PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE18CS312

NOVEMBER 2020: IN SEMESTER ASSESSMENT B Tech FIFTH SEMESTER
TEST – 2

UE18CS312 (4 credit subject) - Data Analytics
Scheme and Solutions

Time: 90 Minutes	Answer All Questions	Max Marks: 40

1. a) Suggest the most appropriate recommender system for each of the following scenarios and briefly explain the reason it is most appropriate. **4 (2+2)**

(i) Mrs. Patel has started a new home-kitchen catering business for those in her apartment complex and wants to know what items she should prepare/ advertise to get good business in the first few weeks.

(ii) One of the teams InternWhiz wishes to develop an app for the class project that will recommend potential companies with internship offers to students, based on the following (i) their specialization domain (ii) programming languages they know (iii) application areas they are keen on (iv) other skills (hobbies such as music, art, etc.).

2 marks each (1 mark for type of recommender system, 1 mark for the reason)

(i) Content based recommender system (collect the information of the type of cuisine families in the apartment complex like) – because ratings are not available in this 'cold start' scenario

(ii) Knowledge based recommender system (either search-based or navigation-based (conversational style or drop down menu) to arrive at the most appropriate matches from the backend) because students would typically know what they are like

1 mark for a constraint based system because students may know what they do not want or case based recommendation if students can share the sort of opportunities they like

Partial/ full credit may be awarded for any other options based on the strength of the rationale presented.

b) Use Manhattan distance to determine whether Kreacher is given a free meal at the inn based on (i) 1-NN and (ii) 3-NN from the following data: **3**

	Can cook	**Grumpy**	**Clean**	**Gluttonous**	**Free meal**
Dobby	1	0	0	1	1
Winky	1	0	1	1	0
Bogrod	1	0	1	0	1
Alguff	0	1	1	1	0
Kreacher	1	1	1	0	**?**

Solution: (1 mark for distance computation)
Distance(Dobby, Kreacher) = 3 (Doby = free meal)
Distance(Winky, Kreacher) = 2 (Winky = no free meal)
Distance(Bogrod, Kreacher) = 1 (Bogrod = free meal)
Distance(Alguff, Kreacher) = 2 (Alguff = no free meal)

1 mark: 1-NN: Yes, gets a free meal
1 mark: 3-NN: No free meal

c) Of a 1000 students who stay at a hostel, a total of 800 students prefer hot chocolate and 150 students prefer cold coffee. 50 of those who prefer cold coffee also enjoy hot chocolate. Compute the support and confidence for the rule given below: **3**

Does not drink hot chocolate → Does not drink cold coffee

Solution:

(1 mark) support count(Coffee and Cocoa) = 50, support count(only cold coffee) = 100,
support count(only cocoa) = 750, support count (neither cold coffee nor hot chocolate = 100

(1 mark) support (do not drink hot chocolate → do not drink cold coffee) = 100/1000 = 0.1

(1 mark) confidence(do not drink hot chocolate → do not drink cold coffee) = 100/200 = 0.5

Full credit can be given if support and confidence and computed correctly, even if the partial computation of support counts is not explicitly shown or shown in the form of a contingency matrix.

2. a) Given the following ratings (on a scale of 1-5, where 1 is the worst and 5 is the best) for various movies, answer the following questions: 4

(i) Which two individuals' ratings would you consider to predict Snoopy's rating for the movie Avengers based on collaborative filtering?

(ii) Compute the rating Snoopy is likely to give the Avengers using collaborative filtering.

	Harry Potter	**Lord of the Rings**	**Avengers**	**Black Panther**
Charlie Brown	3	4	5	NA
Snoopy	5	4	**?**	5
Pattie	2	NA	NA	5
Marcy	2	3	4	NA

(Hint: You may use the following proximity measure: cosine_similarity(**a**,**b**) = ($\mathbf{a}^T\mathbf{b}$)/(||**a**||.||**b**||))

Solution:

(i) (1 mark) Charlie Brown and Marcy since both of watched and rated two movies in common with Snoopy and have also watched and rated the Avengers.

(ii) (1 mark) similarity between Snoopy and Charlie Brown = 31/(5*6.4) = 0.96875

(**note:** here, we take into consideration only HP and LotR for computation of similarity since they are the common movies watched and rated by Charlie Brown and Snoopy)

(1 mark) similarity between Snoopy and Marcy = 22/(6.4*3.6056) = 0.9534

(1 mark) computing the average rating: (5*0.96875 + 4*0.9534)/(5+4) = 4.5 (can be rounded up to 5 and still get credit)

b) With a schematic diagram, briefly explain how the following points are labeled in an iteration of DBSCAN: 3

(i) Core point

(ii) Noise point

Solution:

(1 mark) – A point with minPts or more points within eps distance of it is called a core point

(1 mark) – A point that is within eps of a core point but does not have minPts or more points within eps radius from it is called a border point; a point that is neither core nor border is called a noise point

(1 mark) – schematic diagram of core point and noise point

c) The word 'delicious' appears three times in a food review of a total of 100 words. Assuming there are a 1000 reviews in all and 'delicious' is found in 100 of them, briefly explain the feature "TF-IDF" and compute its value for the word 'delicious' for this data. 3

Solution:

1 mark: Term frequency (TF) is the number of times a word appears in a document

1 mark: Inverse document frequency (IDF) is the inverse of the number of documents in which the word appears out of all the documents there are

TF-IDF is the product of the two.

1 mark: (3/100)*(1000/100) = 0.3

OR if $\log_{10}$ is used for IDF (3/100)*1 = 0.03

Any other variation of the formula can be given credit based on how reasonable it is.

3.	a)	Briefly explain any two challenges posed by sparse data and any two ways of dealing with (storing, processing, etc.) sparse data. Solution (1 mark each): challenges – (i) storage (most of the entries are zero's) (ii) processing (most of the entries are zero's, so most similarity measures like SMC may be overwhelmed by how many common entries there are, masking the real similarity) Two ways to deal with sparse data to overcome the challenges (i) Nonzero values can be stored with row major or column major indices (ii) Similarity measures like Jaccard can be used instead of SMC to give more importance to presence, rather than absence of values	4 (2+2)
	b)	In the scenarios given below, identify whether the omitted third variable is confounding or not (briefly justify your answer). (i) Variable 1: revenue from a course on deep learning, Variable 2: no. of promotion mails sent about the course, Variable 3: gender of the registrants (ii) Variable 1: revenue from a course on fitness and self-defense Variable 2: no. of promotion mails sent about the course, Variable 3: gender of the registrants Solution (open ended, based on the strength of the justification) 1.5 marks each – possible answer choices could be: Scenario (i) not a confounding variable – both male and female students would be interested in learning deep learning (alternatively, is a confounding variable because there is a higher concentration of male students in STEM courses and they are more likely to sign up for this) Scenario (ii) is a confounding variable – female students are more fitness-aware or would sign up for a self-defense course whereas male students are either not into fitness or would rather sign up for a weight training or training for a specific sport (alternatively, is not a confounding variable since both genders are aware of fitness and would want to learn self-defense)	3
	c)	If the market share for the current month between StarSports and ESPN is $\mathbf{u}_0$=[0.4 0.6], what would the market share be **after two months** if it is known that 30% of those who watch StarSports will switch to ESPN every month and 20% of those who watch ESPN will switch to StartSports every month. Write the transition probability matrix **P** clearly. [Hint: you can use use $\mathbf{u}_2 = \mathbf{u}_1\mathbf{P}$ = (($\mathbf{u}_0\mathbf{P}$)**P**) to compute the state of **u** after 2 months.] Solution (1 mark): P = 0.7 0.3 0.2 0.8 After one month: $\mathbf{u}_1$ = **uP =** 0.4*0.7 + 0.6*0.2 = 0.3*0.4+0.8*0.6 = 0.28+0.12 0.12+0.48 = [0.4 0.6] Therefore, $\mathbf{u}_2$ = $\mathbf{u}_1\mathbf{P}$ **= [**0.4 0.6] (2 marks)	3
4.	a)	TagE is a food manufacturining company that plans on increasing its price for the most popular noodles by Rs. 10 and increasing the net weight from 70g to 100g. As a consultant Data Analyst suggest: (i) any two performance indicators that can be used to measure the impact of this change (ii) how would you design an A/B test to help TagE get some feedback on this proposal before they implement the change in all their manufacturing plants? Solution (2 marks each) open ended; possible answers could be: (i) 2 performance indicators: total #packets sold, total revenue from sales (compared with previous number of packets sold and previous revenue from sales) (ii) Identify different regions based on sale figures; in every local region, for about half the locations introduce the change, the other half serves as a control group. Use the two groups to track the difference for the same period of time with all other conditions being the same	4 (2+2)

b) Given the state transition probability matrix P below, identify the absorbing states and write P in the Canonical form. Also compute the Fundamental matrix **F** of **P**.

(Hint: **F** = (**I-Q**)$^{-1,}$ where I is the identity matrix and **Q** is the matrix of probabilities of transitions between non-absorbing states)

3

P =

	A	B	C	D
A	0	0.3	0.3	0.4
B	0	1	0	0
C	0	0	1	0
D	0.8	0.1	0.1	0

Solution: (1 mark) B and C are absorbing states

(1 mark) The canonical form is

	B	C	A	D
B	1	0	0	0
C	0	1	0	0
A	0.3	0.3	0	0.4
D	0.1	0.1	0.8	0

(1 mark) F = inverse(**I-Q**) = inverse of the following matrix

1	-0.4
-0.8	1

=

1.47	0.59
1.18	1.47

(Full credit can be given even if the solution stops at the penultimate step.)

c) In the context of evaluating recommender systems, briefly explain what each of the following mean:

3

(i) Coverage
(ii) Novelty
(iii) Diversity

Solution (1 mark each)

(i) Novelty – how likely is a user to get a recommendation they did not expect

(ii) Coverage – (any one) the fraction of users for whom at least k ratings can be predicted or the fraction of items for which the ratings of at least k users can be predicted or the fraction of items that are recommended to at least one user

(iii) Diversity – if k predictions are made, they must not all be of the same type (i.e., if 3 movies are recommended, then at least one must be from a different genre)