XI'AN JIAOTONG-LIVERPOOL UNIVERSITY

西交利物浦大学

REMOTE CLOSE BOOK

EXAMANSWER SUBMISSION

COVER SHEET

Name	Zizhe (Surname)	Wang (Given Name)
Student ID Number	1929413	
Programme	information and computing science	
Module Title	Big Data Analytics	
Module Code	INT 303	
Module Examiner	Jia Wang/Pengfei Fan	

By uploading or submitting the answers of this Remote Close Book Exam, I certify the following:

- I will act fairly to my classmates and teachers by completing all of my academic work with integrity. This means that I will respect the standards and instructions set by the Module Leader and the University, be responsible for the consequences of my choices, honestly represent my knowledge and abilities, and be a community member that others can trust to do the right thing even when no one is watching. I will always put learning before grades, and integrity before performance.
- I have read and understood the definitions of collusion, copying, plagiarism, and dishonest use of data as outlined in the Academic Integrity Policy, and cheating behaviors in the Regulations for the Conduct of Examinations of Xi'an Jiaotong-Liverpool University.
- This work is produced all on my own and can effectively represent my own knowledge and abilities.

I understand collusion, plagiarism, dishonest use of data, submission of procured work, submission of work produced and/or contributed by others are serious academic misconducts. By uploading or submitting the answers with this statement, I acknowledge that I will be subject to disciplinary action if I am found to have committed such acts.

Signature 王孜哲 Zizhe Wang Date 2023/1/5

Q1

(a)

<1> missing data: we have two solution: deletion or imputation.

~~When the data without some value~~

When the number of data without some value is overcome the number of data with full value, we should use deletion to process data.

~~Whe~~ In other situations, we consider ~~three~~ ways for ~~imputation.~~ ~~missing~~ imputation.

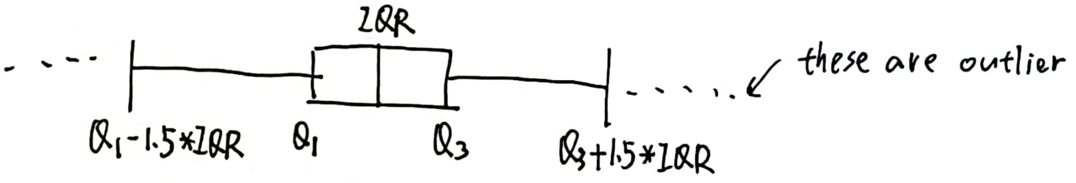
~~missing~~ ① regression imputation

② simple imputation

③ KNN imputation

④ piority Knoledge of this area

<2> ~~data st~~ Outlier valnes: we use IQR:



we draw box chart to delete them

<3> data standardization ~~isson~~ issues

~~we use min-Max scaler~~, ~~we~~

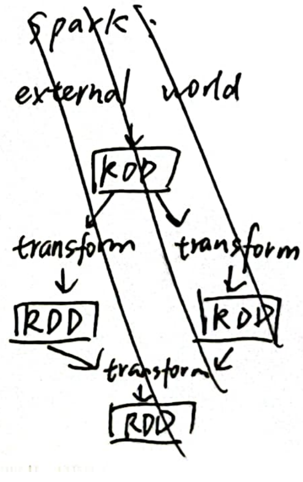
$$S = \frac{\sqrt{\sum_{i=0}^{n}(x_i-\bar{x})^2.}}{n-1}$$

(b) Spark is higher level of Hadoop, Hadoop is constructed by Mapreduce and HDFS.

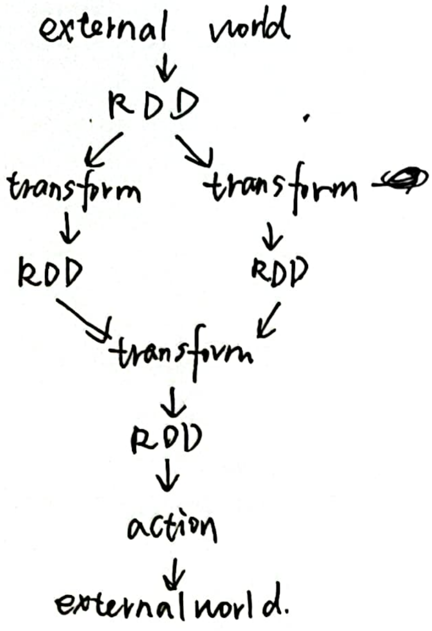
① Mapreduce have three main component:

map, join value by key, redeuce

②. Hadoop have chunk node, master node and client server ~~while~~.



③. spark :

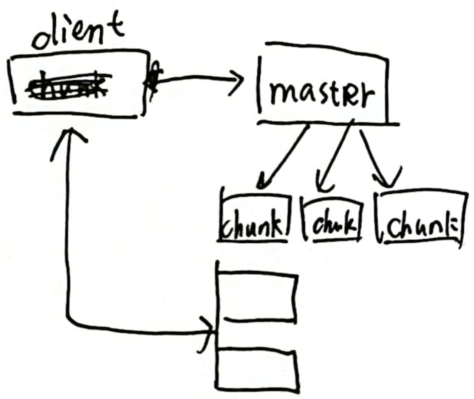


spark use memory , but Hadoop use disk.

transform : join , distinct , union and so on.

action : count , save and so on.

Hardoop infrastructure :



client ask master node and find chunk , then chunk return data to client .

(c). ~~skl~~ sklearn , numpy , pandas ,

Q2. a. 4 2 1

(b).

$$N = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2\sqrt{2} & \sqrt{2} \\ 2\sqrt{2} & -\sqrt{2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

N singular value decomposition :

$$N = U\Sigma V^T = \begin{bmatrix} 2\sqrt{2} & \sqrt{2} \\ 2\sqrt{2} & -\sqrt{2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}$$

(c)

reconstruction error ⇒

~~$= \sqrt{(3-\frac{1}{\sqrt{2}})^2 + (1-\frac{1}{\sqrt{2}})^2 + (1-\frac{1}{\sqrt{2}})^2 + (1\cdot}$~~

$$= \sqrt{(3-3)^2+(1-1)^2+0^2+(1-1)^2+(3-3)^2+(0-0)^2+(0-0)^2+(0-0)^2+(0-1)^2}$$

$$= 1$$

Q3.

(a) Student A is correct. Because it is obrious that if we split data into training and test sets, the min and max value should be different in these two sets. if we not do the normalize togther, the scale standard will be different on each sets. it will result in some error.

(b)

(i). $A = 3 \quad B = 4 \quad C = 4$

(ii) $A = a \quad B = 12.0 \quad C = 12.0$

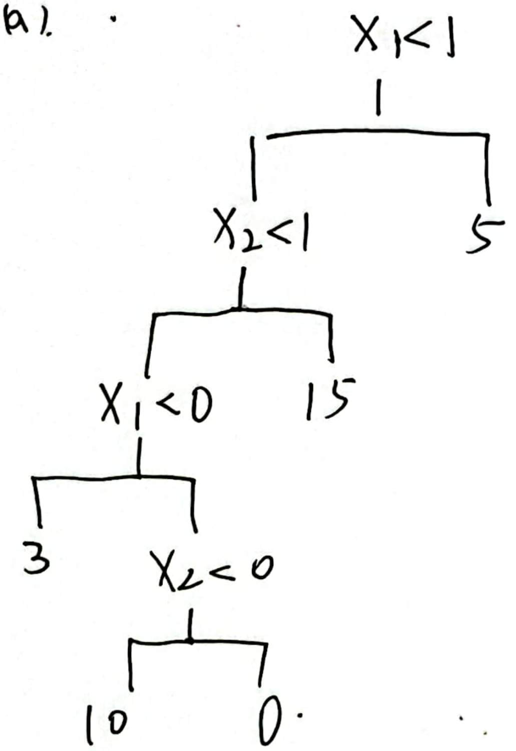
(iii) $a = 13.0 \quad b = 12.3$

(iiii) ~~histogram~~ pie ~~chart~~ chart

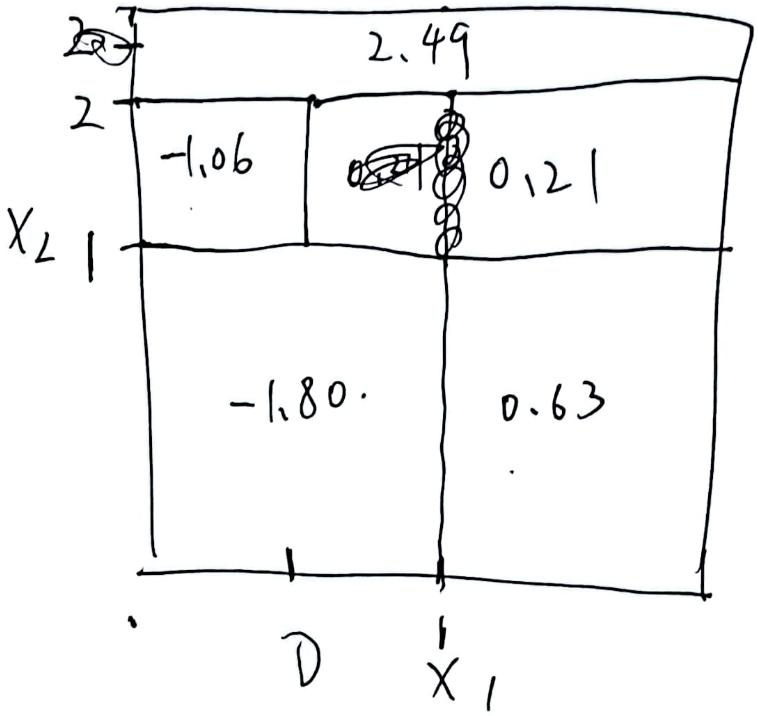
13

Q4

(a).



(b).



Q5

(a). False. We must reduce after all the mappers finished, or we will get wrong number of key-value pairs.

(b) False. Sort and Group should before the reducer.

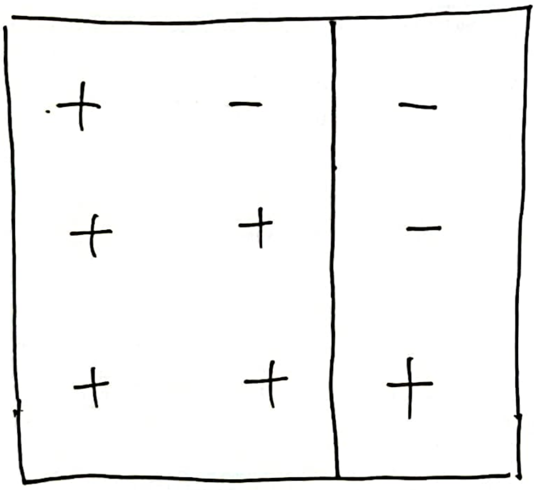
(c) False. ~~it doesn't matter~~ ;it doesn't matter.

(d) False. It up to the request of reducer.

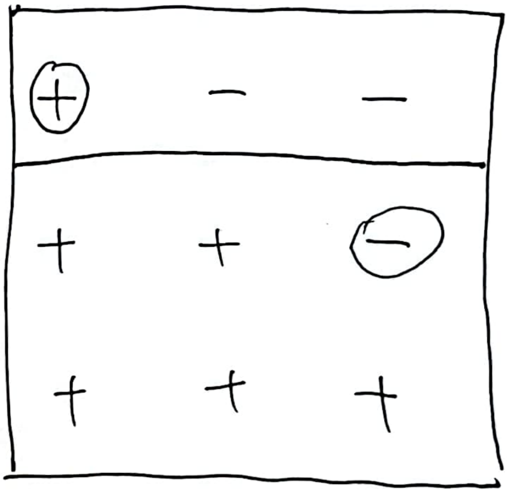
Q6.

(a).

i).



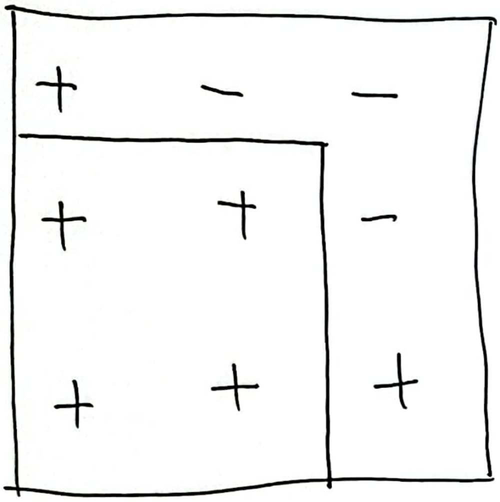
ii)



the wrong classification point's weight will increase.

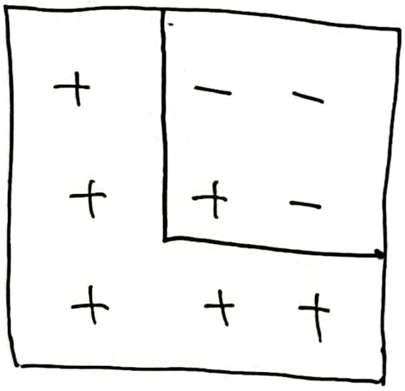
In the first iteration, the "−" in (1, 2) position has increased its weight, so in 2 iteration, it must be classified in right way, the same to "+" in (3,3) position

(iii)

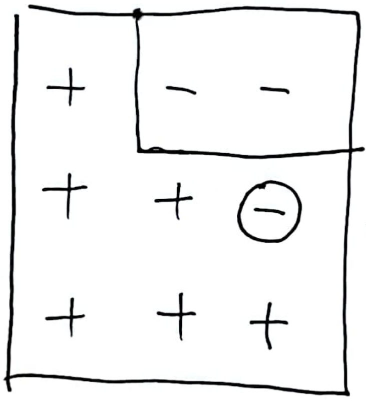


(b)

(i)



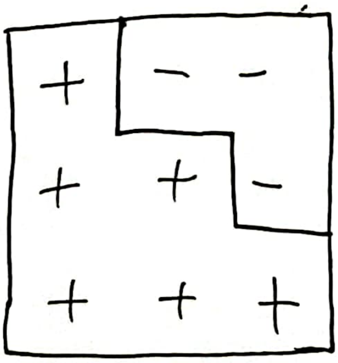
(ii)



the wrong classification point's weight will increase.

In the first iteration, the "+" in (2,2) position has increased its weight, so in 2 iteration, it must be classified in right way.

iii)



Consider the weight of position (2,2) and position (2,3) has increased, and "–" in (2,3) has increased more, so this "–" must be classified in right way.

Q7.

(a) Reader 1 : mean = 1.5 ~~similarity~~

Reader 2 : mean = $\frac{3+1}{2} = 2$ similarity (1,2) = $\frac{-0.5}{\sqrt{1+1}\cdot\sqrt{(\frac{1}{2})^2+(\frac{1}{2})^2}} = -0.5 = -\sqrt{\frac{\cancel{5}}{10}}$

Reader 3 : mean = 1 ~~similarity (1,3)~~ = ~~$\sqrt{1}$~~ similarity = 0

Reader 4 : mean = $\frac{2+1}{2} = 1.5$ similarity (1,4) = $\frac{0.5\times-0.5}{\sqrt{0.5^2+(-0.5)^2}\cdot\sqrt{0.5^2+(-0.5)^2}} = \frac{-\frac{1}{4}}{\frac{1}{4}}$

Reader 5 : mean = $\frac{0+3}{2} = 1.5$ similarity (1,5) = $\frac{-0.5\times1.5}{\sqrt{(0.5)^2+(0.5)^2}\sqrt{(-1.5)^2+(1.5)^2}}$ $= -1$

$$= \frac{-\frac{3}{4}}{\frac{\sqrt{10}}{4}} = -\frac{3}{\sqrt{10}} = -\sqrt{\frac{9}{10}}$$

~~sim~~ sim(1,3) > sim(1,2) > sim(1,5) > sim(1,4)

most similar : reader 3 and reader 2 .

$$\therefore \quad \frac{1.0\times0+3.0\times(-0.5)}{0+(-0.5)} = \frac{-1.5}{-0.5} = 3 .$$

∴ Book1 for reade 1 is rating 3.0.

(b),

	r1	r2	r3	r4	r5
book 1	?	3.0	1.0	2.0	0.0
book 2	2.0	1.0		1.0	
book 3	1.0				3.0

$$\begin{array}{c|ccccc} & r_1 & r_2 & r_3 & r_4 & r_5 \\ \hline & & 1.5 & -0.5 & 0.5 & -1.5 \\ & \frac{8}{3} & -\frac{1}{3} & & -\frac{1}{3} & \\ & -1 & & & & 1 \end{array}$$

∴ book 1 mean = $\frac{3+1+2+0}{4} = \frac{6}{4} = 1.5$ ~~sim~~

book 2 mean = $\frac{2+1+1}{3} = \frac{4}{3}$

book 3 mean = $\frac{1+3}{2} = 2$

sim (1,2) = $\frac{1.5\times(-\frac{1}{3})+(0.5)\times(-\frac{1}{3})}{\sqrt{\frac{1}{9}+\frac{1}{9}+\frac{64}{9}}\cdot\sqrt{(\frac{3}{2})^2\cdot(\frac{1}{2})^2\cdot(\frac{1}{4})\cdot(\frac{3}{2})^2}}$

~~sim (1,3) =~~

sim(1,3) = $\frac{-1.5\times1}{\sqrt{1^2+1^2}\sqrt{(1.5)^2+(0.5)^2+(0.5)^2+(-1.5)^2}} = \frac{-1.5}{2} = -0.75$ = 0 sim(1,2) > sim(1,3)

∴ rating D.0