Assignment: 1

Page
Date

Q:1) What is Data science? What are its applications?

Data Science is a multidisciplinary field that involves using scientific methods, algorithms, processes, and systems to extract knowledge and insights from data.

- It combines elements of statistics, mathematics, computer science, and domain expertise to analyse and interpret data.

Applications of data science are vast and diverse, including:

1. Business Analytics:
Forecasting sales, customer segmentation, market analysis, and decision-making support.

2. HealthCare:
Predictive modeling for disease diagonosis, drug discovery, and personalized medicine.

3. Finance:
Fraud detection, risk assessment, algorithmic trading, and credit scoring.

4. Marketing:
Customer behaviour analysis, targeted advertising, and campaign optimization.

5. Social Media Analysis:
Sentiment analysis, network analysis, and recommendation systems

6. Natural Language Processing:
Language translation, sentiment analysis and chatbots.

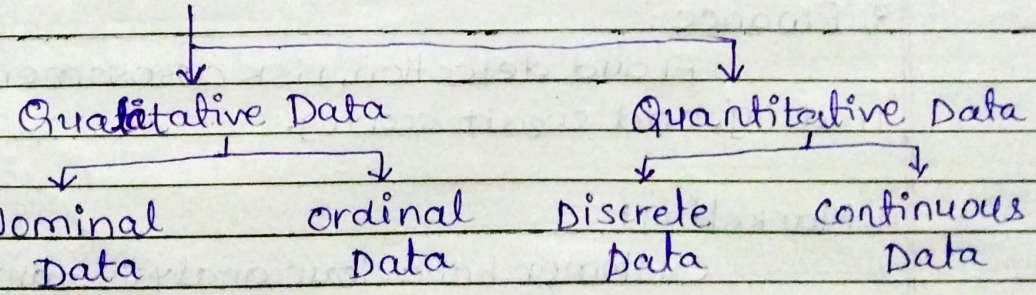
7. Internet of Things (IoT):
Analyzing sensor data for predective maintenance and optimization.

8 Environmental Science:
Climate modeling, resource management, the environmental impact assessment.

Q:2) What are the different types of Data in Data Science? Explain this?

→ Data is a collection of facts and figures. It is a set of characters used to collect & store information for specific purpose.

* Types of Data :-



(1) Qualitative Data :-

1. Nomial Data : (to label variables without any other)
→ Can't arrange in sequence (order)
↳ Ex: Colour, gender, etc.

2. Ordinal Data:
- Having natural ordering (on the scale)
- Arrange in sequence (order)
- Feedback, Rank, Grade, etc.

(2) Quantitative Data:

1. Discrete Data:
→ means separated
→ It contains values under integer or whole number.
→ It can't be broken into decimal or fraction.
Ex. NO. of students in class, etc.

2. Continuous Data:
→ In form of fractional numbers.
→ can be height of a person, height of object, etc.
→ It represents information, that is divided into small parts.

Q:3) calculate Descriptive statistics (mean, median, mode, standard deviation, variance, range) of the following list.
82, 93, 91, 69, 96, 61, 88, 58, 59, 100, 93, 71, 78, 98

Soln Sum = 1137, n = 14

$$\text{mean} = \frac{\text{Sum}}{n} = \frac{1137}{14} = 81.2143 \approx 81$$

Data in Ascending Order:
58, 59, 61, 69, 71, 78, 82, 88, 91, 93, 93, 96, 98, 100

$$\text{median} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$$

$$= \frac{82+88}{2}$$

$$= \frac{170}{2}$$

$$= 85$$

mode = 93

Range = 100 − 58

$= 42$

$$\text{standard deviation (sample)} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\sum (x_i - \bar{x})^2 = (23)^2 + (22)^2 + (20)^2 + (12)^2 + (10)^2 + (3)^2 + (1)^2 + (7)^2 + (10)^2 + (12)^2 + (12)^2 + (15)^2 + (17)^2 + (19)^2$$

$$= 529 + 484 + 400 + 144 + 100 + 9 + 1 + 49 + 100 + 144 + 144 + 225 + 289 + 361$$

$$= 2979$$

$$\text{standard deviation (sample)} = \sqrt{\frac{2979}{13}}$$

$$= \sqrt{229.154}$$

$$= 15.1378$$

$$V(x) = E(x^2) - (E(x))^2$$

$$E(x) = \frac{\sum x_i}{n} = 81$$

$E(x^2) = \frac{\sum x_i^2}{n}$

$= \frac{95319}{14}$

$= 6808.5$

$V(x) = (15.1378)^2$

$= 229.154$

$V(x) = E(x^2) - (E(x))^2$

$= 6808.5 - 6561$

$= 247.5$

Q:4) Calculate sample covariance of given set of numbers.

$x = 2.1, 2.5, 4.0, 3.6 \quad y = 8, 12, 14, 10$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
2.1	8	-0.95	-3	2.85
2.5	12	-0.55	1	-0.55
4.0	14	+0.95	3	2.85
3.6	10	+0.55	-1	-0.55
				4.6

$\bar{x} = 3.05$

$\bar{y} = 11$

$cov(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{n}$

$= \frac{4.6}{4}$

$= 1.15$

Q:5) What is sampling? what are the different sampling techniques?

Ans. Sampling is the process of selecting a subset of a population to study. This is done in order to make inferences about the population as a whole

There are lot of sampling techniques, which are grouped into two categories as
- Probability Sampling
- Non-Probability Sampling

* Non Probability Sampling
- This sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It's alternatively known as Random Sampling.

1) Simple Random Sampling:
A method where every individual or item in the population has an equal chance of being selected for the sample, without any specific criteria or grouping.

2) Stratified Sampling:
Divides the population into mutually exclusive subgroups (strata) and then randomly selects samples from each stratum, ensuring representation of various characteristics within the population.

3) Systematic Sampling:

Selects samples at fixed intervals from the population, often using a random starting point, providing an evenly spread representation.

4) cluster sampling:

Divides the population into clusters, randomly selects some clusters, and then includes all the individuals within the selected clusters in sample

5) Multi-stage sampling:

Involves a combination of various sampling techniques, using multiple stages of selection, suitable for large and complex populations.

* Non-Probability Sampling:

- It does not rely on randomization. This technique is more reliant on the researcher's ability to select elements for a sample. This type of sampling is also known as non-random sampling.

1) convenience Sampling:

Involves selecting individuals who are readily available and easily accessible, leading to potential biases due to the non-random nature of selection.

2) Purposive Sampling:

Also known as judgemental or selective sampling, it involves deliberately choosing specific individuals who possess perticular characteristics or expertise relevant to the study.

3) Quota Sampling:

Selects participants to meet Pre-determined proportions or quotas for certain characteristics, often used when random sampling is difficult.

4) Referral/snowball Sampling:

Relies on initial participants to refer or recruit others from their network, commonly used in studies involving hard-to-reach populations.

Q:6) Explain Hypothesis Testing. Explain any 2 types of testing.

Hypothesis testing is a method in inferential statistics used to make assumptions about a full populations based on a representative sample since observing the entire population is often impossible. It involves establishing hypotheses, such as a difference between two groups or a correlation between variable in the population. The Alternative Hypothesis (HA) represents the thesis to be proven, while the Null Hypothesis (HO) states that nothing new is happening in the population.

The purpose of hypothesis testing is to determine if the Null Hypothesis can be rejected or not. Rejecting the Null Hypothesis does not necessarily prove the Alternative Hypothesis true, but it may lead to acceptance of the Alternative Hypothesis in some cases.

→ Two Types of Hypothesis testing are:

1) T-test:

- The T-test used when the population standard deviation is unknown, and sample size is relatively small.
- The test statistics for that t-test is calculated as the difference between the sample mean and the population mean, divided by the standard ~~error~~ deviation of the mean.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

s: standard deviation
μ: Population mean
$\bar{x}$: Sample's mean
n: Sample's size.

2) Z-Test

- The z-test is used when the population standard deviation is known, or the sample size is large.
- It is often employed when comparing the means of two groups or ~~mean~~ when comparing the mean of a sample to a known population mean.

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$H0: \mu = \mu_0$ or $HA: \mu \neq \mu_0$
μ_0 = Population mean
σ = standard deviation
n = Sample size
$\bar{x}$ = sample mean

Q:7) Difference between classification and clustering

Classification	clustering
- It is used for supervised Learning.	- It is used for unsupervised Learning.
- It is a process of classifying the input instances based on their corresponding class labels.	- grouping the instances based on their similarity without the help of class labels.
- It has labels so there is need of training and testing dataset for verifying the model created.	- There is no need of training and testing dataset.
- more complex as compared to clustering	- less complex as compared to classification.
- Logistic Regression, Naive Bayes classifier, support Vector machines etc.	- K means clustering algo, Fuzzy - C means clustering algo, Gaussian (EM) clustering algo.
- Produces class labels or category predictions for new, unseen data based on the learned model.	- Outputs clusters or groups of data instances that are similar to each other based on their features.

Q:8) Briefly discuss K-means clustering algorithm with its pros and cons.

Ans: K-Mean means is a popular clustering algorithm used to partition data into K clusters based on their similarity.

Algorithm:

1. Choose the number of clusters K.
2. Randomly initialize K cluster centroids.
3. Assign each data point to the nearest centroid, forming K clusters.
4. Calculate the mean of each cluster to obtain new centroid.
5. Repeat steps 3 and 4 untill convergence (centroid do not change significantly) or a maximum number of iterations is reached.

Pros :

- Simple and easy to implement.
- computionally efficient, making it suitable for large databases sets.
- Works well when clusters are well-defined and compact.
- Converges to local optimum, which often provides reasonble results.

Cons :

- Requires the number of clusters K, to be specified beforehand, which may not always be known in advance.
- May produces suboptimal results when dealing with non-linearly separable or overlapping clusters.
- Does not work well with clusters of varying sizes or irregular shapes.

Q.9) Illustrate the steps of ANN algorithm with proper example.

→ To illustrate the steps of the Average Nearest Neighbor (ANN) algorithm we consider Business Analyst data. Analyst use ANN algorithm to perform classification analysis of datasets in his organization database. He compares employees based on the following five features:

1) Age 2) Number of children 3) Annual income 4) Seniority 5) Eligibility to Retire

Step 1: Input the spatial data set.

The database contains the 5 dimensional tuples representing the employees features.

Name	Age	childrens	Annual income	Seniority	Elligible to Retire
Mike	34	1	$120,000	9	0
Liz	42	0	$90000	5	0
JIN	22	0	$60000	2	0
Mary	53	3	$1,80,000	30	1

Step 2: calculate Difference Between Each 2 Employees.

1. Mike & Liz

34	1	120000	9
- 42	0	90000	5
-8	1	30000	4

$: Calculate Difference. no need of Negative number

Same for all pair.

2. Mike & Jin :
   (12, 1, 60000, 7)
3. Mike & Mary :
   (19, 2, 60000, 21)
4. Liz & Jin :
   (20, 0, 30000, 3)
5. Liz & Mary :
   (11, 3, 90000, 25)
6. Jin & Mary :
   (31, 3, 120000, 28)

step 3: Finding Average of each pair of features values.

Mike & Liz = 7503.25
Mike & Jin = 15005
Mike & Mary = 15010.5
Liz & Jin = 7505.75
Liz & Mary = 22509.75
Jin & Mary = 30015.5

step 4: Take pairs that have minimum Average distance. compare that pair with other that is near to it.

- Mike & Liz and Liz & Jin { ∵ 7503.25 ≈ 7505.75 }

So, we can group of Mike, Liz & Jin which will class-0.

- ~~Now, Mary is~~
- So, Mary is now in class-1.

Q:10) What is confusion Matrix ? Explain in detail all its measures.

A confusion matrix is a tabular representation of Prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on set of ~~data~~ test data when true values are known.

The confusion matrix is simple to implement, but the terminologies used in this matrix might be confusing for beginners.

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

- True Positive: when the actual value is positive and predicted is also positive.
- True Negative: when the actual value is Negative and prediction is also Negative.
- False Positive: when the actual is negative but Prediction is positive. Also known as the Type 1 Error.
- False Negative: when the actual is positive but the prediction is Negative. Also known as the Type 2 error.

1. Accuracy:

The accuracy metric is one of the simplest classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{Total number of predictions}}$$

2. Precision:

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below formula.

$$\text{precision} = \frac{TP}{TP + FP}$$

3. Recall or Sensitivity:

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F-scores

F1 score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

$$\text{F1 score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$