Chartered Data Scientists Practice Paper

The below questions give a sense of exam that you can expect on the day of your CDS registration. Please use this practise paper as a yardstick but expect the difficulty level to go up on the exam day.

Section 1: Probability Theory, Statistics and Linear Algebra

1. A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?
i. 10/21 ii. 11/21 iii. 2/7 iv. 5/7

2. The chance that doctor *A* will diagnose a disease *X* correctly is 60 per cent. The chance that a patient will die by his treatment after a correct diagnosis is 40 per cent, and the chance of death by the wrong diagnosis is 70 per cent. A patient of doctor *A*, who had disease *X*, died. What is the probability that this disease was diagnosed correctly?
i. 6/25 ii. 7/25 iii. 6/7 iv. 6/13

3. If $X_1, X_2, ..., X_n$ be a sequence of mutually independent random variables where X_i can take only positive integral values and Sm=i=1mXi(mn), Sn=i=1nXi, E(X)=>0, then
i. E(SmSn)=1
ii. E(SmSn)=0
iii. E(SmSn)=mn
iv. E(SmSn)=

4. Which of the following is the correct formula for calculating the mean of a discrete series by direct method?
i. X=X / f
ii. X=ifiXi / ifi
iii. X=A+ifidxi/ fi
iv. None of these

5. Two matrices *A* and *B* are given below:

$$A = \begin{bmatrix} p & q \\ r & s \end{bmatrix}; \; B = \begin{bmatrix} p^2 + q^2 & pr + qs \\ pr + qs & r^2 + s^2 \end{bmatrix}$$

If the rank of matrix *A* is *N*, what is the rank of matrix *B*?

i. N/2 ii. N-1 iii. N iv. 2N

6. Let A be the *2 × 2* matrix with elements $a_{11} = a_{12} = a_{21} = +1$ and $a_{22} = -1$. Then the eigenvalues of the matrix A^{19} are:
i. 1024 and -1024 ii. 1024√2 and -1024√2
iii. 4√2 and -4√2 iv. 512√2 and -512√2

7. In a class, there are 15 boys and 10 girls. Three students are selected at random. The probability that 1 girl and 2 boys are selected, is:
i. 21/46 ii. 25/117 iii. 1/50
iv. 3/25

8. Three unbiased coins are tossed. What is the probability of getting at most two heads?
i. 3/4 ii.1/4 iii. 3/8 iv. 7/8

9. Let *X* be a continuous random variable with probability density function,

f(x)=kx; 0x1
=k; 1x2
=0; otherwise

The value of kis equal to:
i. 14 ii. 23 iii. 25 iv. 34

10. If *X* and *Y* are two random variables such that their expectations exist and $P(x \leq y) = 1$. Which of the following is correct?
i. $E(X) \leq E(Y)$
ii. $E(X) \geq E(Y)$
iii. $E(X) = E(Y)$
iv. None of the above

11. If the rank of a *5 x 6* matrix *Q* is 4, then which one of the following is true?
i. Q will have four linearly independent rows and four linearly independent columns
ii. Q will have four linearly independent rows and five linearly independent columns
iii. QQ^T will be invertible
iv. Q^TQ will be invertible

12. Consider the following statements:
A. The sum of two singular *n × n* matrices may be non-singular
B. The sum of two *n × n* non-singular matrices may be singular.
Which of the above statements are correct?
i. Both A and B are correct ii. Only A is correct
iii. Only B is correct iv. Both are incorrect

13. In a lottery, there are 10 prizes and 25 blanks. A lottery is drawn at random. What is the probability of getting a prize?
i. 1/10 ii. 2/5 iii. 2/7 iv. 5/7

14. Classical probability is measured in terms of:
i. An absolute value ii. A ratio
iii. Absolute value and ratio both iv. None of the above

15. If *X* is a random variable with mean μ, the E(X-)r is called:
i. Variance iii. r^{th} raw moment
iii. r^{th} central moment iv. None of these

16. *t*-distribution is used to test:
i. The validity of a postulated value of the population mean
ii. To test the significance of sample correlation coefficient
iii. To test the equality of two population means
iv. All of the above

17. Which one of the following statements is correct?
i. If the trace of the matrix is positive and the determinant of the matrix is negative, at least one of its eigenvalues is negative.
ii. If the trace of the matrix is positive, all its eigenvalues are positive.
iii. If the determinant of the matrix is positive, all its eigenvalues are positive.
iv. If the product of the trace and determinant of the matrix is positive, all its eigenvalues are positive.

18. Two eigenvalues of a *3 x 3* real matrix P are (2 + √-1) and 3. What is the determinant of P?
i. 0 ii. 1 iii. 15 iv. -1

Answers of Section 1:

1. i	2. iv	3. iii	4. ii	5. iii	6. iv	7. i	8. iv	9. ii	10. i	11. i	12. i
13. iii	14. ii	15. iii	16. iv	17. i	18. iii						

Section 2: Data Engineering and Database

1. Consider a relational table with a single record for each registered student with the following attributes.
a. Registration_Number: Unique registration number for each registered student
b. UID: Unique Identity number, unique at the national level for each citizen
c. BankAccount_Number: Unique account number at the bank. A student can have multiple accounts or joint accounts. This attributes stores the primary account number
d. Name: Name of the Student
e. Hostel_Room: Room number of the hostel

Which of the following options is not correct?
i. BankAccount_Number is a candidate key
ii. Registration_Number can be a primary key
iii. UID is a candidate key if all students are from the same country
iv. If S is a superkey such that S ∩ UID is NULL then S ∪ UID is also a superkey

2. Department (dept_name, building, budget) and Employee (employee_id, name, dept_name, salary)
Here the dept_name attribute appears in both the relations. Here using common attributes in relation schema is one way of relating _____ relations.
i. Attributes of common
ii. Tuple of common
iii. Tuple of distinct
iv. Attributes of distinct

3. Which of the following statements is not correct?
i. The data dictionary is normally maintained by the database administrator.
ii. Data elements in the database can be modified by changing the data dictionary.
iii. The data dictionary contains the name and description of each data element.
iv. A data dictionary is a tool used exclusively by the database administrator.

4. Which of the following statements is not correct?
i. Non Relational databases require that schemas be defined before you can add data
ii. NoSQL databases are built to allow the insertion of data without a predefined schema
iii. NewSQL databases are built to allow the insertion of data without a predefined schema
iv. All of the above

5. Consider the following SQL queries:
CREATE TABLE Employee(Emp_id NUMERIC NOT NULL, Name VARCHAR(20) , dept_name VARCHAR(20), Salary NUMERIC UNIQUE(Emp_id,Name));
INSERT INTO Employee VALUES(1002, Ross, CSE, 10000)
INSERT INTO Employee VALUES(1006,Ted,Finance,);
INSERT INTO Employee VALUES(1002,Rita,Sales,20000);
What will be the result of the execution of the above query?
i. All statements executed
ii. Error in Create statement
iii. Error in insert into Employee values(1006,Ted,Finance,);
iv. Error in insert into Employee values(1008,Ross,Sales,20000);

6. Which of the following statements is correct about a global transaction?
i. The required data are at one local site and the distributed DBMS routes request as necessary.

ii. The required data are located in at least one nonlocal site and the distributed DBMS routes request as necessary.
iii. The required data are at one local site and the distributed DBMS passes the request to only the local DBMS.
iv. The required data are located in at least one nonlocal site and the distributed DBMS passes the request to only the local DBMS.

7. Which of the following statements is correct about views?
i. The user who creates a view cannot be given update authorization on a view without having update authorization on the relations used to define the view
ii. The user who creates a view cannot be given update authorization on a view without having update authorization on the relations used to define the view
iii. If a user creates a view on which no authorization can be granted, the system will allow the view creation request
iv. A user who creates a view receives all privileges on that view

8. Which of the following statements is not correct?
i. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents
ii. MongoDB has official drivers for a variety of popular programming languages and development environments
iii. When compared to relational databases, NoSQL databases are more scalable and provide superior performance
iv. All of the above

9. To include integrity constraint on a table, which of the following is used?
i. Create Table ii. Modify Table iii. Alter Table iv. Drop Table

10. Which of the following is not an integrity constraint?
i. Not null ii. Positive iii. Unique iv. Check 'predicate'

11. Domain constraints, functional dependency and referential integrity are special forms of:
i. Foreign Key ii. Primary Key iii. Assertion iv. Referential Constraint

12. Which of the following is a wide-column store?
i. Cassandra ii. Riak iii. MongoDB iv. Redis

Answers of Section 2:

1. i 2. iii 3. ii 4. i 5. iv 6. ii 7. iii 8. iv 9. iii 10. ii 11. iii 12. i

Section 3: Exploratory Data Analysis

1. A person has been deputed to find the average income of the factory employees. To provide the correct picture of average income, the person should find out
i. Geometric Mean
ii. Weighted Mean
iii. Progressive Mean
iv. Arithmetic Mean

2. In a frequency distribution of a large number of values, the mode:
i. Largest observation
ii. Smallest value
iii. Observation with the maximum frequency
iv. The maximum frequency of an observation

3. The percentage of items in a frequency distribution lying between upper and lower quartiles is:
i. 80 per cent
ii. 40 per cent
iii. 50 per cent
iv. 25 per cent

4. In a discrete series having *(2K + 1)* observations, the median is:
i. K^{th} observation
ii. $(K + 1)^{th}$ observation
iii. $(K + 2)/2^{th}$ observation
iv. $(2K + 1)/2^{th}$ observation

5. Which of the following statements is not correct?
i. The bars in a histogram touch each other
ii. The bars in a column graph touch each other
iii. There are bar diagrams which are known as broken bar diagrams
iv. Multiple bar diagrams also exist.

6. Proportions of males and females in India in different occupations in the year 2019 can most properly be represented by:
i. Sliding bar diagram
ii. Deviation bar diagram
iii. Sub-divided bar diagram
iv. Multiple bar diagram

7. The data relating to the number of registered allopathic and homoeopathic doctors in six different states can be most appropriately represented by the diagram:
i. Line graph
ii. Histogram
iii. Pie-chart
iv. Double bar diagram

8. When for some countries, the magnitudes are small and for others, the magnitudes are very large, to represent the data, it is preferred to construct:
i. Deviation bar diagram
ii. Duo-directional bar diagram
Iii. Broken bar diagram
iv. Any of these

9. The shape of a trilinear chart is that of a:
i. Cone
ii.Cube
iii. Triangle
iv. Pyramid

10. Histograms can be used only when:
i. Class intervals are equal or unequal
ii. Class intervals are all equal
iii. Class intervals are unequal
iv. Frequencies in the class interval are equal

11. Histograms are suitable for the data presented as:
i. Continuous grouped frequency distribution
ii. Discrete grouped frequency distribution
iii. Individual series
iv. All of the above

12. The immigration and outmigration of people in a number of countries and also the net migration can be better displayed by:

i. Duo-directional column chart
ii. Gross-deviation column chart
iii. Net deviation column chart
iv. Range chart

Answers of Section 3:

1. ii 2. iii 3. iii 4. ii 5. ii 6. i 7. iv 8. iii 9. iii 10. ii 11. i 12. ii

Section 4: Supervised Learning and Unsupervised Learning

1. Which of the following statements is not correct about regularization?
i. Using too large a value of lambda can cause your hypothesis to underfit the data.
ii. Using too large a value of lambda can cause your hypothesis to overfit the data.
iii. Using a very large value of lambda cannot hurt the performance of your hypothesis.
iv. None of the above

2. How can you avoid the bad local optima issue while running a clustering algorithm?
i. Set the same seed value for each run
ii. Use multiple random initializations
iii. Both i and ii
iv. None of the above

3. In which of the following cases will K-means clustering fail to give good results?
A. Data points with outliers
B. Data points with different densities
C. Data points with nonconvex shapes

i. A and B
ii. B and C
iii. A and C
iv. All of the above

4. Which of the following is a reasonable way to select the number of principal components "k"?
i. Choose k to be the smallest value so that at least 99% of the variance is retained.
ii. Choose k to be 99% of m (k = 0.99*m, rounded to the nearest integer).
iii. Choose k to be the largest value so that 99% of the variance is retained.
iv. Use the elbow method

5. You run gradient descent for 15 iterations with α=0.3 and compute $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ decreases quickly and then levels off. Based on this, which of the following conclusions seems most plausible?
i. Rather than using the current value of a, use a larger value of a (say α=1.0)
ii. Rather than using the current value of a, use a smaller value of a (say α=0.1)
iii. α=0.3 is an effective choice of learning rate
iv. None of the above

6. Let a feature P can take certain values 1, 2, 3 and 4 and represents platform numbers at a railway station. Which of the following statements is true?
i. Feature P is an example of nominal variable
ii. Feature P is an example of ordinal variable
iii. Both i and ii.
iv. Neither i nor ii.

7. Which of the following hyperparameters, when increased, may result in overfitting of a random forest model?
A. Number of trees
B. Depth of tree
C. The learning rate
i. All of these
ii. Only A and B
ii. Only B
iv. Only B and C

8. How can we find the global minima in the K-Means algorithm?

A. By running the algorithm for different centroid initialization
B. By adjusting the number of iterations
C. By finding out the optimal number of clusters

i. All of these
ii.Only A and B
iii. Only B and C
iv. Only A and C

9. What is the consequence between a node and its predecessors while creating a Bayesian network?

i. Functionally dependent
ii. Dependant
iii. Conditionally independent
iv. Both conditionally dependent and independent

10. Computational learning theory analyzes the sample complexity and computational complexity of:

i. Unsupervised Learning
ii. Inductive learning
iii. Forced based learning
iv. Weak learning

11. Decision trees are appropriate for the problems where:

i. Attributes are both numeric and nominal
ii. Target function takes on a discrete number of values.
iii. Data may have errors
iv. All of the above

12. Which of the following is the advantage of Decision Trees?

i. Possible Scenarios can be added
ii. Use a white-box model, If the given result is provided by a model
iii. Worst, best and expected values can be determined for different scenarios
iv. All of the above

13. Why is second-order differencing in time-series needed?

i. To remove stationarity
ii. To find the maxima or minima at the local point
iii. Both i and ii
iv. None of the above

14. While performing regression or classification, which of the following is the correct way to preprocess the data?

i. Normalize the data → PCA → normalize PCA output → training
ii. Normalize the data → PCA → training
iii. PCA → normalize PCA output → training
iv. None of the above

15. Which of the following is an example of feature extraction?

i. Constructing a bag of words vector from an email
ii. Applying PCA projects to a large high-dimensional data
iii. Removing stop words in a sentence
iv. All of the above

16. Which of the following is true about Naive Bayes?

i. Assumes that all the features in a dataset are equally important
ii. Assumes that all the features in a dataset are independent
iii. Both i and ii

iv. None of the above

17. The factors which affect the performance of a learner system does not include:
i. Representation scheme used
ii. Training scenario
iii. Type of feedback
iv. Good data structures

18. Different learning models do not include:
i. Memorization ii. Analogy iii. Deduction iv. Introduction

19. What does the Bayesian network provide?
i. A complete description of the domain ii. Partial description of the domain
ii. A complete description of the problem iv. None of these

20. Which of the following is an example of active learning?
i. News Recommender system ii. Dust cleaning machine
iii. Automated vehicle v. None of the mentioned

21. Which of the following is also called exploratory learning?
i. Supervised learning ii. Active learning
iii. Unsupervised learning iv. Reinforcement learning

22. Which of the following is a widely used and effective model for machine learning based on the concept of bagging?
i. Decision Tree ii. K-Nearest Neighbors iii. AdaBoost iv. Random Forest

23. To find the minimum or the maximum of a function, we set the gradient to zero because:
i. The value of the gradient at extrema of a function is always zero
ii. Depends on the type of problem
iii. Both i and ii
iv. None of the above

24. Which of the following is a disadvantage of decision trees?
i. Factor Analysis ii. Decision trees are robust to outliers
iii. Decision trees are prone to overfit iv. None of these

Answers of Section 4:

1. iv	2. ii	3. iv	4. i	5. iii	6. ii	7. ii	8. i	9. iii	10. ii	11. iv	12. iv
13. iii	14. ii	15. iv	16. iii	17. iv	18. iv	19. i	20. i	21. ii	22. iv	23. i	24. iii

Section 5: Neural Networks and Deep Learning

1. Which of the following neural network models use unsupervised learning?
A. Autoencoders
B. Generative Adversarial Networks
C. Self-Organizing Maps
i. All of these ii.Only A and B
iii. Only B and C iv. Only A and C

2. What is differential Hebbian learning?
i. Synaptic strength is proportional to the correlation between the firing of post & presynaptic neuron
ii. Synaptic strength is proportional to the correlation between the firing of postsynaptic neuron only
iii. Synaptic strength is proportional to the correlation between the firing of presynaptic neuron only
iv. Synaptic strength is proportional to changes in correlation between the firing of post & presynaptic neuron

3. What is competitive learning?
i. Learning laws which modulate the difference between synaptic weight & output signal
ii. Learning laws which modulate the difference between synaptic weight & activation value
iii. Learning laws which modulate the difference between actual output & desired output
iv. None of these

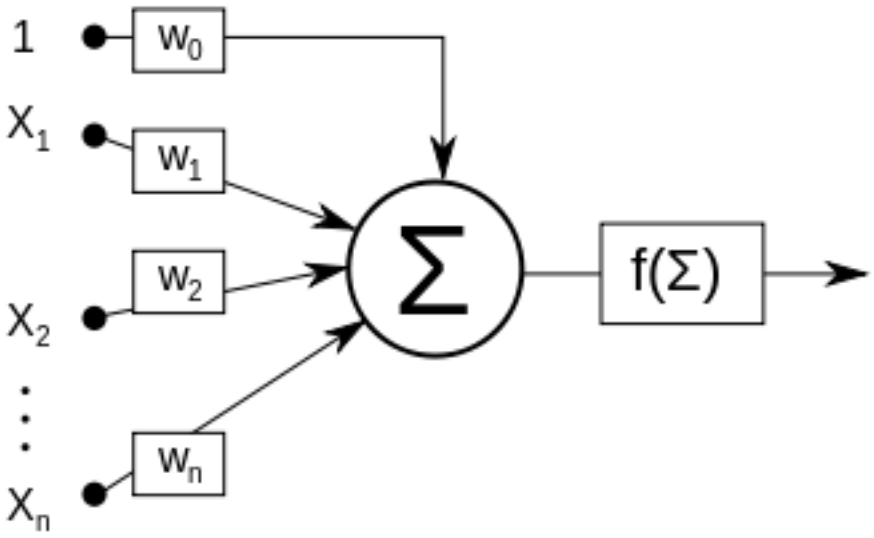
4. Which of the following statements is true when you use 1 x 1 convolution in a Convolutional Neural Network?
i. It can help in dimensionality reduction
ii. It can be used for feature pooling
iii. It suffers less overfitting due to small kernel size
iv. All of the above.

5. Consider the following statements:
A. It is possible to train a neural network well by initializing all the weights as 0
B. It is possible to train a neural network well by initializing biases as 0
Which of the above statements is correct?
i. Only A is correct ii. Only B is correct
iii. Both A and B are correct iv. None of these are correct

6. Suppose we wish to predict the probabilities of n classes (p_1, p_2, ..., p_n) such that sum of p over all n equals to 1, which of the following activation function should be used at the output layer of the neural network?
i. Softmax ii. ReLu ii. Sigmoid iv. Tanh

7. Ability to learn how to do tasks based on the data given for training or initial experience is called:
i. Self-organization ii. Adaptive learning
iii. Auto Associativity iii. None of these

8. What is the name of the neural network model given in the following diagram?



i. The Rosenblatt Perceptron Model
ii. McCulloch-Pitt's Model
iii. Widrow's Adaline Model
iv. Autoassociative Network

9. What is the fundamental difference between the Adaline and the perceptron model?
i. Weights are compared with output
ii. Sensory units result is compared with output
iii. Analog activation value is compared with output
iv. All of the above

10. Which of the following elements is not available in the architecture of a Gated Recurrent Unit (GRU), a variant of LSTM Recurrent Neural Network?
i. Cell ii. Input Gate iii. Output Gate iv. Forget Gate

11. What is the objective of the backpropagation algorithm?
i. To develop a learning algorithm for multilayer feedforward neural network
ii. To develop a learning algorithm for single-layer feedforward neural network
iii. To develop a learning algorithm for multilayer feedforward neural network, so that network can be trained to capture the mapping implicitly
iv. None of these

12. Which of the following statements is correct regarding backpropagation rule?
i. It is used in the feedback neural network
ii. The actual output of the network is determined by computing the outputs of units for each hidden layer
iii. Hidden layers output in the network is not important, they are only meant for supporting input and output layers
iv. None of these

13. What are the issues where biological neural networks prove to be superior to artificial neural networks?
i. Robustness and fault tolerance ii. Flexibility
iii. Collective computation iv. All of these

14. The activation function of a neuron in the neural network can be:
i. Linear ii. Non-linear iii. Can be linear or non-linear iv. None of these

15. What is Adaline in neural network models?

i. Adaptive Least Infinite Network
ii. Adaptive Linear Network
iii. Automatic Linear Element
iv. Adaptive Linear Element

16. Widrow-Hoff Learning in neural networks is a special case of:
i. Hebbian Learning
ii. Perceptron Learning
iii. Delta Learning
iv. None of these

17. What is the other name of Widrow-Hoff Learning in neural networks?
i. Hebb Rule
ii. Least Mean-Square Rule
iii. Sum of Squares Rule
iv. None of these

18. Correlation Learning Law is the special case of:
i. Hebb Rule
ii. Perceptron learning rule
iii. Least Mean-Square Rule
iv. Delta learning rule

Answers of Section 5:

1. i	2. iv	3. i	4. iv	5. iii	6. i	7. ii	8. ii	9. iii	10. iii	11. iii	12. ii
13. iv	14. iii	15. iv	16. iii	17. ii	18. i						

Section 6: Natural Language Processing

1. Which of the following techniques can be used for the purpose of keyword normalization where a keyword is converted into its base form?

A. Lemmatization
B. Stemming
C. Levenshtein

i. Only A and B ii. Only B and C iii. Only A and C iv. All of these

2. Which of the following statements are correct about topic modelling?

A. It is a supervised learning technique
B. LDA (Linear Discriminant Analysis) can be used to perform topic modelling
C. Selection of the number of topics in a model does not depend on the size of data
D. The number of topic terms is directly proportional to the size of the data

i. All of these
ii. Only B and D
ii. Only C and D
iv. None of these

3. In Latent Dirichlet Allocation model for text classification purposes, what do alpha and beta hyperparameters represent-

i. Alpha is the number of topics within documents and beta is the number of terms within topics
ii. Alpha is the density of terms generated within topics and beta is the density of topics generated within terms
iii. Alpha is the number of terms within documents and beta is the number of terms within topics
iv. Alpha is the density of topics generated within documents and beta is the density of terms generated within topics

4. Which of the following statement is/are correct about Word2Vec model?

i. The architecture of word2vec consists of only two layers – a. continuous bag of words and b. skip-gram model
ii. Continuous bag-of-word (CBOW) is a Recurrent Neural Network model
iii. Both CBOW and Skip-gram are shallow neural network models
iv. All of the above

5. What is/are the main challenge(s) faced in natural language processing?

i. Handling ambiguity of sentences
ii. Handling tokenization
iii. Handling POS-Tagging
iv. All of the above

6. Which of the following models can be used for the purpose of document similarity?

i. Training a word 2 vector model on the corpus that learns context present in document
ii. Training a bag of words model that learns the occurrence of words in document
iii. Creating a document-term matrix and using cosine similarity for each document
iv. All of the above

7. What can be the role of natural language processing in collaborative filtering and content-based filtering algorithms?

i. Feature Extraction from text
ii. Measuring Feature Similarity
iii. Engineering Features for vector space learning model
iv. All of these

8. What is the main difference between Conditional Random Field (CRF) and Hidden Markov Model (HMM)?

i. CRF is Generative whereas HMM is Discriminative model
ii. CRF is Discriminative whereas HMM is Generative model
iii. Both CRF and HMM are Generative model
iv. Both CRF and HMM are Discriminative model

9. What is the field of Natural Language Processing?
i. Computer Science ii. Artificial Intelligence iii. Linguistics iv. All of these

10. In which of the following areas the natural language processing can be applied?
i. Automatic text summarization
ii. Automatic question-answering system
iii. Information retrieval
iv. All of the above

11. In linguistic morphology, which of the following is the process for reducing inflected words to their root form?
i. Stemming ii. Rooting iii. Text-proofing iv. Both i and ii.

12. Which of the following techniques is not part of the flexible text matching?
i. Soundex ii. Metaphor iii. Edit Distance iv. Keyword Hashing

Answers of Section 6:
1. i 2. iv 3. iv 4. iii 5. i 6. iv 7. iv 8. ii 9. iv 10. iv 11. i 12. iv

Section 7: Computer Vision

1. Let an input signal *x = (2, 3, 80, 6)* and we are using a window size of three with one entry immediately preceding and following each entry. What will be the output signal *y* if we are using a median filter?
i. *y = (2, 6, 80 ,3)*
ii. *y = (3, 6, 6, 3)*
iii. *y = (2, 6, 6, 3)*
iv. *y = (2, 6, 6, 2)*

2. Which of the following steps may be used to avoid boundary issues in image processing?
A. Avoid processing the boundaries, with or without cropping the signal or image boundary afterwards
B. Fetching entries from other places in the signal. With images, for example, entries from the far horizontal or vertical boundary might be selected
C. Shrinking the window near the boundaries, so that every window is full
i. All of the above
ii. Only A and B
iii. Only B and C
iv. Only A and C

3. In geometric mean filters where alpha is equal to 1, then it works as a:
i. Notch filter
ii. Bandpass filter
iii. Wiener filter
iv. Inverse filter

4. To avoid the negative values taking absolute values in Laplacian image doubles:
i. Thickness of line
ii. Thinnes of line
iii. Thickness of edge
iv. Thinnes of edge

5. If pixels in the image are very different in colour or intensity from their surrounding pixels; the defining characteristic is that the value of a noisy pixel bears no relation to the colour of surrounding pixels. What is the type of noise present in the image?
i. Gaussian noise
ii. Salt and pepper noise
iii. Shot noise
iv. Quantization noise

6. In _____, each pixel in the image will be changed from its original value by a (usually) small amount. A histogram, a plot of the amount of distortion of a pixel value against the frequency with which it occurs, shows a normal distribution of noise.
i. Gaussian noise
ii. Salt and pepper noise
iii. Shot noise
iv. Quantization noise

7. One method to remove noise is by convolving the original image with a mask that represents a _____ or smoothing operation.
i. Band-pass filter
ii. High-pass filter
iii. Low-pass filter
iv. Narrow-pass filter

8. The method to remove noise by evolving the image under a smoothing partial differential equation similar to the heat equation is called:
i. Linear smoothening filtering
ii. Non-local means
iii. Anisotropic diffusion
iv. None of the above

9. In image processing, a median filter is an example of:
i. Linear filter
ii. Nonlinear filter
iii. Kernel filter
iv. None of these

10. A continuous image is digitised at _____ points.
i. Random ii. Vertex iii. Contour iv. Sampling

11. What is the term referred to the transition between continuous values of the image function and its digital equivalent?
i. Sampling ii. Quantization iii. Rasterization iv. None of these.

12. The dynamic range of the imaging system is a ratio where the upper limit is determined by:
i. Saturation ii. Noise ii. Brightness iv. Contrast

Answers of Section 7:
1. iii 2. i 3. iv 4. i 5. ii 6. i 7. iii 8. iii 9. i 10. iv 11. ii 12. i

Section 8: Deployment and Model Management

1. Which of the following may be the reasons when A model with thousands of features to attain an accuracy of more than 90% on evaluation might not be good enough for deployment?
A. Portability
B. Scalability
C. Operationalization
i. All of the above
ii. Only A and B
iii. Only B and C
iv. Only A and C

2. Which of the following are the ways of training machine learning models into production?
A. One-off
B. Batch
C. Real-Time/Online
i. All of the above
ii. Only A and B
iii. Only B and C
iv. Only A and C

3. What is the batch training of machine learning models?
i. A model is trained ad-hoc and pushed to production until its performance deteriorates enough that they are called upon to refresh it.
ii. The training that allows having a constantly refreshed version of your model based on the latest train.
iii. Both i and ii.
iv. None of these

4. What is the application of gain and lift charts in machine learning model evaluation?
i. It measures the performance of classification models
ii. It measures the performance of local resources
iii. It checks the rank ordering of the probabilities
iv. None of these

5. Which of the following machine learning architectures will be employed when training and persisting are done offline while prediction is done in real-time?
i. Train by batch, predict by batch, serve through a shared database
ii. Train by batch, predict on the fly, serve via REST API
iii. Train, predict by streaming
iv. Train by batch, predict on mobile

6. Which of the following layers in a machine learning architecture is used to monitor production models where it is checked how closely the predictions on live traffic matches the training predictions?
i. Scoring layer
ii. Evaluation layer
ii. Feature layer
iv. Data Layer

7. Which of the following layers in a machine learning architecture transforms features into predictions?
i. Scoring layer
ii. Evaluation layer
ii. Feature layer
iv. Data Layer

8. What is entanglement in machine learning model development?
i. There are some data inputs which are unstable and change over time

ii. When models are constantly iterated on and subtly changed, tracking config updates whilst maintaining config clarity and flexibility becomes an additional burden
iii. If we have an input feature which we change, then the importance, weights or use of the remaining features may all change as well
iv. Machine learning systems require cooperation between multiple teams, which can result in no single team or person understanding how the overall system works, teams blaming each other for failures, and general inefficiencies

9. The process in which we integrate a machine learning model into an existing production environment to make practical business decisions based on data is called:
i. Model verification ii. Model evaluation
iii. Model deployment iv. Model scraping

10. The machine learning model deployment is the _____ stage of machine learning life cycle.
i. First ii. Last iii. Second last iv. None of these

11. What is a base-line machine learning model?
i. A model having the minimum possible number of features but with good evaluation measures.
ii. A model having the maximum possible number of features but with fewer evaluation measures.
iii. A model having the minimum possible number of features and no evaluation measures.
iv. None of these

12. What does CI/CD stand for?
i. Combined Integration and combined deployment
ii. Continuous integration and continuous deployment
iii. Classical integration and classical deployment
iv. None of these

Answers of Section 8:
1. i 2. i 3. ii 4. iii 5. ii 6. ii 7. i 8. iii 9. iii 10. ii 11. i 12. ii

Section 9: Python and R

1. What is the output of the following Python code?
print([i.lower() for i in "HELLO"])
i. ['h', 'e', 'l', 'l', 'o']
ii. 'hello'
iii. ['hello']
iv. hello

2. Let a list in Python L = [1, 2, 2, 3]. What will be the output of print(L*2)?
i. [1, 2, 2, 3, 1, 2, 2, 3] ii. [2, 4, 4, 6]
iii. [2, 4, 4, 6, 2, 4, 4, 6] iv. [1, 4, 4, 9]

3. Consider the following Python code:

```
S = [['him', 'sell'], [90, 28, 43]]

print(S[0][1][1])
```

What is the output if the above code is executed?
i. 'i' ii. 'e' ii. '90' iv. 'h'

4. How many times 'Welcome to Python!' will be printed if the following Python is executed?

```
a=0
while a<10:
   print('Welcome to Python!')
   pass
```

i. 9 ii. 10 iii. 11 iv. Infinite number of time

5. What is the output of the following R code?

```
x <- c("a", "b")
as.numeric(x)
```

i. [1] 1 2
ii [1] TRUE TRUE
iii. [1] NA NA (Warning message: NAs introduced by coercion)
iv. [1] NaN

6. Which of the following is the use of id() function in Python?
i. Every object does not have a unique id.
ii. id returns the identity of the object
iii. Both i and ii
iv. Neither i nor ii

7. What will be the output of the following Python code?

```
str = 'hello'
print(str[:-2])
```

i. he ii. hel iii. lo iv. llo

8. What will be the output of the following R code snippet?

```
paste("a", "b", se = ":")
```

i. "a+b" ii. "a=b" iii. "a b:" iv. None of these

9. You can check to see whether an R object is NULL with the _____ function.
i. is.null() ii. is.nullobj() iii. null() iv. as.nullobj()

10. What is the class defined in the following R code?

```
y <-  c(FALSE, 2)
```

i. Character ii. Numeric iii. Logical iv. Integer

11. In Python, which of the following keywords is used with function?
i. define ii. fun iii. def iv. function

12. Which of the following is not a core data type in Python?
i List ii. Dictionary iii. Tuple iv. Class

13. Given a function that does not return any value, What value is thrown by default when executed in the shell.
i. int ii. bool iii. void iv. None

14. In R programming, which of the following functions is used to create matrices by row binding?
i. rjoin() ii. rbinding() iii. rowbind() iv. rbind()

15. In R programming, what is the function used to test objects (returns a logical operator) if they are NA?
i is.na() ii. is.nan() iii. as.na() iv. as.nan()

Answers of Section 9:
1. i 2. ii 3. ii 4. iv 5. iii 6. ii 7. ii 8. iv 9. i 10. ii 11. iii 12. iv
13. iv 14. iv 15. i

Section 10: Business and Data Science

1. In a data science project, who are the key stakeholders in the business understanding phase?
A. Business end-users
B. Data analysts
C. Business analysis
i. All of the above
ii. Only A and B
iii. Only B and C
iv. Only A and C

2. "Miscommunication between data scientists and the data engineer, leading to poor identification of necessary and available data sources". This issue refers to which of the following phases of a data science project?
i. Business understanding
ii. Data understanding
iii. Data preparation
Iv. Model deployment

3. What are the characteristics of best decision models?
i. Accurately reflect relevant characteristic if the real-world object or decision
ii. Are mathematical models
iii. Replicate all aspects of the real-world object or decision
iv. Replicate the characteristic of a component in isolation from the rest of the system

4. Which of the following is not a characteristic of data?
i. Statistics are collected by enumeration or estimation
ii. Statistics are placed in relation to each other
iii. Human being
iv. Comparative study

5. In business analysis, which of the following is a performance management tool that recapitulates an organization's performance from several standpoints on a single page?
i. Balanced Scorecard
ii. Data Cube
iii. Dashboard
iv. All of these

6. Which of the following is the process of basing an organization's actions and decisions on actual measured results of performance?
i. Institutional performance management
ii. Gap analysis
iii. Slice and Dice
iv. None of these

7. Which of the following correctly specifies the outcome of engagement of stakeholders in the data understanding phase?
i. how important it is to have a clean database for a correct analysis
ii. Illustrate the possibilities with well-designed examples and set realistic expectations
iii. Picture the benefits of the data project
iv. None of these

8. The essence of decision analysis is:
i. Breaking down complex situations into manageable elements
ii. Choosing the best course of action
iii. Finding the root cause of why something has gone wrong
iv. Thinking ahead to avoid negative consequences

9. Which type of analytics gain insights from historical data with reporting, scorecards, clustering etc.
i. Decisive ii. Descriptive iii. Predictive iv. Prescriptive

10. What is to be used when faced with the decision of how to arrange furniture in a room?
i. Mathematical model ii.Mental model
iii. Physical model iv. Visual model

11. The method that takes data from a given data set and rather than looking at all users as one unit, it breaks them into related groups for analysis, is known as:
i. Behavioural analytics ii. Cohort Analysis
iii. Collection analytics iv. Contextual data modelling

12. Which of the following is the characteristic of data?
i. Aggregate of fact ii. Numerically expressed
iii. Affected by various causes iv. All of these

13. Which of the following analytics recommend decisions using optimization, simulation etc.?
i. Decisive ii. Descriptive iii. Predictive iv. Perspective

14. Which of the following analytics supports human decisions with visual analytics that the user models to reflect reasoning?
i. Decisive ii. Descriptive iii. Predictive iv. Perspective

15. Which of the following represents data?
i. A single value ii. Only two values in a set
iii. A group of values in a set iv. None of these

16. Which type of analytics uses statistical and machine learning techniques?
i. Decisive ii. Descriptive iii. Predictive iv. Perspective

17. Data which is taken from the publication, "Agricultural Situation in India" will be considered as:
i. Primary data ii. Secondary data
Iii. Primary and secondary data iv. Neither primary nor secondary data

18. Which of the following examples does not constitute an infinite population?
i. A population consisting of odd numbers
ii. The population of weights of newly born babies
iii. The population of heights of 15-year old children
iv. The population of heads and tails in tossing a coin successively

Answers of Section 10:

1. iv	2. ii	3. i	4. iii	5. i	6. i	7. iv	8. ii	9. ii	10. iv	11. ii	12. iv
13. iv	14. i	15. iii	16. iii	17. ii	18. iii						