CS 4376/5376

Homework Assignment #1

Probability Theory

DUE: Sunday, Sept 13 at 11:59 PM

The assignment should be turned in using blackboard as either a word document or pdf file. Scanned work is acceptable, as long as it is legible.

1) A coin is flipped three times. Event A consists of observing exactly 2 heads. Event B consists of observing one or more tails.

a) List the members of event A

* {HHT, THH, HTH}

b) List the members of event B

* {HHT,THH,HTH,TTH,THT,TTT,HTT}

c) Find the union of event A and B

* {HHT,THH,HTH,TTH,THT,TTT}

d) Find the intersection of event A and B

* {HHT, THH, HTH}

e) Find P(A|B) => P(A and B) / P(B) => 3/8/7/8 => 0.4285

2) Let P(A)=0.3, P(B)=0.7, and P(A ^ B) = 0.2. Find the probability that:

P(A)\*P(B) = 0.21 which is not equal to P(A and B) so dependent

a) Either A or B will occur: P(AvB) = P(A) + P(b) – P(A and B) => 1-0.2 =0.8

b) Neither A or B will occur: 1-P(AvB) => 0.2

c) A will occur, and B will not occur=> P(A and !B) => P(A) – P(A and B) => 0.3-0.2 => 0.1

d) A will occur, given that B has occurred

P(A|B) => P(A and B) / P(B) => 0.2/0.7 => 0.2857

e) A will occur, given that B has not occurred: P(A|!B) = P(A and !B)/P(!B) => P(A and !B)/1-P(B)=> 0.1/0.3 => 1/3

3) Consider the following joint probability distribution over two events. Answer the following questions based on this table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Left | Center | Right |
| Top | 0.25 | 0.1 | 0.15 |
| Middle | 0 | 0.1 | 0.05 |
| Bottom | 0.15 | 0.05 | 0.15 |

a) What is the probability of Right occurring, P(R)? 0.35

b) What is the probability of Center and Top occurring, P(C ^ T)? 0.1

c) What is the probability of Top and either Left or Center occurring,

P(T ^ (L v C))? 0.35

d) What is the probability of Center given Top, P(C | T)? P(C and T) / P(T) => 0.1/0.5 => 0.2

e) What is the marginal probability distribution over the rows (Top, Middle, Bottom)? Top => 0.5, Middle=> 0.15, Bottom=>0.35

f) Suppose you get a value based on which row occurs in a sample, with the following values: Top = 30, Middle = 10, Bottom = 50. What is the expected value of sampling based on the distribution shown in the table and these values? 30\*0.5 + 10\*0.15+50\*0.35 => 15+1.5+17.5 => 34

4) Compute the expected value of a card drawn at random from a shuffled desk; aces count as 1, all face cards count as 10, and numeric cards count as the value shown. Disregard the suit.

Expected Value = value of card \* p(of card)

E[aces] = 4\*(4/52) => 16/52=> 0.3076

E[all\_numeric] = (2+…+10) \*(4/52)=>4.1538

E[face\_cards] = (10)\*(12/52) => 2.30

E[Deck] => 0.3076+4.1538+2.3076=>~6.769

5) You are selling security insurance services to a small business and wish to know how to price your policy. In particular, you will reimburse the client for the costs of any attack that results in more than $10,000 of damage, up to a limit of $1,000,000. You know based on your data that attacks that result in more than $1,000,000 occur during a given year no more than 5% of the time, and attacks costing more than $250,000 but less than $1,000,000 occur no more than 15% of the time. How much should you charge the client on a yearly basis to ensure that you do not lose money on the insurance policy in expectation? Note that your information is incomplete, so you should make pessimistic assumptions where necessary to guarantee your profits.

* Assumption: The missing probability of 80 percent goes to attacks that cost between 10,000 and 250,000 (due to pessimistic assumptions)
* Assumption: I assume every range of cost will be determined by the upper end of cost, like in the case between 250,000 and 1,000,000 the price is determined by 999,999 (due to pessimistic assumptions)

Two interpretations:

Interpretation 1: You don’t pay if the attack surpasses $1M

* More than a million=> 0(cost) \* 0.05
* Between $250,000 and $1M=> 0.15\*999,999 => 150,000
* Between $10,001 and $250,000=> $250,000\* 0.8=> 200,000
* Charging fee to not lose money: $350,000

Interpretation 2: You pay $1M if the attack surpasses $1M

* More than a million=> 0.05 \*1,000,000 => 50,000
* Between $250,000 and $1M=> 0.15\*999,999 => 150,000
* Between $10,001 and $250,000=> $250,000\* 0.8=> 200,000
* Charging fee to not lose money: $400,000

6) Consider the following scenario in email filtering. You know based on your data that 10% of messages sent out are phishing attempts. Now, you look for a few keywords to improve your estimate of whether a particular message is a phishing email. In particular, you see two interesting words in a message you are considering that might indicate that the message is or is not a phishing email. These words and the likelihood that they appear in both a phishing and non-phishing message are listed below. Apply Bayes rule to update the probability that the message is a phishing email once for each word and give the final probability that the message is a phishing email. Note that your prior for the second update is based on the posterior probability after applying Bayes rule for the first word.

Phishing Not Phishing

|  |  |  |
| --- | --- | --- |
| Urgent | 0.15 | 0.02 |
| Password | 0.10 | 0.01 |

Following:

Diagram

Description automatically generated

* Word: Urgent
  + P(Phising|Urgent) = P(Urgent|Phising)\*P(Phishing)/P(Urgent)
  + P(Phising|Urgent) = (0.15/0.25)\*0.1/0.12 => 0.5
* Word: Password
  + P(Phising|Password) = P(Password|Phising)\*P(Phishing)/P(Password) =>
  + P(Phising|Password) = (0.10/0.6) \*0.1/0.11 => 0.15
* Final:
  + Probability that the message is a phising message is 0.65 since the words are independent, we simply can add the joint probabilities and obtain a probability for the message.

8) (GRAD STUDENTS ONLY) Many authentication systems use biometrics (e.g., a fingerprint) to identify individuals. Suppose that a system is designed to detect N different “features” of a fingerprint, and each of these features can take on D different values. Further, assume that for any given fingerprint these values are all equally likely for every feature. Suppose you have P individuals randomly sampled from the population. N, D, and P are variables. For given values of these variables, you want to know how likely it is if you have a sample of P individuals that at least one of them will be able to authenticate as someone else in the sample.

1. Write down an equation you can use to approximate the solution to this problem. (*Hint: this is very similar to detecting a hash collision; use Google).*

*\*Approximation which becomes better as we increase the size of possible values set*

Equation: Text

Description automatically generated

K is the number of samples that will exist, and N is the number of options that can exist

1. What is the likelihood for N=10, D=5, and P=100?

P = 100 => k

DN => Number of options => 510

Likelihood = 1 – e-100(99)/5^10=> 0.0005

1. For N=10, D=5, and P=1,000?

P = 1000 => k

DN => Number of options => 510

Likelihood = 1 – e-1000(99)/5^10=>0.049

1. For N=20, D=10, and P=10,000?

P = 10,000 => k

DN => Number of options => 1020

Likelihood = 1 – e-10000(99)/10^20=>4.999^-13

9) (GRAD STUDENTS ONLY) We wish to transmit an *n*-bit message to a receiving agent, which is the case for the low-level messaging on most computer networks. Unfortunately, no network technology is perfect, and it is possible for bits to be corrupted (i.e., flipped from the intended value). We want to analyze the impact of this, and understand how using parity can protect against these errors.

Suppose that the bits in the message are independently corrupted (flipped) during transmission with probability *e.*  With an extra parity bit sent along with the original information, a message can be corrected by the receiver if at most one bit in the entire message (including the parity bit) has been corrupted. Suppose we want to ensure that the correct message is received with probability at least 1 – *d*. What is the maximum feasible value of *n*? Calculate this value for the case *e* =0.0002 and *d* = 0.01. Note that the total length of the message including the parity bit is *n+1.*

My reasoning for this problem is simple. Knowing that every bit can be independently corrupted with probability 0.0002 means there is a 0.9998 probability of not being corrupted. Hence, all we have to find is how many bits are needed to be multiplied by each other to reach the desired threshold of security that we received the message correctly – based on the theory that P(A and B) = P(A)\*P(b)- in this case then we say:

(0.9998)n = 1-d => (0.9998)n = 0.99

I used wolfram alpha to perform the calculation but the calculation can be done with logs => n\*log\_10(0.9998) = log(0.99) = > n= log(0.99)/log(0.9998) => ~51 bits:

Graphical user interface

Description automatically generated