CSE 5243: Homework #1

Deadline: **11:59PM on 09/14/2022**.
No late submissions will be accepted.

Instructions and Notes.

1. Please submit an electronic copy of your homework on Carmen. **Only** when there is something wrong with Carmen, email it to the instructor (sun.397@osu.edu).
2. Instead of just giving the final answers, provide the intermediate steps for each problem so that you can still get some points even if the final answers are wrong.
3. This homework will take up 10% of your final grade.

Problem 1. (15 points)

(1) Assume a biased die lands on each face with the following probabilities:

face	1	2	3	4	5	6
P(face)	0	.1	.3	.2	0.1	0.3

Assume you win a game if the die shows even. What is the probability that you win? Is this better or worse than a fair die (i.e., a die with equal probabilities for each face)? (5 points)

(2) Recall that the expected value $E[X]$ for a random variable X is $E[X] = \sum_{x \in Values(X)} x * P(X = x)$, where $Values(X)$ is the set of values X may take. Similarly, the expected value of any function f of random variable X is $E[X] = \sum_{x \in Values(X)} f(x) * P(X = x)$.

Now consider the function below (called the "indicator function").

$$\delta(X = a) := \begin{cases} 1 & \text{if} X = a \\ 0 & \text{if} X \neq a \end{cases}$$

Let X be a random variable which can have values 2, 4 or 8 with probabilities p_2, p_4 and p_8 respectively. Calculate $E[\delta(X = 8)]$. (5 points)

(3) Consider the coin toss problem. Suppose we observe that k of n tosses are

Heads. Define p as the probability of obtaining *Heads* in a single toss of the coin. Which value of p makes this outcome *most likely*? Show the Maximum Likelihood Estimation procedure how you get the estimate of p. (5 points)

Problem 2. (15 points)

Alice living in place X goes to see the doctor for a cough. She is asked to do a blood test for swine flu, which we assume in this problem affects 3 in $10,000$ people in place X. The test is 97% accurate in the sense that the probability of a false positive is 3%. The probability of a false negative is 2%. Alice tests positive.

(a) What is the new probability that she has swine flu? Show how you get the answer. (10 points)

(b) Now imagine that Alice visited place Y for vacation recently, and it is known that 1 in 100 people who visited place Y recently come back with swine flu. Given the same test result as above, what is your estimate for the probability Alice has the disease? (5 points)

Problem 3. (35 points)

A flu is going around and it is believed that 3 in 1,000 people now have it. John just had a flu test and the result was positive. The test can accurately identify 97% of patients who have flu. If a patient doesn't have flu, 99% of the time the test result will be negative.

1. What is the probability that John has flu? Show how you get to the answer. (5 points)
2. John didn't believe the test result and just had the same test one more time. The result was still positive. Now what is the new probability that John has flu? (15 points)
3. What if the result of the second test was negative? (15 points)

Problem 4. (20 points)

Let X be a random variable denoting age. Consider a random sample of size n = 20. $X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$.

1. Find the mean, median, and mode of X. (10 points)
2. Let us use the normal distribution to model the random variable X. Write down its probability density function (use the sample mean and standard deviation). (10 points)

Problem 5. (15 points)

Similarity/distance between data points plays an important role in data analysis. However, the results can vary depending on the similarity/distance measure used, and in practice one should choose a measure that works best for the specific type of data and analysis under investigation.

Suppose we have the following 2-D data set:

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

1. Consider the data as 2-D data points. Given a new data point, $x = (1.4, 1.6)$ as a query, rank the points based on similarity (most similar/closest ones first) with the query using Euclidean distance, Manhattan distance, Jaccard similarity, and cosine similarity. (8 points)

2. Normalize the data set to make the Euclidean norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points. (7 points)