	UNIT I
SR.NO	QUESTION
1	Which of the following is the most important language for Data Science?
[a]	Java
[b]	Ruby
[c]	R
[d]	None of the mentioned
	Correct Option:[c]
2	Which of the following is one of the key data science skills?
[a]	a) Statistics
[b]	b) Machine Learning
[c]	c) Data Visualization
[d]	d) All of the mentioned
	Correct Option:[d]
3	Which of the following is characteristic of Processed Data?
[a]	Data is not ready for analysis
[b]	All steps should be noted
[c]	Hard to use for data analysis
[d]	None of the mentioned
	Correct Option:[b]
4	What are the five V's of Big Data?
[a]	Volume
[b]	Velocity
[c]	Variety
[d]	All the above
	Correct Option:[d]
5	Point out the correct statement.
[a]	Machine learning focuses on prediction, based on known properties learned from the training data
[b]	Data Cleaning focuses on prediction, based on known properties learned from the training data

[c]	Representing data in a form which both mere mortals can understand and get valuable insights is as much a science as much as it is art
[d]	None of the mentioned
	Correct Option:[d]
6	Which of the following step is performed by data scientist after acquiring the data?
[a]	Data Cleansing
[b]	Data Integration
[c]	Data Replication
[d]	All of the mentioned
	Correct Option:[a]
7	Which of the following focuses on the discovery of (previously) unknown properties on the data?
[a]	Data mining
[b]	Big Data
[c]	Data wrangling
[d]	Machine Learning
	Correct Option:[a]
8	A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?
[a]	12
[b]	24
[c]	48
[d]	72
	Correct Option:[d]
Ans:	***Maximum possible different examples are the products of the possible values of each attribute and the number of classes;***
	3 * 2 * 2 * 2 * 3 = 72
9	What is unsupervised learning?
[a]	features of group explicitly stated
[b]	number of groups may be known
[c]	neither feature & nor number of groups is known

[d]	none of the mentioned
	Correct Option:[c]
10	Supervised Learning differ from unsupervised clustering in that supervised learning requires
[a]	at least one input attribute
[b]	input attribute to be categorical
[c]	at least one output attribute
[d]	output attribute to be categorical
	Correct Option:[b]
11	Reinforcement learning is-
[a]	Unsupervised learning
[b]	Supervised learning
[c]	Award based Learning
[d]	None of the mentioned
	Correct Option:[c]
12	What are different types of attributes?
[a]	Nominal
[b]	Ordinal
[c]	Spacial
[d]	All of the above
	Correct Option:[d]
13	Important Characteristics of Structured Data are:
[a]	Generality
[b]	Dimensionality
[c]	Resolution
[d]	All of the Above
	Correct Option:[d]
14	The correct way of pre processing the data should be-
[a]	Imputation ->feature scaling-> training
[b]	Feature scaling->imputation->training
[c]	Feature scaling->label encoding->training

[d]	None
	Correct Option:[a]
15	Supervised Learning is
[a]	learning with the help of examples
[b]	learning without teacher
[c]	learning with the help of teacher
[d]	learning with computers as supervisor
	Correct Option:[c]
16	Which of the following are ML methods?
[a]	based on human supervision
[b]	supervised Learning
[c]	semi-reinforcement Learning
[d]	All of the above
	Correct Option:[a]
17	Data science is the process of diverse set of data through ?
[a]	organizing data
[b]	processing data
[c]	.analysing data
[d]	All of the above
	Correct Option:[d]
18	Which of the following are correct component for data science?
[a]	Data Engineering
[b]	Advanced Computing
[c]	Domain expertise
[d]	All of the above
	Correct Option:[d]
19	Which of the following is not a part of data science process?
[a]	Discovery
[b]	Model Planning
[c]	Communication Building

[d]	Operationalize
	Correct Option:[c]
20	Which of the following is not a application for data science?
[a]	Recommendation Systems
[b]	Image & Speech Recognition
[c]	Online Price Comparison
[d]	Privacy Checker
	Correct Option:[d]
20	Which of the following focuses on the discovery of (previously) unknown properties on the data?
[a]	Data mining
[b]	BigData
[c]	Data wrangling
[d]	Machine Learning
	Correct Option:[a]
	UNIT II
SR.NO	QUESTION
1	
[a]	
[b]	
[c]	
[d]	
2	
[a]	
[b]	
[c]	
[d]	

3	Previous probabilities in Bayes Theorem that are changed with help of new available information are classified as _____
[a]	independent probabilities
[b]	posterior probabilities
[c]	interior probabilities
[d]	dependent probabilities
	Correct Option:[a]
4	Mutually Exclusive events _____
[a]	Contain all sample points
[b]	Contain all common sample points
[c]	Does not contain any sample point
[d]	Does not contain any common sample point
	Correct Option:[d]
5	If the values taken by a random variable are negative, the negative values will have _____
[a]	Positive probability
[b]	Negative Probability
[c]	May have negative or positive probabilities
[d]	Insufficient data
	Correct Option:[a]
6	The variable that assigns a real number value to an event in a sample space is called _____
[a]	Random variable
[b]	Defined variable
[c]	Uncertain variable
[d]	Static variable
	Correct Option:[a]
7	A jar containing 8 marbles of which 4 red and 4 blue marbles are there. Find the probability of getting a red given the first one was red too.
[a]	4/13
[b]	2/11
[c]	3/7

[d]	8/15
	Correct Option:[c]
Explaination	Answer: c Explanation: Suppose, P (A) = getting a red marble in the first turn, P (B) = getting a black marble in the second turn. P (A) = $\frac{4}{8}$ and P (B) = $\frac{3}{7}$ and P (A and B) = $\frac{4}{8} * \frac{3}{7} = \frac{3}{14}$ P(B/A) = $\frac{P(A \text{ and } B)}{P(A)} = \frac{\frac{3}{14}}{\frac{1}{2}} = \frac{3}{7}$.
8	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
9	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
10	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
11	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]

12	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
13	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
14	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
15	
[a]	
[b]	
[c]	
[d]	
	Correct Option:[d]
	UNIT III
SR.NO	QUESTION

1	Which of the following is required by K-means clustering?
[a]	defined distance metric
[b]	number of clusters
[c]	initial guess as to cluster centroids
[d]	all of the mentioned
	Correct Option:[d]
2	Significant Bottleneck in the Apriori algorithm is
[a]	Finding frequent itemsets
[b]	Pruning
[c]	Candidate generation
[d]	Number of iterations
	Correct Option:[c]
3	With Bayes theorem the probability of hypothesis HÂ¾ specified by P(H) Â¾ is referred to as
[a]	conditional probability
[b]	an a priori probability
[c]	a bidirectional probability
[d]	a posterior probability
	Correct Option:[b]
4	Which of the following clustering requires merging approach?
[a]	Partitional
[b]	Hierarchical
[c]	Naive Bayes
[d]	None of the mentioned
	Correct Option:[b]
5	If two variables V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?
	1. If V1 and V2 has a correlation of 1, the cluster centroids will be in a straight line
	2. If V1 and V2 has a correlation of 0, the cluster centroids will be in straight line
[a]	1 only
[b]	2 only
[c]	1 and 2

[d]	None of the above
	Correct Option:[a]
Reason	If the correlation between the variables V1 and V2 is 1, then all the data points will be in a straight line. Hence, all the three cluster centroids will form a straight line as well.
6	What techniques can be used to improve the efficiency of apriori algorithm?
[a]	Hash-based techniques
[b]	Transaction Increases
[c]	Sampling
[d]	Cleaning
	Correct Option:[a]
7	Cluster is----------
[a]	Group on a training data set to transform or simplify data in order to prepare it for a machine-learning algorithm
[b]	Group of similar objects that differ significantly from other objects
[c]	Symbolic representation of facts or ideas from which information can potentially be extracted
[d]	Both a and b
	Correct Option:[b]
8	The number of iterations in apriori _____
[a]	increases with the size of the data
[b]	decreases with the increase in size of the data
[c]	increases with the size of the maximum frequent set
[d]	decreases with increase in size of the maximum frequent set
	Correct Option:[c]
9	Which of the following are interestingness measures for association rules?
[a]	recall
[b]	lift
[c]	accuracy
[d]	compactness
	Correct Option:[b]
10	The apriori property means

[a]	If a set cannot pass a test, its supersets will also fail the same test
[b]	To decrease the efficiency, do level-wise generation of frequent item sets
[c]	To improve the efficiency, do level-wise generation of frequent item sets
[d]	If a set can pass a test, its supersets will fail the same test
	Correct Option:[a]
11	Produce dependency rules which will predict occurrence of an item based on occurrences of other items.
[a]	Sequential Pattern Discovery
[b]	Association Rule Discovery
[c]	Both a and b
[d]	Otherwise
	Correct Option:[b]
12	Which of the following statement is true about outliers in Linear regression?
[a]	Linear regression is sensitive to outliers
[b]	Linear regression is not sensitive to outliers
[c]	Can't say
[d]	None of these
	Correct Option:[a]
Reason	The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.
13	Three companies A, B and C supply 25%, 35% and 40% of the notebooks to a school. Past experience shows that 5%, 4% and 2% of the notebooks produced by these companies are defective. If a notebook was found to be defective, what is the probability that the notebook was supplied by A?
[a]	$44/69$
[b]	$25/69$
[c]	$13/24$
[d]	$11/24$
	Correct Option:[b]

	Explanation: Let A, B and C be the events that notebooks are provided by A, B and C respectively. Let D be the event that notebooks are defective Then, P(A) = 0.25, P(B) = 0.35, P(C) = 0.4 P(D|A) = 0.05, P(D|B) = 0.04, P(D|C) = 0.02 P(A| D) = (P(D| A)*P(A))/(P(D| A) * P(A) + P(D| B) * P(B) + P(D| C) * P(C)) = (0.05*0.25)/((0.05*0.25)+(0.04*0.35)+(0.02*0.4)) = 2000/(80*69) = 25⁄69.
14	Which of the following is correct about the Naive Bayes?
[a]	Assumes that all the features in a dataset are independent
[b]	Assumes that all the features in a dataset are equally important
[c]	Both
[d]	All of the above
	Correct Option:[c]
15	Naïve Bayes Algorithm is a _____ learning algorithm.
[a]	Supervised
[b]	Reinforcement
[c]	Unsupervised
[d]	None of these
	Correct Option:[a]
16	Examples of Naïve Bayes Algorithm is/are
[a]	Spam filtration
[b]	Sentimental analysis

[c]	Classifying articles
[d]	All of the above
	Correct Option:[d]
17	Naïve Bayes algorithm is based on _____ and used for solving classification problems.
[a]	Bayes Theorem
[b]	Candidate elimination algorithm
[c]	EM algorithm
[d]	None of the above
	Correct Option:[a]
18	Disadvantages of Naïve Bayes Classifier:
[a]	Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.
[b]	It performs well in Multi-class predictions as compared to the other Algorithms.
[c]	Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
[d]	It is the most popular choice for text classification problems.
	Correct Option:[a]

SINHGAD INSTITUTE OF TECHNOLOGY, LONAVALA

Department of Electronics and Telecommunication

MCQ PRELIM EXAM