Reg. No. : □□□□□□□□□□□□

Question Paper Code : 20730

B.E./B.Tech. DEGREE EXAMINATION, NOVEMBER/DECEMBER 2018.

Seventh Semester

Information Technology

IT 6006 — DATA ANALYTICS

(Common to Computer Science and Engineering)

(Regulations 2013)

(Also common to PTIT 6006 – 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. Why you need to tame big data?

2. State the benefits of analytic sandbox.

3. Distinguish between linear and nonlinear regression.

4. What are the parameters used to characterize any fuzzy membership function?

5. State example for stream sources.

6. What is the storage requirement for the DGIM algorithm?

7. What are the strength and weakness of clique?

8. If we use a triangular matrix to count pairs and n, the number of items is 20, what pair's count is in a [100].

9. State the map-reduce strength and weakness.

10. What are the benefits of visual data exploration?

PART B — (5 × 13 = 65 marks)

11. (a) (i) What make a great analysis? State reason with example. (6)

(ii) Distinguish between attrition and Response modeling. (7)

Or

(b) What are prediction error? State and explain the prediction error in regression and classification with suitable example.

12. (a) (i) Distinguish between supervised and unsupervised learning with example. (6)

(ii) Given the following 3D input data, identify the principal component. 1 1 9; 2 4 6 ; 3 7 4; 4 11 4; 5 9 2. (7)

Or

(b) Consider a Kohonen self-organizing net with two cluster units and five input units. The weight vectors for the cluster Units are given by

$W_1 = [1.0, 0.9, 0.7, 0.5, 0.3]$

$W_2 = [0.3, 0.5, 0.7, 0.9, 1.0]$

Use the square of Euclidean distance to find the winning cluster unit for the input pattern $x = [0.0 ,0.5, 1.0, 0.5, 0.0]$. Using a learning rate of 0.25, find the new weights for the winning unit. [Hint : the winner unit is one with smaller index].

13. (a) (i) Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What is the third moment of this stream? (6)

(ii) State and explain the function of Bloom Filtering with example. (7)

Or

(b) (i) Suppose that A, B, C, D, E, and F are all the items. For a particular support threshold, the maximal frequent item sets are {A, B, C} and {D, E}. What is the negative border? (6)

(ii) State and explain the real-time analytics platform applications. (7)

14. (a) (i) Given a one dimensional dataset {1, 5, 8, 10, 2}, use the agglomerative clustering algorithms with the complete link with Euclidean distance to establish a hierarchical grouping relationship. By using the maximal lifetime as the cutting threshold, how many clusters are there? What is their membership in each cluster? (6)

(ii) State and explain the clustering in non-Euclidean space with example. (7)

Or

(b) Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the *k*-means algorithm for 1 epoch only. At the end of this epoch show

(i) The new clusters

(ii) The centres of the new clusters

(iii) How many more iterations are needed to converge? Draw the result for each epoch.

15. (a) (i) What are the relationships between parallel databases and Big data with respect to three V's? (6)

(ii) What is the features of MapR distribution? State and explain the architecture for MapR. (7)

Or

(b) (i) What is the purpose of sharding? Explain the process of sharding in MongoDB. (6)

(ii) What are the visualization techniques used visualizing data? Explain any two approaches. (7)

PART C — (1 × 15 = 15 marks)

16. (a) Here is a collection of twelve baskets. Each contains three of the six items 1 through 6.

{1, 2, 3} (2, 3, 4} {3, 4, 5} {4, 5, 6}

{1, 3, 5} {2, 4, 6} {1, 3, 4} (2, 4, 5}

{3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}

Suppose the support threshold is 4. On the first pass of the PCY Algorithm we use a hash table with 11 buckets, and the set {i, j} is hashed to bucket i×j mod 11.

(i) By any method, compute the support for each item and each pair of items.

(ii) Which pairs hash to which buckets?

(iii) Which buckets are frequent?

(iv) Which pairs are counted on the second pass of the PCY Algorithm?

Or

(b) Explain any two clustering techniques with suitable example.