# CSE 558 — Data Science Assignment 1

**Name:** Bhavya Narnoli
**Student Number:** 2021316
**Due Date:** September 10 ,2023, 11:59 PM
**Assignment:** Number 1

**Problem 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.
ans) Answer is in ques1.ipynb

**Problem 2** For n points x1, x2, . . . , xn .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, . . . , 1000.

(a) ans ) answer is in ques2.ipynb , even the last argument part written in markdown.

**Problem 3** Enter question 3 data

(a) probability measure

To determine whether the measure $P(A) = \frac{|A|}{\Omega}$ is a probability measure, we need to check if it satisfies the three axioms of a probability measure:

**Non-negativity:** For any event $A$, $P(A) \geq 0$. Since both $|A|$ and $\Omega$ are non-negative values, $P(A)$ is non-negative for all events $A$. This property holds.

**Normalization:** $P(\Omega) = 1$. We need to check if $P(\Omega) = 1$. If $\Omega$ is the sample space, then $|\Omega| = \Omega$, and

$$P(\Omega) = \frac{\Omega}{\Omega} = 1.$$

So, the normalization property holds if $\Omega$ is the sample space.

**Additivity:** To check additivity, consider a sequence of disjoint events $A_1, A_2, A_3, \ldots$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

This shows that the measure $P(A) = \frac{|A|}{\Omega}$ satisfies all three axioms of a probability measure. Therefore, it is indeed a probability measure as long as $\Omega$ represents the sample space.

$A_1, A_2, A_3, \ldots$. We need to check if:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Substituting $P(A) = \frac{|A|}{\Omega}$, we get:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{|\bigcup_{i=1}^{\infty} A_i|}{\Omega}$$

Now, we know that for disjoint events $A_1, A_2, A_3, \ldots$, the cardinality of their union is the sum of their individual cardinalities:

$$\left|\bigcup_{i=1}^{\infty} A_i\right| = \sum_{i=1}^{\infty} |A_i|$$

Therefore, we have:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\sum_{i=1}^{\infty} |A_i|}{\Omega} = \sum_{i=1}^{\infty} \frac{|A_i|}{\Omega} = \sum_{i=1}^{\infty} P(A_i)$$

This shows that the measure $P(A) = \frac{|A|}{\Omega}$ is additive for a sequence of disjoint events.

So, the measure $P(A) = \frac{|A|}{\Omega}$ satisfies all three axioms of a probability measure. Therefore, it is indeed a probability measure as long as $\Omega$ represents the sample space.

(b) To prove the Inclusion-Exclusion Principle, we will use mathematical induction.

**Base Case (n = 2,3):** For two events, $A_1$ and $A_2$ we have:

$$P(A_1 \cup A_2) = \sum_{i=1}^{2} P(A_i) - \sum_{1 \leq i < j \leq 2} P(A_i \cap A_j)$$

For three events, $A_1$ and $A_2$, $A_3$ we have:

$$P(A_1 \cup A_2 \cup A_3) = \sum_{i=1}^{3} P(A_i) - \sum_{1 \leq i < j \leq 3} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3)$$

This is the standard formula for the union of two events.

$$P\left(\bigcup_{i=1}^{k} A_i\right)$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = \sum_{i=1}^{2} P(A_i) - \sum_{1 \leq i < j \leq k} P(A_i \cap A_j)$$

$$+ \sum_{1 \leq i < j < k \leq k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{k-1} P(A_1 \cap A_2 \cap \ldots \cap A_k)$$

**Inductive Step (Assumption):** Trivial holds for n = 2 , n = 3
Assume that the principle holds for $n = k$:

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) \leq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq k} P(A_i \cap A_j \cap A_k)$$

**Inductive Step (n = k + 1):**
Now, let's add the $(k+1)$th event, $A_{k+1}$, and prove the principle for $n = k+1$:

$$P(A_1 \cup A_2 \cup \ldots \cup A_k \cup A_{k+1})$$

$$\begin{aligned}
&P((A_1 \cup A_2 \cup \ldots \cup A_k) \cup A_{k+1}) \\
&= (P(A_1 \cup A_2 \cup \ldots \cup A_k)) + P(A_{k+1}) - P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k)) = \\
&\leq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k} P(A_i \cap A_j \cap A_t) + P(A_{k+1}) \\
&- P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k))
\end{aligned}$$

Equation 1 [It came from the assumption taken for the elements till k]

Using the inclusion-exclusion principle for $n = k$, we have:
Now, expand $P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k))$ using set operations. $P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k)) =$

$$\sum_{1\leq i<j<k} P((A_{k+1} \cap A_1) \cup (A_{k+1} \cap A_2) \cup (A_{k+1} \cap A_3) \ldots \cup (A_{k+1} \cap A_k))$$

$$\sum_{1\leq i<k} \cup P((A_{k+1} \cap A_i)) \geq \sum_{1\leq i<k} P((A_{k+1} \cap A_i)) - \sum_{1\leq i<j<t} P((A_{k+1} \cap A_i \cap A_j))$$

[ taken from lower bound formula above in question] putting in the above eqn , since the value after applying negative signs both side , found is even more than LHS , it doesn't affect the sign in eqn 1 it will become

$$\leq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k} P(A_i \cap A_j \cap A_t) + P(A_{k+1}) - P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k))$$

$$\begin{aligned}
\leq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t<k} P(A_i \cap A_j \cap A_t) + P(A_{k+1}) - \\
\sum_{1\leq i<k} P((A_{k+1} \cap A_i)) + \sum_{1\leq i<j<k} P((A_{k+1} \cap A_i \cap A_j))
\end{aligned}$$

$$\leq \sum_{i=1}^{k+1} P(A_i) - \sum_{1\leq i<j\leq k+1} P(A_i \cap A_j) + \sum_{1\leq i<t\leq k+1} P(A_i \cap A_j \cap A_t)$$

From the above statements This completes the proof for $n = k + 1$. By mathematical induction, the statement is true.

We see that the given statement is also true for n=k+1. Hence we can say that by the principle of mathematical induction this statement is valid for all events

$$A_1, A_2 \ldots A_n \subset \Omega$$

---

(c) Prove second part of b question

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) \geq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) -$$

$$\sum_{1\leq i<j<k<l\leq n} P(A_i \cap A_j \cap A_k \cap A_l)$$

**Base Case**: Holds true for n = 2 , n= 3 , n = 4
For n = 4

$$P(A_1 \cup A_2 \cup_3 \cup A_4) = \sum_{i=1}^{4} P(A_i) - \sum_{1\leq i<j\leq 4} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq 4} P(A_i \cap A_j \cap A_k) -$$

$$P(A_1 \cap A_2 \cap A_3 \cap A_4)$$

for n = 3 ,

$$P(A_1 \cup A_2 \cup_3) = \sum_{i=1}^{3} P(A_i) - \sum_{1\leq i<j\leq 3} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3)$$

Similarly for 2 , 1
This is the standard formula for the union of two events.

$$P\left(\bigcup_{i=1}^{k} A_i\right)$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = \sum_{i=1}^{2} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j)$$

$$+ \sum_{1\leq i<j<k\leq k} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{k-1} P(A_1 \cap A_2 \cap \ldots \cap A_k)$$

**Inductive Step (Assumption):** Trivial holds for n = 2 , n = 3
Assume that the principle holds for $n = k$:

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) \geq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k} P(A_i \cap A_j \cap A_t)$$

$$- \sum_{1\leq i<j<t<l\leq k} P(A_i \cap A_j \cap A_t \cap A_l)$$

**Inductive Step (n = k + 1):**
Now, let's add the $(k+1)$th event, $A_{k+1}$, and prove the principle for $n = k+1$:

$$P(A_1 \cup A_2 \cup \ldots \cup A_k \cup A_{k+1})$$

TO PROVE

$$P(A_1 \cup A_2 \cup \ldots \cup A_k + 1) \geq \sum_{i=1}^{k+1} P(A_i) - \sum_{1\leq i<j\leq k+1} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k+1} P(A_i \cap A_j \cap A_t)$$

$$- \sum_{1\leq i<j<t<l\leq k+1} P(A_i \cap A_j \cap A_t \cap A_l)$$

Proof :

$P((A_1 \cup A_2 \cup \ldots \cup A_k) \cup A_{k+1})$
$= (P(A_1 \cup A_2 \cup \ldots \cup A_k)) + P(A_{k+1}) - P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k)) =$

$$\geq \sum_{i=1}^{k} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k} P(A_i \cap A_j \cap A_t) - \sum_{1\leq i<j<t<l\leq k} P(A_i \cap A_j \cap A_t \cap A_l)$$

$$+P(A_{k+1}) - P(A_{k+1} \cap (A_1 \cup A_2 \cup \ldots \cup A_k))$$

$$\geq \sum_{i=1}^{k+1} P(A_i) - \sum_{1\leq i<j\leq k} P(A_i \cap A_j) + \sum_{1\leq i<j<t\leq k} P(A_i \cap A_j \cap A_t) - \sum_{1\leq i<j<t<l\leq k} P(A_i \cap A_j \cap A_t \cap A_l)$$

$$-P((A_{k+1} \cap A_1) \cup (A_{k+1} \cap A_2) \cup \ldots (A_{k+1} \cap A_k))$$

the term : [ using union property proved in last question 3) part b )]

$$P((A_{k+1} \cap A_1) \cup (A_{k+1} \cap A_2) \cup \ldots (A_{k+1} \cap A_k))$$

$$\leq \sum_{1\leq i\leq k} P((A_{k+1} \cap A_i)) - \sum_{1\leq i<j\leq 1k+} P((A_{k+1} \cap A_i \cap A_j))$$

$$+\sum_{1\le i<j<t\le k} P((A_{k+1}\cap A_i\cap A_j\cap A_t))$$

putting this back in eqn above , doesn't affect the sign , since after applying negative sign both sides the rhs is even lesser than lhs of this equation of the term

$$\ge \sum_{i=1}^{k+1} P(A_i) - \sum_{1\le i<j\le k} P(A_i\cap A_j) + \sum_{1\le i<j<t\le k} P(A_i\cap A_j\cap A_t) - \sum_{1\le i<j<t<l\le k} P(A_i\cap A_j\cap A_t\cap A_l)$$

$$-\sum_{1\le i\le k} P((A_{k+1}\cap A_i)) + \sum_{1\le i<j\le k} P((A_{k+1}\cap A_i\cap A_j))$$

$$-\sum_{1\le i<j<t\le k} P((A_{k+1}\cap A_i\cap A_j\cap A_t))$$

$$P(A_1\cup A_2\cup\ldots\cup A_k+1) \ge \sum_{i=1}^{k+1} P(A_i) - \sum_{1\le i<j\le k+1} P(A_i\cap A_j) + \sum_{1\le i<j<t\le k+1} P(A_i\cap A_j\cap A_t)$$

$$-\sum_{1\le i<j<t<l\le k+1} P(A_i\cap A_j\cap A_t\cap A_l)$$

From the above statements This completes the proof for $n = k+1$. By mathematical induction, the statement is true.

We see that the given statement is also true for n=k+1. Hence we can say that by the principle of mathematical induction this statement is valid for all events

$$A_1, A_2 \ldots A_n \subset \Omega$$

---

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

(a) Over-expect how many times you need to roll the die until you see the number $|\sqrt{k}|$ on its upward face.

For a distribution , like this assuming it's geometric distribution

where ith success probability is denoted by P( X == i ) = ( 1 - p)$^(i-1)$ * p , where p denotes the probability of success of event in this case p = 1/k where , p denotes sucess in seeing the number $|\sqrt{k}|$ on its upward face. Since, it's given it's unbiased, every face has a equal probability of p = 1/k.

E[X == i] = i * P(X == i) = i * ( 1 - p)$^i - 1$ * p

$$\sum_{i=1}^{k} P(X == i) * i$$

= 1 * ( 1 - p)$^0$ * p + 2 *( 1 - p)$^1$ * p + 3 * ( 1 - p)$^2$ * p + 4 * ( 1 - p)$^3$ * p + 5 * ( 1 - p)$^4$ * p + 6 *( 1 - p)$^5$ * p + .. .+ n * ( 1 - p)$^($n$-1$) * p [Equation 1]

Multiply -(1-p) both sides

-(1-p) * E[X == i] = -( 1 - p ) *i * P(X == i) = -i * ( 1 - p)$^($i$)$ * p
= 1 * -( 1 - p)$^1$ * p - 2 *( 1 - p)$^2$ * p - 3 * ( 1 - p)$^3$ * p - 4 * ( 1 - p)$^4$ * p - 5 * ( 1 - p)$^5$ * p - 6 *( 1 - p)$^6$ * p + .. .- n * ( 1 - p)$^($n$)$ * p [Equation 2]
Adding the two equations ,
p* E[X] = p + p * ( 1 - p)$^1$ + p * ( 1 - p)$^2$ + p * ( 1 - p)$^3$ + .. ..
E[X] = 1 + (1-p) + ( 1 - p)$^2$ + ( 1 - p)$^3$ + ..
E[X] = 1 / (1 - ( 1 - p)) = 1/p
E[X] = 1 / (1/k) = k

The expected number of trials until the first success occurs is given by = k

(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. The event of observing the ith unique coupon follows a geometric distribution, whose probability of success is (pi) = (k-(i-1))/k.
Consider the random variable Xi
So, E[Xi] = 1/pi (coupons needed to see ith unique coupon).
As proven above , E[Xi] = j , with probability , (1 - p$_i$)$^($i-1$)$ * p$_i$,
$E[\sum_{i=1}^{k} X_i] = \sum_{i=1}^{k} E[X_i]$
$= \sum_{i=1}^{k} k/(k-(i-1)) = klog(k)/log(1)$

(c) Let $k = 3$ and the die is biased, i.e., $P(1) = P(3) = \frac{1}{4}$ and $P(2) = \frac{1}{2}$. Over expectation, how many times do you need to roll the die until you see every number from 1 to 3 at least once on its upward face?

In this scenario, we have a biased 3-faced die with the following probabilities:

$$P(1) = \frac{1}{4}$$
$$P(2) = \frac{1}{2}$$
$$P(3) = \frac{1}{4}$$

We want to find the expected number of rolls needed to see every number from 1 to 3 at least once on the die's upward face. This problem can be solved using a similar approach to the Coupon Collector's Problem as done in previous question proved expectation of geometric variable 1/p for p: sucess of event.

Let $E(X)$ be the expected number of rolls needed. We calculate the expected number of rolls to collect each number:

$$\text{Expected number of rolls to collect 1 } (E_1) = \frac{1}{P(1)} = 4$$
$$\text{Expected number of rolls to collect 2 } (E_2) = \frac{1}{P(2)} = 2$$
$$\text{Expected number of rolls to collect 3 } (E_3) = \frac{1}{P(3)} = 4$$

Now, we have collected all three numbers, so the total expected number of rolls to see all the numbers at least once is:

$$E(X) = E_1 + E_2 + E_3 = 4 + 2 + 4 = 10$$

So, in this specific scenario with a biased 3-faced die, you would, on average, need to roll the die 10 times until you see every number from 1 to 3 at least once on its upward face.