Question Paper Code : 91762

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2019

Seventh Semester

Information Technology

IT6006 – DATA ANALYTICS

(Common to Computer Science and Engineering)

(Regulations 2013)

(Also common to PTIT6006 – Data Analytics for B.E. (Part-Time), Seventh Semester – Computer Science and Engineering – Regulations 2014)

Time : Three Hours   Maximum : 100 Marks

Answer ALL questions

PART – A   (10×2=20 Marks)

1. Define prediction error for a set of sensitive and insensitive instances.
2. What is the need for big data analytics and how it evolved in data science ?
3. Define support and confidence of an itemset.
4. How are moments estimated ?
5. Present an instance to understand Simpson's paradox.
6. What is a data stream management system ?
7. How to determine whether a point is closer to a cluster or not ?
8. What is limited-pass algorithm ?
9. What is inter-cluster data copying ?
10. What is difference between conventional NAS and HDFS ?

PART – B   (5×13=65 Marks)

11. a) i) Explain the big data dimensions and the data value chain.   (7)

   ii) Elaborate the features of any 3 modern data analytical tools.   (6)

(OR)

   b) i) List the differences between the process of reporting and analysis.   (7)

   ii) Write a short note on sampling and their characteristics.   (6)

12. a) i) Discuss the various components of time series and elaborate the Autoregressive Moving Average (ARMA) model. (7)

    ii) Explain the rule induction using sequential covering algorithm and perform rule based classification for an example scenario. (6)

    (OR)

    b) Discuss the Fuzzy decision tree algorithm and stopping criterion for decision making using an example. (13)

13. a) i) Explain and analyze the algorithm which eliminates most of the tuples that do not meet the criteria during filtering of streams. (7)

    ii) Explain the algorithm to determine distinct elements appearing in a stream. (6)

    (OR)

    b) Explain the methods of selecting a subset of a stream so that queries about the selected subset can have the answers statistically as a representative of the stream as a whole. (13)

14. a) Elaborate the process of hierarchical clustering in Euclidian and non-Euclidian space. (13)

    (OR)

    b) Write a short note on K-Means algorithm and explain how it could be used for handling streaming data. (13)

15. a) Explain and illustrate the various core components of Hadoop framework. (13)

    (OR)

    b) Explain the following related to No-SQL :

    i) Features-Schema less property of No-SQL. (3)

    ii) Difference from the conventional RDBMS. (3)

    iii) Any one type of No-SQL database. (7)

PART – C   (1×15=15 Marks)

16. a) Justify why SVM is effective on high dimensional data and discuss the polynomial kernel function for multiple classes. (15)

    (OR)

    b) Design a real-time sentiment analysis framework for online movie review application and explain the various steps involved to understand the opinion of the viewer. (15)