Xi'an Jiaotong-Liverpool University
西交利物浦大学

PAPER CODE	EXAMINER	DEPARTMENT	TEL
CSE313		**Computer Science and Software Engineering**	

1ST SEMESTER 2018/19 EXAMINATIONS

BACHELOR DEGREE – Year 4

Big Data Analytics

TIME ALLOWED: Two Hours

INSTRUCTIONS TO CANDIDATES

1. Total marks available are 100. Marks for this examination account for 70% of the total credit.
2. The numbers on the right indicate the marks available.
3. Answer ALL questions in all sections.
4. Answers should be written in the answer booklet(s) provided and clearly mark question numbers and write "**Answer:**".
5. The university approved calculator – CASIO FS82ES/83ES can be used.
6. Only answers in English are accepted.

THIS PAPER MUST NOT BE REMOVED FROM THE EXAMINATION ROOM

Part A. Big Data Analytics foundations and Social issues (20 marks)

A1. What are the three major benefits Big Data Analytics can bring to a business organization? [6]

A2. Explain three levels of data definition with a consideration of data as an object. [3]

A3. Use the simplest statistical term to explain the four basic data attribute types. [4]

A4. What are the basic measurements in data quality? Explain the view of data quality as "Fitness for use" and why is that? [7]

Part B. Big Data Analysis Platforms (20 marks)

B5. **[Distributed Computing]** What is the CAP theorem in distributed computing? Provide at least one example for each option. **[5]**

B6. **[Hadoop]** Explain the difference between NameNode and DataNode in HDFS. **[5]**

B7. **[MapReduce]** Give a high-level explanation of data-flow in MapReduce (between major components). **[5]**

B8. **[Apache Spark]** Specify the three major steps in programming with RDDs. **[5]**

Part C. Data Analysis Methods and Algorithms (60 marks)

C9. **[Data Exploration and Normalization]** Suppose we have the following values for salary (in thousands of dollars), shown in an increasing order: 30, 37, 46, 51, 52, 52, 56, 60, 64, 70, 72, 73, 78. **[10]**

1. Discuss the skewness of the distribution. **5**

2. Do Z-Score Normalization on value 73. **5**

C10. [Similarity analysis] Suppose we have two documents X and Y (represented using vector). $X = (3, 0, 5, 0, 2, 6, 0, 2, 0, 2)$ and $Y = (0, 7, 0, 2, 1, 0, 0, 3, 0, 0)$. Compute the cosine similarity between the two documents (vectors) and discuss how similar they are. **[10]**

C11. [Cluster] Consider a sample data that consists of 6 two-dimensional points shown in Figure 1. The x and y coordinates of the points and the Euclidean distances between them are shown in Tables 1 and 2 respectively. Apply **single link** method to the given dataset and draw results using dendrogram and nested clusters, comment on the results. **[20]**

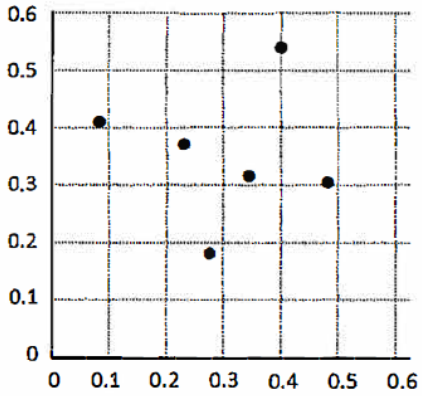

Figure 1. Sample data set with 6 two-dimensional points.

Table 1. x, y coordinates of 6 points

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 2. Euclidean distance matrix for 6 points

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

C12. [Prediction] Suppose we have a group of 40 students spending between 0 and 10 hours studying for an exam. The table 3 shows the number of hours each student spent studying, and whether they passed (1) or failed (0). **[20]**

Table 3. The number of hours each student spent studying, and the results.

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50
Pass	0	0	0	0	0	0	1	0	1	0
Hours	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.25
Pass	1	0	1	0	1	0	1	1	0	1
Hours	5.50	5.75	6.00	6.25	6.50	6.75	6.75	7.00	7.25	7.50
Pass	0	1	0	0	1	0	1	0	1	0
Hours	7.75	8.00	8.25	8.50	9.00	9.25	9.50	9.75	10.0	10.0
Pass	1	0	1	0	1	1	1	1	1	1

I feed these data into statistical software and get the output shown in Figure 2 (appendix).

Build a prediction model to show how the number of hours spent studying affects the probability that the student will pass the exam (present your model by equation or graph). For a student who studies 4 and 6 hours, what are the probabilities of passing the exam? Discuss your result.

Appendix: (see next page)

```
Descriptives...

18 cases have Y=0; 22 cases have Y=1.

 Variable     Avg       SD
     1        5.1250    3.0187
Overall Model Fit...
  Chi Square=    4.9685;  df=1;  p=    0.0258

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...
 Variable     Coeff.    StdErr       p          O.R.      Low  --  High
     1        0.2495    0.1189    0.0358       1.2834    1.0167    1.6202
Intercept    -1.0485    0.6725    0.1190

Predicted Probability of Outcome, with 95% Confidence Limits...
    X                Y     Prob        Low  --  High
    0.5000           0    0.2842     0.1050     0.5733
    0.7500           0    0.2970     0.1159     0.5767
    1.0000           0    0.3102     0.1276     0.5803
    1.2500           0    0.3237     0.1403     0.5842
    1.5000           0    0.3375     0.1537     0.5883
    1.7500           0    0.3516     0.1681     0.5928
    1.7500           1    0.3516     0.1681     0.5928
    2.0000           0    0.3660     0.1832     0.5976
    2.2500           1    0.3806     0.1991     0.6029
    2.5000           0    0.3954     0.2157     0.6086
    2.7500           1    0.4104     0.2329     0.6148
    3.0000           0    0.4256     0.2504     0.6216
    3.2500           1    0.4409     0.2683     0.6290
    3.5000           0    0.4563     0.2863     0.6372
    4.0000           1    0.4874     0.3217     0.6559
    4.2500           1    0.5030     0.3389     0.6665
    4.5000           1    0.5186     0.3553     0.6779
    4.7500           1    0.5341     0.3710     0.6903
    5.0000           0    0.5496     0.3857     0.7034
    5.2500           1    0.5650     0.3994     0.7172
    5.5000           0    0.5803     0.4121     0.7316
    5.7500           1    0.5954     0.4237     0.7465
    6.0000           0    0.6103     0.4344     0.7616
    6.2500           0    0.6250     0.4440     0.7768
    6.5000           1    0.6395     0.4528     0.7919
    6.7500           1    0.6538     0.4607     0.8067
    7.0000           0    0.6678     0.4680     0.8212
    7.2500           1    0.6815     0.4745     0.8352
    7.5000           0    0.6949     0.4805     0.8486
    7.7500           1    0.7079     0.4860     0.8613
    8.0000           0    0.7206     0.4911     0.8733
    8.2500           1    0.7330     0.4958     0.8846
    8.5000           0    0.7451     0.5001     0.8952
    0.0000           1    0.2595     0.0857     0.5670
    9.2500           1    0.7790     0.5114     0.9223
    9.5000           1    0.7895     0.5147     0.9299
    9.7500           1    0.7997     0.5178     0.9369
   10.0000           1    0.8095     0.5207     0.9432
   10.0000           1    0.8095     0.5207     0.9432
   10.0000           1    0.8095     0.5207     0.9432
```

Figure 2. Statistical results output

THE END OF THE PAPER