}{365})$
- **For a large n and a small x,**

$$(1+\frac{x}{n})=e^{x/n}$$

- **So, Pr of no collision is,**

$$\prod_{i=1}^{k-1}(1-\frac{i}{365})\approx\prod_{i=1}^{k-1}e^{-i/365}=e^{-\frac{k(k-1)}{730}}$$

Let us compute for the class

- **Probability of a collision is:** $1-e^{-\frac{k(k-1)}{730}}$
- **Let this be Є=0.5**
- **Thus,**

$$1-e^{-\frac{k(k-1)}{730}}=0.5$$

$$\therefore\frac{k(k-1)}{730}=\ln(2)$$

$$\therefore k^2-k=730\ln(2)$$

$$\therefore k\approx\sqrt{730\ln(2)}\approx 23$$

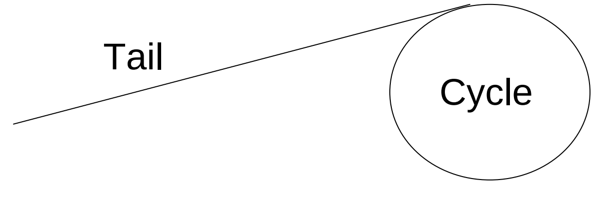
Thus, in a random room of 23 people, the probability that there are two persons with the same birthday is 0.5 !!! Seems to be a paradox

Applications of the Paradox

- **Deciding the bit length of Hash functions.**
- **Digital Signature Schemes are more than 128 bits.**
- **Index Computation (probabilistic) algorithms to solve the Discrete Logarithm Problems.**

Cycle Finding Algorithms

- **Consider a function, F from S to itself**
- **Starting from X_0 in S generate a sequence by using $X_{i+1}=F(X_i)$**
- **Goal is to find a collision, $X_i=X_j$**



The Birthday Approach

- **Note if F is random, the Birthday Paradox comes into play and we expect a collision after $2^{n/2}$ points, if S has 2^n points.**
- **Assume that the cycle's structure is:**
  - **a tail from X_0 to X_{s-1}**
  - **a loop from X_s to X_{s+l}**
- **How to detect the cycle?**

A Tree based Approach

- **Start storing the sequence elements in a binary search tree, as long as there is no duplicate.**
- **Thus, the first duplicate occurs when X_{s+l} is to be inserted, as then already X_s is in the tree.**
- **Time Complexity: O((s+l)log(s+l))**
- **Space Complexity: O(s+l)**
- **Running time is optimal.**
- **Space requirement is high.**

Floyd's Cycle Finding Algorithm

- **Define $Y_0=X_0$ and $Y_{i+1}=F(F(Y_i))$**
- **Input initial sequence X_0 and max iterations M**

$x = X_0, y = X_0$
for i from 1 to M do
  $x = F(x)$
  $y = F(F(y))$
  if $x == y$
   Output 'Collision between i and 2i'
   exit
  end if
end for
output Failed

Measuring Information

- **L={a_1,a_2,…,a_n} : Language of n different symbols.**
- **Independent probabilities:**

  Pr[a_1],Pr[a_2],…,Pr[a_n]
- **Probabilities satisfy:** $\sum_{i=1}^{n} \Pr[a_i] = 1$

Entropy

- **Entropy of the source, S:**

$$H(S) = \sum_{i=1}^{n} \Pr[a_i] \log_2 \left(\frac{1}{\Pr[a_i]}\right)$$

- **Number of bits required per source output**

Properties of Entropy

- **If S outputs a_1 with probability 1:**
  H(S)=0
- **If S outputs n symbols with equal probability 1/n, that is S is a source of a uniform distribution:**

$$H(S) = \frac{1}{n} \sum_{i=1}^{n} \log_2 n = \log_2 n$$

- **H(S) can be thought as the amount of uncertainty or information in each output from S.**

Points to Ponder

- **Suppose that four digit PINs are randomly distributed. How many people must be in a room such that the probability that two of them have the same PIN is at least ½ ?**

References

- **W. Mao, “Modern Cryptography: Theory and Practice”, Prentice Hall**
- **A. Joux, “Algorithmic Cryptanalysis”, CRC**
- **Johannes A. Buchmann, “Introduction to Cryptography”, Springer**

Next Days Topic

- **Classical Cryptosystems**