Homework 2

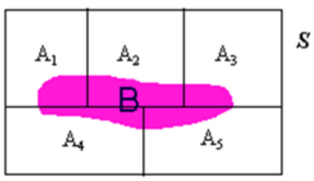
Statistical Foundations

Natural Language Processing

Instructor: Arjun Mukherjee

Q1: Prove Bayes' rule for the following partition $A_1, A_2, \ldots A_5$ 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: Define $f(x) = \begin{cases} \frac{1}{2}\lambda e^{-\lambda x}; x \geq 0 \\ \frac{1}{2}\lambda e^{\lambda x}; x < 0 \end{cases}$

where λ is a fixed positive constant. Show that $f(x)$ is a valid PDF. [4 points]

Q8: 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]

Q9: 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 $E[X]$.

Extra credit [10 points]: Using the expression computed above for the expected number of children, set p = 1/2 and sum the series using a simple program assuming the maximum number of children to be 1000. What is the expected value? What happens if we set p = 1/4, what is the expected value? What is the expected number of children when we set p = 1/10? From these can you figure out the analytical/closed form of $E[X]$?

Q10: A standard drug is known to be effective 80% of the cases when it was administered to schizophrenia patients. A new drug was once tested on a sample of 100 schizophrenia patients and found to be effective in 85 cases. We want to know whether the new drug is more effective than the standard drug.

Hint: Suppose we assume that the new drug is equally effective as the standard drug. Now the experiment of testing the drug on a sample of 100 patients can be modeled as Bernoulli distribution with repeated Bernoulli trials with success probability 0.8, i.e., if $X = \#\ of\ effective\ cases$, then $X \sim Bin(n = 100, p = 0.8)$. Now assuming the new and standard drug are equally effective, let us compute the probability of observing 85 or more successes as follows.

$$P(X \geq 85) = \sum_{x=85}^{100} \binom{100}{x} .8^x .2^{100-x} = .1285.$$

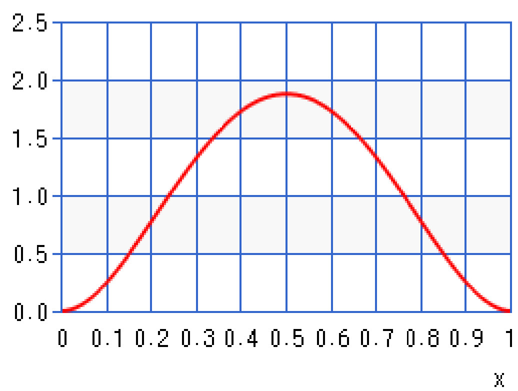
From the above result, can you say that the new drug is more effective than the standard drug? Why or why not? Justify your reasoning [4 points].

Q11: 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., $X \sim Poisson(\lambda = 5)$. 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.

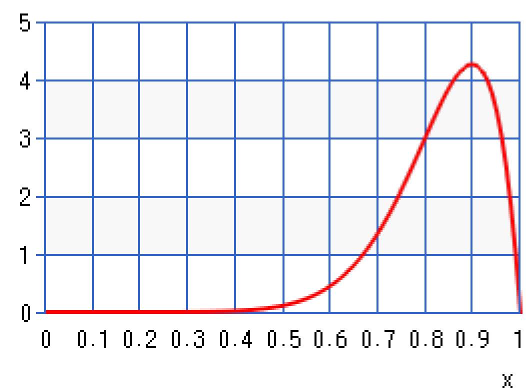
Q12: Write a pseudocode to simulate a biased coin toss with a probability of occurrence of an head as p. Specifically write the function/method `BiasedCoinToss(p)` which return Head with probability p and Tail with probability $(1-p)$. You are given the method/function `Math.rand()` which returns you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

Q13: Write a pseudocode to simulate the categorical distribution. Specifically, simulate Cat (4, <0.1, 0.2, 0.4, 0.3>), i.e., write the method for `int SimulateCategorical(int n=4, double [] dist = {0.1, 0.2, 0.4, 0.3})` which should return 1 with probability 0.1, 2 with probability 0.2, 3 with probability 0.4 and 4 with probability 0.3. 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 $X \sim Uni(0,1)$. [8 points]

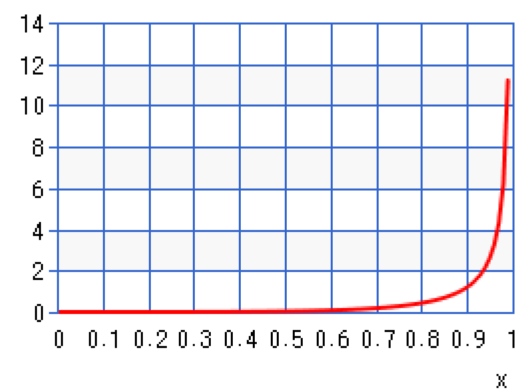
Q14: Recall that if $X \sim Beta(\alpha, \beta)$, then the mean value of the random variable is given by $E[X] = \frac{\alpha}{\alpha+\beta}$. Using this result which of the following density plots (PDF) on the y-axis for different values of x, best describes the distribution $Beta(\alpha = 0.1, \beta = 6)$? Provide a justification for your choice. [10 points].



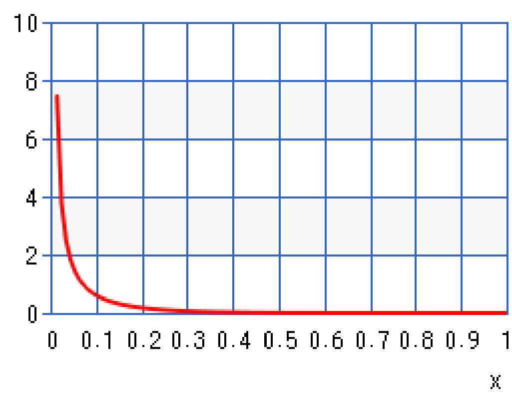
(Plot A)



(Plot B)



(Plot C)



(Plot D)