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

Module Code	Examiner	Department	Tel
INT303	Jia WANG	Department of Intelligent Science	9047

1^{st} SEMESTER 21-22 FINAL EXAMINATION

Undergraduate - Year 4

Big Data Analytics

TIME ALLOWED: *2 hours*

INSTRUCTIONS TO CANDIDATES

1. This is a blended close-book exam and the duration is 2 hours.
2. Total marks available are 100. This accounts for 70% of the final mark.
3. Relevant and clear steps should be included in the answers.
4. Only English solutions are accepted. For online students, answers need to be handwritten and fully and clearly scanned or photographed for submission as one single PDF file via LEARNING MALL.
5. Online students should use the format “Module Code-Student ID.filetype” to name their files before submitting to Learning Mall. For example, “INT303-18181881.pdf”.

Question 1

1. [Big Data Conception] For the following questions give short answers (Note: Explain in 3–5 lines maximum, no rigorous proof is required.)

(a) What is data science? How does it relate to and differ from statistics? **(5 Marks)**

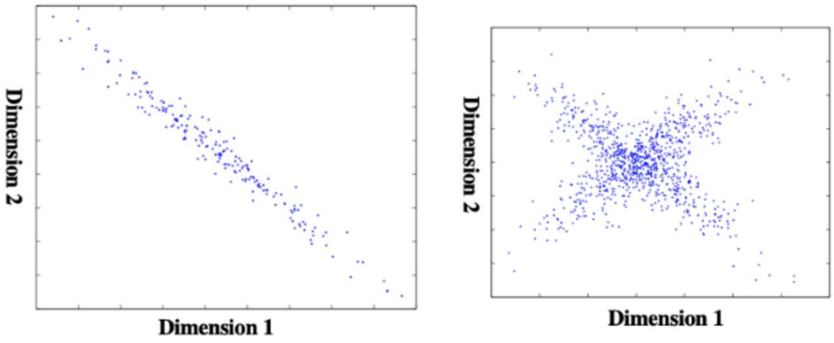
(b) Why do we need Hadoop for big data analytics? **(5 Marks)**

(c) Based on your experience, name the three best tools used for data analysis. **(5 Marks)**

(15 Marks)

Question 2

2. [Dimension Redundancy] Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D dataset plots, draw the first and second principal components on each plot.



(10 Marks)

Question 3

3. [Data Grammar] Following is a preview of the DataFrame **df**. The header of the DataFrame contains x,y,z. This DataFrame also includes some missing values in NaN.

x	y	z
1	NaN	1
2	NaN	2
NaN	1	3

(a) What is data cleaning? State the reasons why data cleansing is critical to the big data analysis process. **(6 Marks)**

(b) What are the expected outcomes of the following command? please fill in the blank.

(i) df.notna().sum(); x=____, y=____, z=____. **(3 Marks)**

(ii) df.isna().any(); x=____, y=____, z=____. **(3 Marks)**

(iii) df.notna().sum(); 0=____, 1=____, 2=____. **(3 Marks)**

(15 Marks)

Question 4

4. [Distance Measures] Calculate the following distance measures between the two vectors, $v1 = [0, 1, 1, 0, 0, 0, 1]$, and $v2 = [1, 0, 1, 0, 1, 0, 0]$.

(a) What is the Jaccard distance between two vectors? **(5 Marks)**

(b) What is the Cosine distance between two vectors? (You can use arccos(x) to present the answer). **(5 Marks)**

(10 Marks)

Question 5

5. [MapReduce] Determine whether the following statements are true or false. Please remember to justify your answers briefly.

(a) Each mapper/reducer must generate the same number of output key/value pairs as it receives on the input. **(3 Marks)**

(b) The output type of keys/values of mappers/reducers must be of the same type as their input. **(3 Marks)**

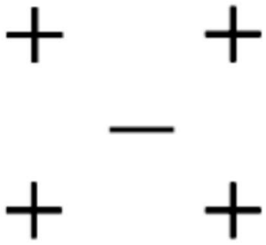
(c) The inputs to reducers are grouped by key. **(3 Marks)**

(d) It is possible to start reducers while some mappers are still running. **(3 Marks)**

(12 Marks)

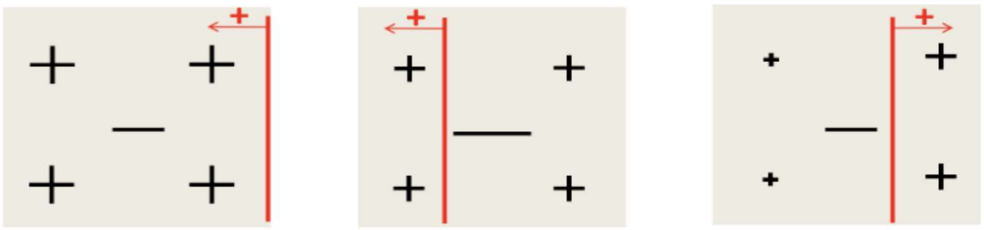
Question 6

6. [Boosting] Consider training a boosting classifier using decision stumps on the following dataset plot:



(a) Which examples will have their weights increased at the end of the first iteration? Explain the reasons. (4 Marks)

(b) How many iterations will it take to achieve zero training error? Explain the reasons. ***[Hint: the following figures may help.]*** (4 Marks)



Hint: These figures help solve the question Q6(b).

(c) Why do we want to use “weak” learners when boosting? **(2 Marks)**

(10 Marks)

Question 7

7. [Recommender Systems] Consider a dataset containing information about movies: genre, director and release decade. We also have information about which users have seen each movie. The rating for a user on a movie is either 0 or 1.

Here is a summary of the database:

Movie	**Release decade**	**Genre**	**Director**	**Total number of ratings**
A	1970s	Humor	D_1	40
B	2010s	Humor	D_1	500
C	2000s	Action	D_2	300
D	1990s	Action	D_2	25
E	2010s	Humor	D_3	1

Consider user U1 is interested in the time period 2000s, the director D2 and the genre Humor. We have some existing recommender system R that recommended the movie B to user U1. The recommender system R could be one or more of the following options:

- User-user collaborative filtering.
- Item-item collaborative filtering.
- Content-based recommender system.

(a) Given the above dataset, which one(s) do you think R could be? (If more than one option is possible, you need to state them all.) Explain your answer. **(6 Marks)**

(b) If some user U2 wants to watch a movie, under what conditions can our recommender system R recommend U2 a movie? If R recommends a movie, how to do it? If R cannot recommend a movie, please explain why it cannot be recommended. State any additional information R might want from U2 for predicting a movie for this user, if required. **(10 Marks)**

(c) Item-item collaborative filtering is seen to work better than user-user because users have multiple tastes. But this also means that users like to be recommended a variety of movies. Given the genre of each movie (there are 2 different genres in the dataset) and an item-item collaborative filtering recommender system that predicts k top-movies to a user (k can be an input to the recommender), suggest **at least three ways** to find top 5 movies to a user such that the recommender will try to incorporate movies from different genres as well. (Note: Explain in 3–5 lines maximum, no rigorous proof is required.) **(12 Marks)**

(28 Marks)

THE END OF EXAM