Data Science
CSE558
Assignment - 1

Submit by 2359 of 10/09/23 Maximum marks: 50

- **While answering the questions, you are encouraged to discuss with your classmates. 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, 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 5 marks per delayed day. Based on a quick viva, you will be graded.**

Best wishes!

---

1. Understand the features of the dataset called Auto MPG that can be found here. Download the dataset 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 and save this in a file. **(2)**
   (b) Let $D = \{x_1, x_2, \ldots, x_n\}$ be n objects consists 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. **(4)**
   (c) You might 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. **(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 $\mu = \frac{\sum_{i=1}^{50} y_i}{50}$. **(2)**

   (c) Say in the first iteration you got $s_1^2 = 21, s_2^2 = 25$ & $s_3^2 = 31$ and in the second iteration 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. **(3)**

3. Let $\Omega \neq$ be a finite set (say sample space). Let $\mathcal{F}$ be a discrete σ-algebra of Ω. So, $\mathcal{F} = \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ be the power set of Ω, i.e., $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$.

   (a) Define a measure $\mathbf{P}(A) = \frac{|A|}{|\Omega|}$ for every event $A \in \mathcal{P}(\Omega)$. Prove or disprove if the measure $\mathbf{P}()$ is a probability measure. **(4)**

   (b) Let $A_1, A_2, \ldots, A_n \subset \mathcal{P}(\Omega)$ be some events in $\mathcal{F}$. We know that,

$$\sum_{1 \leq i \leq n} \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}\left(A_i \bigcap A_j\right) \leq \mathbf{P}\left(\bigcup_{1 \leq i \leq n} A_i\right) \leq \sum_{1 \leq i \leq n} \mathbf{P}(A_i).$$

   Using the principle of inclusion & exclusion prove the following tighter bounds,

$$\begin{aligned}
\mathbf{P}\left(\bigcup_{1 \leq i \leq n} A_i\right) &\leq \sum_{1 \leq i \leq n} \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}\left(A_i \bigcap A_j\right) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}\left(A_i \bigcap A_j \bigcap A_k\right) \\
\mathbf{P}\left(\bigcup_{1 \leq i \leq n} A_i\right) &\geq \sum_{1 \leq i \leq n} \mathbf{P}(A_i) - \sum_{1 \leq i < j \leq n} \mathbf{P}\left(A_i \bigcap A_j\right) + \sum_{1 \leq i < j < k \leq n} \mathbf{P}\left(A_i \bigcap A_j \bigcap A_k\right) \\
&\quad - \sum_{1 \leq i < j < k < l \leq n} \mathbf{P}\left(A_i \bigcap A_j \bigcap A_k \bigcap A_l\right).
\end{aligned}$$

(6)

4. There is a biased k-faced die, numbered 1 to k. 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.

   (a) Consider $k = 4$ and randomly roll the die 4 times and calculate the sum of the upward face value. Repeat this task for 1000 times and plot a frequency distribution histogram. Compute the Bowley's coefficient of the observed sample. **(3)**

   (b) For the same $k = 4$, now randomly roll the die 8 times and calculate the sum of the upward face value. Repeat this task for 1000 times and plot a frequency distribution histogram. Compute the Bowley's coefficient of the observed sample. **(2)**

   (c) Repeat (a) for $k = 16$. Argue the observed changed in the Bowley's coefficient for all the three cases. **(5)**

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

   (a) Over expectation how many times you need to roll the die until you see the number $\lfloor\sqrt{k}\rfloor$[1] 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) Let $k = 3$ and the die is biased, i.e., $\mathbf{P}(1) = \mathbf{P}(3) = \frac{1}{4}$ and $\mathbf{P}(2) = \frac{1}{2}$. Over expectation how many times you need to roll the die until you see every number from 1 to 3 at least once on its upward face. **(5)**

[1] *Floor function*, e.g., $\lfloor 4.234 \rfloor = 4$.