Homework 1: Warm Up

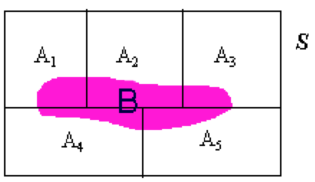
Statistical Foundations

COSC 6342: Machine Learning

Instructor: Arjun Mukherjee

Part I: Basic Probability Review for Machine Learning

Q1: Prove Bayes' rule for the following partition $A_1, A_2, \dots A_5$ of the sample space S and for the set B (shaded in purple), i.e., show that
$P\ \ A_i|B\ \ = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{5}(P(B|A_j)P(A_j))}$: [4 points].



Hint: Recall that the definition of conditional probability is $P\ \ A|B\ \ = \frac{P(A\cap B)}{P(B)}$. Use it on both LHS and RHS of the Bayes' identity.

Q2: Two biased coins, one with $P\ \ Head\ \ = u$ and another with $P\ \ Head\ \ = w$ are tossed together and independently. [2+3+5 points]

(a) Compute the following in terms of u and : $p_0 = P(0\ Head\ occurs), p_1 = P(1\ Head\ occurs), p_2 = P(2\ Head\ occurs)$.

(b) Is it possible to choose valid value of u and w so that we have $p_0 = p_2$? What values of u and w ensure $p_0 = p_2$?

(c) Is it possible to choose valid value of u and w so that we have $p_0 = p_1 = p_2$? Why or why not?

Q3: Two players A and B alternatively and independently flip a coin and the first player to obtain a head wins. Assume that A starts the game (flips first) and the coin is fair, i.e, $P(Head)\ \ = 1/2$. [2+3+2+3 points]

(a) What is the sample space of this experiments/game? For notational convenience, define a Head and Tail outcome of the toss when A flips as H_a and T_a respectively. Similarly for B, H_b and T_b. Express the sample space in terms of H_a, H_b, T_a, T_b.

(b)What is the probability that A wins the game? Hint: Recall from high school algebra that the summation of an infinite geometric series, $S = a + ar + ar^2 + \cdots = \frac{a}{1-r}$ where $a > 0$ and $0 < r < 1$.

(c) What is the probability that B wins the game?

(d) Suppose that the coin is biased with $P\ \ Head\ \ = p$. What is the probability that A wins (in terms of p)?

Q4: Suppose 5% of men and 0.25% of women in some African tribe are color blind. A random person is chosen form the tribe and was examined to be color blind. What is the probability that the person is male? We also know that the tribe is female dominated and there are twice as more females as males, i.e., $P\ \ M\ \ = \frac{1}{3};\ P\ \ F\ \ = 2P\ \ M\ \ = \frac{2}{3}$. [4 points]. Hint: Apply Bayes rule.

Q5: Consider an antique telegraph system for transmitting natural language. It has some coding scheme which converts English text to a sequence of dots $(\cdot)$ and dash $(-)$ such that all messages have dots and dashes in the proportion of 3:4. Being so old, it invariably causes some erratic transmissions! Specifically, it is known to cause a dot to become a dash with probability 1/4 (i.e., $P\ -\ rcvd \mid \cdot\ sent) = 1/4$) and to cause a dash to become a dot with probability 1/3 (i.e, $P\ \cdot\ rcvd \mid -\ sent) = 1/3$).

(a) What is the probability that a dot was received given that a dot was sent?

(b) What is the probability that a dash was received given that a dash was sent?

(c) What is the probability that a dash was sent given that a dash was received? [2+2+4 points]

Q6: Standardized tests (with multiple choice questions having one correct answer) revel an interesting application of basic probability. Suppose a test has 20 questions, each question having 4 choices/options (with exactly one correct answer). If a student guesses each question, then this process/experiment can be modeled as a sequence of 20 independent events. Each event is a Bernoulli trial with a success probability of 1/4. What is the probability that the student gets at least 10 answers correct out of the 20 questions assuming that he is guessing the answers of all the questions? [5 points]

Hint: This can be solved by directly applying the Binomial distribution's result for prob. Of k successes in n Bernoulli trials.

Q7: Consider a sequence of independent coin flips with $P\ Head\ = p$ (i.e., Bernoulli trials). Define the random variable,

$X = The\ length\ of\ the\ run\ started\ by\ the\ first\ trials$. For e.g., $X = 3$ if either TTTH or HHHT is observed. What is the probability distribution of X? i.e., What values x can the random variable X take and what are the probability of $P(X = x)$ for all allowable values of X? [4 points]

Q8: A couple decides to have children until a daughter is born. What is the expected number of children for this couple assuming the genetic/environmental effects dictate that the probability of a girl child is p? You are only required to provide the correct algebraic expression for the expected number of children. You need not solve the expression. [4 points]

Hint: First identify the sample space and assign probabilities to each event. Then define the random variable X as the number of children and compute $E[X]$.

Q9: Suppose you enter a chocolate chip cookie factory. You are told that the random variable, X = number of chocolate chips dropped on a cookie has a Poisson distribution with an average of 5 chocolate-chips/cookie, i.e., X ~ $Poisson(\lambda = 5)$. Compute the probability that a random cookie has at least 2 chips. You are only required to provide and expression and need not solve it.
[5 points]
Hint: Compute $1 - P(at\ most\ 1\ chips\ on\ a\ cookie)$ as this avoids infinite summation and is easier to evaluate.

Part II: Using Code Simulation to arrive at answers to Real world Stochastic Problems

Here we learn solving real-world stochastic problems using computer programming. For instance, if we have to solve the following stochastic problem:

Determine the probability of obtaining at least a six in 4 tosses of a fair die.

We would set up the following experiment (programmatically). Below is a barebones BASIC script that leverages a simple Random Number Generator (which is similar to what we see in our recent HLL e.g., Math.Random() in Java or a similar language).

Take a fair die and toss it 4 times. Note whether a six has turned up in the four tosses. If at least one six is observed, then count it as a "yes" or "1", if no six is observed in any of the four tosses, then count it as a "no" or "0". For a fairly reliable answer, the procedure must be repeated N times where N is at least a hundred or even a thousand. The probability of obtaining at least a six in four tosses of a die is then the number of "yes's" or "1's" divided by N.

```
        RANDOMIZE TIMER
        INPUT "Number of trials"; N
        S=0
        For I=1 to N
                For J=1 to 4
                U=INT(RND*6+1)
                If U=6 THEN 100
                Next J:  GOTO 200
100     S=S+1
200     NEXT I
        PRINT "Probability of at least 1 six in 4 tosses in a die ="; S/1000
        END
```

We run several experiments using a computer simulation and can arrive at the answer. Using similar analogy solve these empirically using an actual computer simulation program (you can use any language of your choice as long as you arrive at the correct numerical probability answer within 2 decimal places):

Q10: What is the probability that in a randomly selected group of N people (say N=25 people), there will be at least 2 people with the same birthday?

Q11: Alexendar's Dilemma - I owe my son Alexander $50 per month for doing chores around the house. Instead of giving him the $50, each month I will let him reach into a bag which contains a $100 bill and five $10 bills, and draw two of the bills. Should Alexander go with this scheme, or in other words, is my scheme fair? If Alexander is a risk taker, then he will have a chance of making $110 instead of $50 per month if he is lucky. On the other hand, if he is risk averse, then he figures that he will more likely be getting $20 rather than $50 per month. So the question is: In the long run, will he be better off taking the $50 or go with my scheme?

Q12: Write a pseudocode to simulate a biased coin toss with a probability of occurrence of an head as p. Specifically write the function/method `BiasedCoinToss(p)` which return Head with probability p and Tail with probability $(1-p)$. You are given the method/function `Math.rand()` which returns you a uniformly distributed random variable in [0, 1], i.e., $X =$ `Math.rand()` and $X \sim Uni(0,1)$. [5 points]