22BDAH101 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. FIRST SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Data Warehousing and Data Mining

Duration:3 Hours Max Marks:70

PART A

Answer any FOUR of the following **(4×5= 20)**

1) What are the different types of learning in machine learning? Explain each of them in brief.

2) Explain briefly how to improve the efficiency of Apriori-based mining algorithm.

3) Explain the characteristics of Partitioning method and Density-based method in Clustering.

4) Explain the concept of Mining Sequence Data.

5) Explain different types of biclusters with suitable illustrations.

PART B

Answer any FIVE questions selecting at least one question from each unit

(5×10= 50)

UNIT-I

6) Elaborate on Proximity Measures for Nominal Attributes.

7) Describe the four methods for the generation of concept hierarchies for nominal data with suitable examples.

UNIT-II

8) Describe the concept of prediction mining in cube space with a suitable example.

9) Explain OLAP operations for multidimensional data with suitable example.

UNIT-III

10) Elaborate Naive Bayesian Classification with an illustration.

11) Explain SVM algorithm with respect to :

a. Linearly Separable Data b. Linearly Inseparable Data

UNIT-IV

12) Elaborate on Mining Graphs and Networks with illustrations.

13) Explain proximity-based methods and clustering-based methods for detecting outlier.

****

22BDAH102 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. FIRST SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Python Programming for Data Analytics

Duration:3 Hours Max Marks:70

PART A

Answer any FOUR of the following **(4×5= 20)**

1) Develop a Python program to load a csv file and find
   (i) Head (ii) Tail (iii) Median (iv) Standard Deviation (v) Size

2) Explain Pandas index object with suitable examples.

3) Explain NumPy's sort() function with its syntax.

4) Describe indexing and selection with respect to Time Series.

5) Briefly explain about the Text and Annotation function available in matplotlib.

PART B

Answer any FIVE questions selecting at least one question from each unit

(5×10= 50)

UNIT-I

6) a) Explain the basic Data types in python (4)
   b) List the general rules that need to be followed while creating a variable.(6)

7) Describe the term actual arguments. Explain the four types of actual arguments used in a function call.

UNIT-II

8) (a) Explain Pandas Multiply Indexed Series. (5)
   (b) Describe the function of read_csv()? Explain with an example. (5)

9) (a) Explain concatenation along an axis with suitable examples. (5)
   (b) Discuss on reshaping and pivoting in pandas. (5)

UNIT-III

10) Enumerate and explain various options of pivot tables.

11) Describe the following with suitable examples:
    (a) GroupBy Mechanics (b) filtering and transformation (5+5)

UNIT-IV

12) (a) Explain the concept of subplots in Matplotlib. (5)
    (b) Describe the usage of matplotlib as a visualization tool. (5)

13) (a) Describe visualization with Seaborn. (5)
    (b) Explain any five map-specific methods used to plot data on maps. (5)

****

22BDAH103 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. FIRST SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Quantitative Techniques for Data Science & Data Visualization

Duration:3 Hours Max Marks:70

PART A

Answer any FOUR of the following **(4×5= 20)**

1) If $A = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 5 \\ 0 & 1 & 2 \end{bmatrix}$ Solve A+A' and A-A' and prove that $(\frac{A+A'}{2}) + (\frac{A-A'}{2}) = A$

2) Distinguish between Raw moments and Central moments. How do these moments explain the characteristics of any data distribution?

3) What do you mean by conditional Probability? State Addition and Multiplication Theorem of Probability.

4) a. Distinguish between unbiased and consistent test procedure. (3)
   b. State NP Lemma for finding the best critical region. (2)

5) "Histograms and qq plots are the important visualization techniques in data analysis". Justify.

PART B

Answer any FIVE questions selecting at least one question from each unit **(5×10= 50)**

UNIT-I

6) Obtain the solution for the following equations by
   a) matrix method and b) Cramer's method.
   i) 4 (y – x) = 5z - 22
   ii) 3z + 4x = 6y + 2
   iii) z - 3y = 14 - 10x

7) What is a set? Explain different types of sets with an example.

UNIT-II

8) Describe line diagram, vertical bar diagram, horizontal bar diagram, pie diagram and Histogram with examples.

9) a) State Bayes Theorem. (2)

b) Cheryl has two bags. Bag 1 has 6 red and 3 blue balls and bag 2 has 7 red and 8 blue balls. Cheryl draws a ball at random and it turns out to be red. Determine the probability that the ball was from the bag 1. (8)

UNIT-III

10) The correlation coefficient between height and weight is 0.424 for 90 vegetarians and 0.578 for 130 Nonvegetarians. Test whether the two correlation coefficients differ significantly.

11) a) Write down the p.m.f, mean and variance of Poisson distribution. (3)

b) If X is a Poisson variate such that $P(X=2) = 30\ P(X=4) + 9\ P(X=6)$. Find $P(X \geq 2)$, $P(X<2)$ and also find the mean and SD. (7)

UNIT-IV

12) Why the Polishing and color palettes are needed for data visualization ? Explain its role on data visualization along with various types of Graphical formats available in ggplot2.

13) What is displaying distribution ? Explain.

****

22BDAS104 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. FIRST SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Optimization Techniques

Duration:3 Hours Max Marks:70

PART A

Answer any FOUR of the following **(4×5= 20)**

1) Discuss the origin and development of OR. What are the limitations of OR? How has the computer helped in popularising OR?

2) A certain paint requires 2 ingredients A and B. Paint manufacturer has to produce 100 kg of a particular paint using the ingredients A and B. Ingredient A costs Rs. 300 per KG While B costs Rs. 500 keg's. As per the blending requirement not more than 40 kgs of ingredient A and at least 30 kgs of ingredient B must be used. Formulate this as LPP so as to minimize the cost of paint blended.

3) What is the role of OR in decision making? Explain.

4) What do you understand by "unboundness" in linear programming? When does this situation arise?

5) Explain the different terms used in queuing models.

PART B

Answer any FIVE questions selecting at least one question from each unit:

(5×10= 50)

UNIT-I

6) Explain the developments in OR before the Second World War.

7) How does Operations Research helps the companies to find the better solutions to their problems? Explain with an example.

UNIT-II

8) A Television repairman finds that the time spent on his jobs has an exponential distribution with mean 30 minutes. If he repairs the sets in the order in which they come in, and if the arrivals of sets are approximately Poisson with an average rate of 10 per 8 hours day what is the repairs man idle time each day? Find the expected number of units in the system and in the queue?

9) Solve the problem using simplex method.

Maximizing profit $7x+5y$

Subject to $2x+y \leq 100$

$4x+3y \leq 240$

$x+y \geq 0$

UNIT-III

10) Describe Concave and Convex functions in detail.

11) What is Non-linear Programming? Explain its classification in detail.

UNIT-IV

12) Explain the steps in Decision Theory approach.

13) What the advantages and limitations of Simulation Technique.

****

21BDH301 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. THIRD SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Data Analytics using SPARK

Duration: 3 Hrs Max Marks:70

PART - A

I. Answer any EIGHT of the following: (2×8= 16)

a) Describe the concept of Resilient Distributed Dataset or RDD.

b) Write code to demonstrate initializing SparkContext in Python.

c) Enumerate four Industry specific use cases of Apache Spark.

d) Discuss the role of the driver component in Spark distributed mode.

e) What do you understand by RDD's being immutable in spark?

f) Explain the combineByKey() pair RDD operation with an example.

g) Which is the class used for the primary configuration mechanism in Spark?

h) Mention the four sections that the Spark UI contains for providing information as applications execute.

i) List any four machine learning algorithms supported by MLlib.

j) Discuss the term Checkpointing in Spark SQL.

PART - B

Answer any FOUR questions : (6×4= 24)

2. Explain the components for distributed execution in Spark with an illustration.

3. Describe four data types that are specific to MLlib package.

4. Discuss the common supported file formats in Apache Spark.

5. Discuss the three methods that need to be implemented when implementing custom partitioners

6. Explain the benefit of running Spark on YARN.

7. Consider a dataframe containing movie details: Movie_ID, Title, Category, User_ID, Rating. Write queries for the following in Spark SQL: 1. List the details of movies that have been reviwed by user_ID 1, 2 and 3 in increasing order of rating. 2. List the 10 lowest ranked moves. 3. Retrieve the category of the movies that user_id = 5 has watched the most and display the movie details for the movies that are best rated in this category.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. Consider two RDDs, RDD1 and RDD2 containing string data. Describe 4 operations that belong to the mathematical set using these RDD's.

9. Highlight any 10 features of Apache Spark

10. Discuss any six Numeric RDD operations used in descriptive statistics on RDDs. Write a python program to remove outliers from a log file containing log data. Consider suitable outlier of your choice.

11. Explain the role of memory, cores, executors and the number of local disks on the effect of completion time of the application.

12. Explain the Spark Streaming process with a detailed workflow diagram.

****

21BDH302 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. THIRD SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Artificial Intelligence

Duration:3 Hrs Max Marks:70

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) State the two definitions of Artificial Intelligence from the rationalist approach.

b) Enumerate the four ways to evaluate an algorithms performance.

c) Give examples to differentiate between atomic sentences and complex sentences in propositional logic.

d) Using a game example, highlight the basic principle followed in partially observable games.

e) Why is Hill-climbing sometimes called greedy local search?

f) With examples differentiate between Relations and Functions in First Order Logic.

g) With an example, discuss the concept of Markov chain.

h) Describe a Dynamic Bayesian Network.

i) Discuss the fundamental idea behind current-best hypothesis search.

j) Discuss the term overfitting.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. Consider the Wumpus world example. Write the PEAS description of its task environment.

3. With an example of the 8-puzzle problem, describe the concept of Heuristics in search problems.

4. Explain the different elements required to formally define a game as a kind of search problem

5. Consider the example where the student's grade, depends not only on his/her intelligence but also on the difficulty of the course, represented by a random variable D whose domain = {easy, hard}. The student asks his professor for a recommendation letter. The professor only looks at the students grade, and writes the letter for him based on that information alone. The quality of the letter is a random variable L, whose domain is {strong, weak}. The actual quality of the letter depends stochastically on the grade. The domain also has the student's Exam score dependent on his intelligence. All of the variables except Grade are binary-valued, and Grade is ternary-valued. Draw the Bayesian network and compute the number of entries for the joint distribution of this example.

6. Explain the term Sample Space and Belief State with an example each.

7. With an example explain the generalization process called Relevance Based Learning.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. Explain the concept of depth-limited search strategy and write the pseudocode for its recursive implementation.

9. Given the full joint distribution for the toothache, cavity, catch world, compute the following:
   1. P(cavity v toothache)
   2. P(cavity)
   3. P(cavity | toothache)
   4. P(¬cavity | toothache)
   5. P(toothache)

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

10. Illustrate the procedure to convert the sentence "Everyone who loves all animals is loved by someone" to Conjunctive Nomal Form.

11. Discuss the working principle of Expectation-Maximization algorithm with a suitable example.

12. Differentiate between a Utility-based agent, Q-learning agent and Reflex agent.

****

21BDS304 Reg No : .....

CHOICE BASED CREDIT SYSTEM

M.Sc. THIRD SEMESTER DEGREE EXAMINATION JANUARY 2023

BIG DATA ANALYTICS

Cloud Computing

Duration:3 Hrs Max Marks:70

PART - A

I. Answer any EIGHT of the following: **(2×8= 16)**

a) Define Microsoft Azure.

b) Describe the term Privileged instructions

c) List the various tasks performed by scheduler in IaaS.

d) List the three services provided by foundation services.

e) List the methods and properties of IService interface.

f) Define Explicit threading.

g) Describe the term ' Embarrassingly Parallel Applications'.

h) Discuss the term data intensive computing.

i) Define Sphere.

j) Describe web role in Azure operating system.

PART - B

Answer any FOUR questions : **(6×4= 24)**

2. Describe the concept of virtualization and cloud computing.

3. Discuss on security, trust and privacy as a challenge in cloud computing.

4. Elaborate on hybrid cloud deployment mode.

5. Explain HTC and MTC.

6. Elaborate on the operations used in task libraries.

7. Describe on Resource naming and buckets hosted on S3 distributed storage.

PART - C

Answer any THREE questions : **(10×3= 30)**

8. Explain with an illustration the cloud computing reference model.

9. Explain different types of cloud.

10. Explain the limitations of Aneka thread compared to local thread.

11. Elaborate on programming abstractions in MapReduce Programming model.

12. Elaborate on social networking applications that uses cloud computing technologies.

****