Indian Institute of Technology Mandi
भारतीय प्रौद्योगिकी संस्थान मण्डी

IC252-Data Science 2
Assignment- 06

General instructions:

- Utilize Python programming language for implementation.
- Ensure the program is well-documented to enhance comprehension.
- Employ functions and loops for efficient code organization.
- Implement error handling to manage invalid inputs or unexpected scenarios.
- Optimize the code for performance and readability where applicable.

1. (a) Calculate the entropy of a fair coin. Now suppose that the coin is biased, i.e., probability of head is not equal to 0.5. Plot the entropy curve where $x-$axis represents the probability of head and $y-$axis is the corresponding entropy.

   (b) Generate and plot two Gaussian distributions with different mean and variance. Calculate the KL divergence and cross-entropy between these two distributions. Repeat this experiment for different mean and variance, and observe the value of KL divergence and cross-entropy when these two distributions:

   i. overlap each other,
   ii. partially overlap,
   iii. do not overlap.

2. (a) Simulate a random number generator for the following distributions:

   i. Uniform distribution
   ii. Normal distribution
   iii. Truncated exponential distribution

   Generate a sample dataset of 1000 points for each case. Plot the histogram of the samples and the density function of the given distributions in a single subplot.

   (b) Let

   $$f_X = \frac{1}{40}(2x+3), \quad 0 < x < 5$$

   be a density function. Generate a random number simulator for this density function and sample 1000 random draws. Plot the graph of given density function and histogram plot of the drawn samples in a single figure.

3. Generate atleast five different random samples of size n from Normal distribution(where the mean and standard deviation are given as input by the user) and

   (a) Plot any three random samples in a single plot and compare them with the parent distribution to visually assess their similarities or differences.

   (b) Calculate the mean and variance of the random samples and compare them with the statistics of the parent distribution.

   (c) Observe how the characteristics of the random samples change as the sample size n increases. Check for any trends or patterns observed in the plots and statistical summaries.

4. Repeat the same procedure as asked in $Q3$ by sampling from Exponential distribution, where the parameter λ is given as input by the user.

5. Let the marks of students in a class are normally distributed with mean 65 and standard deviation 15. Then,

   (a) Generate five random samples of size 50 and calculate their statistics. Also, plot the histogram of these samples and the total population in a single graph.

   (b) Again generate random samples of size 100, 150 and 250, and observe their statistics and histogram.

(c) Think of any other sampling strategy different form random sampling. Use it to generate samples of different sizes from the population. Calculate their statistics and plot the histogram.

(d) Compare the statistics of the random sampling and the sampling from part *c*. Are they close to what is in the population? Is there any bias one way or the other? Is the variance too high or too low?