বাংলাদেশ ইউনিভার্সিটি অব প্রফেশনালস্

সেকশন/গ্রুপ B (Section-B)

ইনভিজিলেটরের স্বাক্ষর

মোট পৃষ্ঠা সংখ্যা 11 টি

BSc. in CSE-17 Final Exam Fall-2020Dec. পরীক্ষা(Examination), 20 20

বিষয় (Subj): Data Warehousing & Data Mining পত্র/কোর্স নং (Paper/Course No): CSE-453

পত্র/কোর্সের নাম (Paper/Course Name): CSE-17 কেন্দ্র (Center): MIST

রেজিঃ নম্বর (Regn No): 131401170018 শিক্ষাবর্ষ (Session): 2019-2020

রোল নম্বর (Roll No): 201714018 তারিখ (Date): 06-12-2020

INSTRUCTIONS FOR EXAMINEE

1. Examinees are forbidden to write their names either on outer cover page or anywhere of the answer scripts. In case of violation, the answer script will not be evaluated.
2. Examinees must mention their roll and registration number along with session on the outer cover page of the answer scripts clearly. Otherwise, answer scripts may not be evaluated.
3. Students will write his examination roll number on the top left corner and section-A/B on the top right corner of each page. All pages must be numbered chronologically at the bottom center in x of y format. (for example: 1 of 21)
4. All rough works should be done in the same paper used as answer scripts. Answer scripts should be submitted intact. Papers used for rough work should be pen through by the examinees.
5. In no case, an examinee will be allowed to start the examination half an hour after the commencement of examination.
6. Examinees must abide by the instructions of chief invigilator if there are no definite instructions on any subject/matter.
7. No examinee will be allowed to leave the examination session until an hour has elapsed from the commencement of examination.
8. Legal action will be taken against the examinees those are caught for copying and found guilty for any breach of discipline as per rule.

পরীক্ষক কর্তৃক পূরণীয়

প্রশ্ন নম্বর	প্রদত্ত নম্বর
১	
২	
৩	
৪	
৫	
৬	
৭	
৮	
৯	
১০	
১১	
১২	
১৩	
১৪	
মোট	

পরীক্ষকের স্বাক্ষর

নিরীক্ষকের স্বাক্ষর

Continued......

INSTRUCTIONS FOR EXAMINEE

9. Smoking is strictly prohibited during examination.

10. The Camera of the examinee MUST always be ON during the examination and answer script submission. If Camera is OFF then that online examination will be treated as CANCELLED.

11. The answer scripts submitted beyond specified time will be treated as CANCELLED.

12. The examinee has to share his/her computer screen to the invigilator throughout the examination time.

13. The focus of the camera should be such that the invigilator(s) can see the script and examinee with his/her surroundings.

14. The examinee will send his/her scanned examination script in PDF format to the following e-mail addresses:

(a) e-mail address of subject invigilator/examiner.

(b) Central Database Scheme (coursecode@mist.ac.bd)
Example: EECE433@mist.ac.bd

15. The examinee has to preserve the original answer script of every examination and be ready to submit whenever asked for.

16. Answer script should be the A4 size papers with a cover page provided by Department. Examinee has to fill up his/her necessary details on the cover page. Section A and section B must be clearly marked on the cover page like. **Section A** or **Section B**

17. Examination duration for each subject will be two hours (section-A for one hour + section B for One hour). In between students will get 20 minutes time to submit the answer script of section A and 10 minutes time to issue the question for section B . After completion of 01 hour examination time for section B, students will get 20 minutes to submit the answer script of section B.

18. After completion of written examination (online/physical), viva will be conducted by the respective faculty of that subject.

Section-B

Ans. to the ques. no.-05(a)

Drawing the graph based on the given webpages as actors (nodes) and relationships (links or hyperlinks):

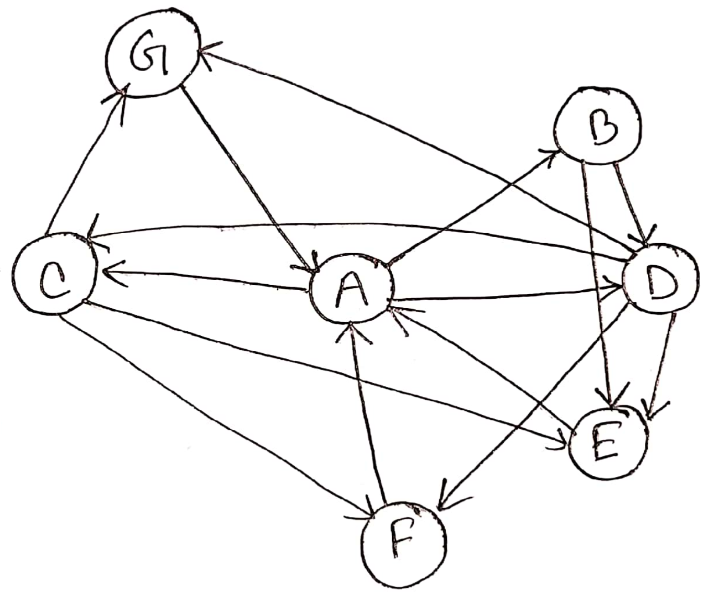


fig: Web pages as Actors.

If designed search Engine uses degree centrality as measuring rank of the web pages, only outlinks will be counted for each page. Since,

$$\text{Degree Centrality, } C_D = \frac{\Sigma d_o}{n-1}$$

where, $n = $ ~~6~~7 (Total ~~6~~7 web pages)

Here, '→' as Link (hyperlink)

Now, For A.com :

$$CD_A = \frac{3}{7-1} = 0.5$$

(∵A outlinks to 3 webpages)

(∵A → B, C, D)

For B. com:

$$CD_B = \frac{2}{7-1} = 0.33$$

(∵B → D, E)

For C. Com :

$$CD_C = \frac{3}{7-1} = 0.5$$

(∵C → E, F, G)

For D. com :

$$CD_D = \frac{4}{7-1} = 0.67$$

(∵ D → C, E, F, G)

For E. Com :

$$CD_E = \frac{1}{7-1} = 0.17$$

(∵ E → A)

For F. Com :

$$CD_F = \frac{1}{7-1} = 0.17$$

(∵ F → A)

For G. Con! $CD_G = \frac{1}{7-1} = 0.17$ (∵ G → A)

So, sorting the degree centrality values we will rank pages.

rank	Webpage	C_D
1	D.com	0.67
2	(A.com) & (C.com)	0.5
3	B.com	0.33
4	(E.com), (F.com), (G.com)	0.17

So, rank will be ~~pa~~ given high priority to D.com (as more outlinks) then,

A.com & C.com as equally then,

B.com then,

E.com, F.com, G.com eqully and as least 3 ranked webpages.

Ans. to the ques. no.-05(b)

Various steps involved in a classification process are discussed below:

① Model construction:

In this process dataset is split into Training and Test dataset and then fit the classification model with the training dataset to train the model to classify. In this step the classification model is trained carefully to classify categories for even unseen (unknown) data.

② Model usage:

In this step, classification model is tested against the test dataset to know the accuracy of the model. Test dataset is new to model so it is used to test the limits of the model. Then the classification model is used for classification tasks.

Major approaches for carrying out classification are:

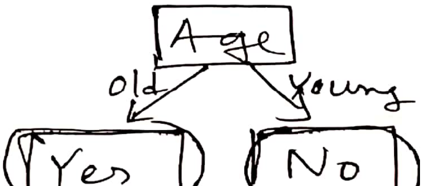
① Probability: here feature set consists of one Attribute. Example:

P(cancer | P34 = "H")

② Naïve Bayes: here, assumption of ~~intere~~ Attributes are not related is considered and classification with multiple feature set is done using following formula (with Laplacian Coefficient) Adding 1

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

③ Decision tree: Decision trees can also be used to classify from the tree leaf nodes where, nodes represent Attributs and brances represent values of Attributes.



④ Rule-Based: IF.. Then -- Rules to classify.
IF age is "Old" Then give "home = "Yes" :

Ans. to the ques. no.-05(c)

When agglometrative method of clustering is used the distance between two clusters can be measured either with ① single link or, ② Average distance

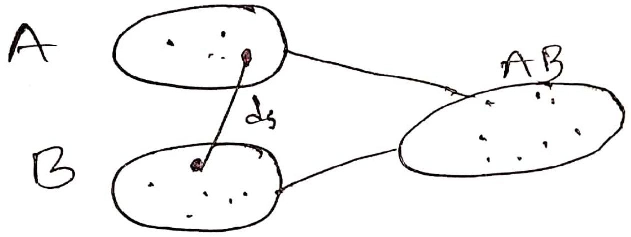


fig: single Link

here the above, we can see A cluster and B cluster is merged to form AB cluster. Here, single link is used. where, ds is the smallest distance that for Both A and B cluster points, is used to merge.

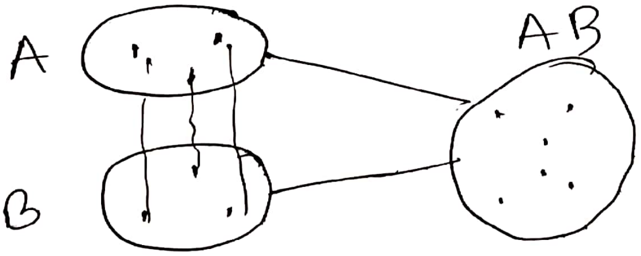
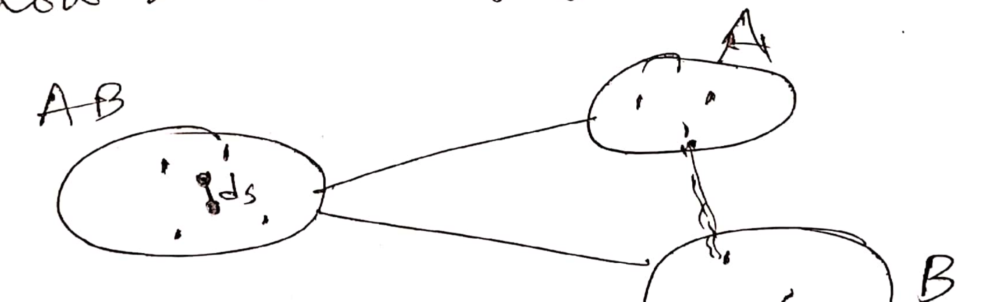


fig Avg. distance

Here, the Avg. distance for both A cluster and B cluster is used to merge to AB cluster.

When using divisive method for clustering the distance between two cluster using ① single Link ② Avg. distance is similar to agglomerative approach but instead of merging, spliting is used here.
below are the figures:



since ds is smaller,

fig: Single Link

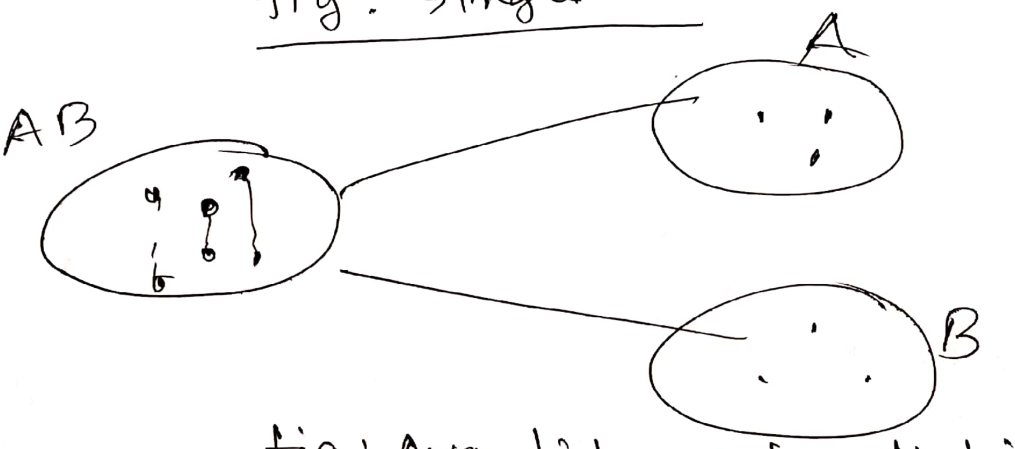


fig: Avg distance for divisive

Ans. to the ques. no.- 07(a)

The data with two attributes:

	Play Chess	Don't Play chess	Total
Like Science Fiction	250(180)	200(270)	450
Don't Like Sci-fi	50(120)	250(180)	300
Total	300	450	750

We can find χ^2 (chi-square) to get the correlation between two attributes.

First we calculate all the expected values using the formula; (inside brackets)

$$e_{ij} = \frac{count(A=a_i) \times Count(B=b_i)}{n}$$

So,

~~$\chi^2 =$~~

$$\chi^2 = \frac{(250-180)^2}{180} + \frac{(200-270)^2}{270} + \frac{(50-120)^2}{120} + \frac{(250-180)^2}{180}$$

$= 113.426$

the χ^2 (chi-square) value is not that much high. So, there exists a very mild or weak co-relation between the two attributes (Play chess, Like sci-fi). So, I will not drop any of the attributes ~~[illegible]~~ from the database.

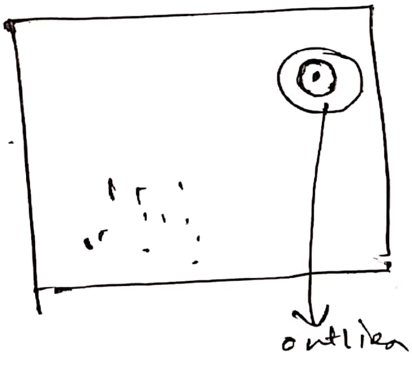
Ans. to the ques. no.-07(b)

Outliers are data objects that deviates significantly from other data objects as if they were generated by a different mechanism.

various kinds of outliers are:

① Global outliers:

~~[illegible]~~ from all other points. outlier is outside.

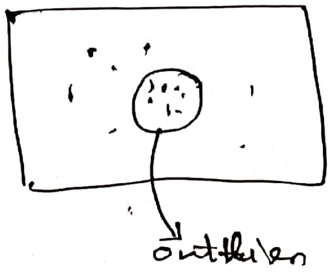


② Contextual outliers!

only outlier when context is given

③ Collective outliers;

outlier is collectively deviate from other data points.



P.T.O

outliers are not errors. where a noise is an error induced when data entry.

Ans. to the ques. no.-07 (c)

Disperstion of data can be measured with standarddeviation. with the formula:

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N}}$$

standard deviation represents the dispertion of data. Also, median and mode and boxplot, Histogram can also be used to measure the

dispertion of data.

Cluster can also help data to find deviation and noises.

Ans. to the ques. no-07 (d)

Noisy ~~data~~ data handle in Data mining. can be done several ways:

① Completly delete the data object.

② Replace the noise with Average value of the Attribute.

③ Replace the noise value with "Unknown" (new class) if category.

④ Use model to predict value for the noise.

and many othe ways to handle noise in Data Mining task.

—x—