Data Science
CSE558
Assignment - 1

Submit by 2359 of 20/09/24 Maximum marks: 50

- **Form a group of two person and submit your assignment. Only one of the group member should submit the assignment before deadline.**
- **you are encouraged to discuss with other groups. Mention their names in your submission. If you have used the internet to understand the solution you are writing, then mention the URL.**
- **Write answers and code of your own; do not copy from others. We will follow the standard plagiarism policy, which can be found link**
- **Clearly state any extra information (such as assumption or encoding) used to reach your answers from the given questions. Write all the steps that you followed.**
- **Prepare one zip file with all your answers to the theory questions (pdf from latex/word), files of code, saved data and plots. Name it "roll_no-DSA1", e.g., "20001-DSA1". Submit the zip file through Google Classroom. A delay in submission would cost you deduction in marks as per class policy. You will be graded based on a viva. If you submit images of your handwritten notes then you will be evaluated on at max 50% of the full marks for the question.**

Best wishes!

---

1. Understand the features of the dataset called Auto MPG that can be found here. Download the dataset "Auto MPG" from this excel file. Here, the last feature, 'car name', has been removed.
   (a) For discrete attributes, apply a one-hot encoding, and for non numeric ordinal attributes, apply integer mapping. Handle outliers/noise and save this in a file. **(1)**
   (b) Let $D = \{x_1, x_2, \ldots, x_n\}$ be n objects consisting of d features, i.e., for every $1 \leq i \leq n$, $x_i = [x_{i1}, x_{i2}, \ldots, x_{id}] \in \mathbb{R}^d$. The variance σ^2 of D is defined as
   $$\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \bar{x}\|_2^2,$$
   where for any $a = [a_1, \ldots, a_d] \in \mathbb{R}^d$, the $a^T a = \|a\|_2^2 = \sum_{1\leq i\leq d}(a_i)^2$ and the *mean* $\bar{x} = \frac{1}{n}\sum_{1\leq i\leq n} x_i$. Now, use the file you have saved in (a) and compute the mean $\bar{x}$ and *variance* σ^2 of the data in it. **(2)**
   (c) Notice that the variance of the data is highly dominated by few features compared to other features. So, normalize each feature of the saved data with its mean and variance. Now compute the variance of the normalized data. **(3)**
   (d) In the given dataset, with a 5% level of significance, test if the number model year has any effect on the number of cylinders. You may use appropriate tables from here. **(4)**

2. For n points $\{x_1, x_2, \ldots, x_n\}$ its variance is $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$, where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Consider a population, consist of 1,00,000 points uniformly distributed between 0.01 and 1000; for example, your population will be $D = \{0.01, 0.02, 0.03, \ldots, 1000\}$.

   (a) Compute σ^2 of the population D. Let's call σ^2 the *true variance* of the population D. **(1)**

   (b) Use sampling with replacement, to randomly sample 50 points $\{y_1, \ldots, y_{50}\}$ from the population D, i.e., for $1 \leq i \leq 50$, $y_i \in D$. Compute s_1^2, s_2^2 & s_3^2, defined as,

$$s_1^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n+1} \qquad s_2^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n} \qquad s_3^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n-1}.$$

   Here $n = 50$, so $\mu = \frac{\sum_{i=1}^{50} y_i}{50}$. **(1)**

   (c) Say in the first iteration (of 50 random samples) you got $s_1^2 = 21, s_2^2 = 25$ & $s_3^2 = 31$ and in the second iteration (of another 50 random samples) you got $s_1^2 = 18, s_2^2 = 22$ & $s_3^2 = 27$, then after the second iteration maintain, $Avgs_1^2 = \frac{21+18}{2} = 19.5, Avgs_2^2 = \frac{25+22}{2} = 23.5$ & $Avgs_1^2 = \frac{31+27}{2} = 29$. Repeat (b) for multiple iterations and maintain the average scores, i.e., $Avgs_1^2, Avgs_2^2$ & $Avgs_3^2$. **(2)**

   (d) Use three different scatter plots to visualize the change in $Avgs_i^2$, for $1 \leq i \leq 3$ over increasing number of iterations and compare it with σ^2 the *true variance* of D. **(2)**

   (e) Repeat (b), (c) & (d) multiple times and notice among $Avgs_1^2, Avgs_2^2$ and $Avgs_3^2$ which score approaches to the *true variance* much quickly or frequently. Argue its reason. **(2)**

3. Consider you have an k-faced die, numbered 1 to k.

   (a) Let the die be unbiased then over expectation how many times you need to roll the die until you see the number $\lfloor\sqrt{k}\rfloor$ on its upward face. **(2)**

   (b) Over expectation how many times you need to roll the die until you see every number from 1 to k at least once on its upward face. **(3)**

   (c) Consider a 3 faced geometric die i.e., $k = 3$, $\mathbf{P}(1) = \mathbf{P}(3) = \frac{1}{4}$ and $\mathbf{P}(2) = \frac{1}{2}$. What is the expected number of rolls you need on this die to see every number from 1 to 3 at least once on its upward face. **(7)**

   (d) Consider a general k-faced geometric die, where the probability that the upward face is i from a random roll is $\frac{1}{2^{i-1}}$ for $2 \leq i \leq k$ and probability that upward face is 1 is $\frac{1}{2^{k-1}}$. So, $P(1) = P(k) = \frac{1}{2^{k-1}}$, $P(2) = \frac{1}{2}, P(3) = \frac{1}{4}$ and so on. Write a program and show how the exact number of rolls changes as k increases. If you have used a closed form solution for (c) then check if it matches your plot. **(5)**

4. Download the dataset "Hurricane" from this excel file. Write programs for the following tasks.

   (a) With a 1% level of significance conduct t-test for correlation coefficient between "Max. sustained winds(mph)" and "Minimum pressure(mbar)". **(3)**

   (b) With a 5% level of significance test if the "Max. sustained winds(mph)" of hurricane depends on the month of its occurrence. **(5)**

   (c) With a 10% level of significance conduct test if "Max. sustained winds(mph)" follows a Poisson distribution. **(7)**