Student name: Nguyễn Minh Tâm Student ID: 1952968

Family name, Given name

Department of Computer Science and Engineering

Intelligent Systems

Midterm Exam - Trimester 1, 2021

Writing time

1 hour

Reading time

10 minutes

Question 1

A company has collected quarterly sales for the past two years which are shown in the following table. The company wants to forecast the next year's seasonal sales.

Index	Time	Sales ($)	Index	Time	Sales ($)
1	Spring 2017	4836	5	Spring 2018	5412
2	Summer 2017	5890	6	Summer 2018	6138
3	Fall 2017	6510	7	Fall 2018	6666
4	Winter 2017	7564	8	Winter 2018	8184

a) Let A_1 and A_2 be the actual total sales (i.e., the sum of all four seasonal sales) in 2017 and 2018, respectively. Assume current time is t, the n-moving average (MV) technique makes forecast for the time $t+1$ by taking the average of previous n actual values where $n \leq t$. (The formula can be written as $F_{t+1} = \frac{\sum_{i=t-n+1}^{t} A_i}{n}$ where A_i is the i-th actual data). Predict the total sales for 2019 using MV with 2 actual values A_1 and A_2.

A1 = 4836 + 5890 + 6510 + 7564 = 24800
A2 = 5412 + 6138 + 6666 + 8184 = 26400
Predict total sales for 2019 = (A1 + A2)/2 = 25600

b) In general, we would expect the total sales gets increased in both 2018 and 2019 if the economy situation has been keeping going well since 2017. Use this and your answer to sub-question **a)** to explain the limitation of moving average (MV) method in forecasting.

Total sales for 2017 = 24800
Total sales for 2018 = 26400
Total sales for 2019 = 25600

In general, we expect the total sales for 2019 will be increased, more than 2018. However, moving average method cannot catch up the trend well because moving average method takes the average of total sales in the past, so the result cannot be bigger than the past total sales.

c) Calculate the average seasonal sales for both 2017 and 2018.

The average seasonal sales for 2017 = 24800 / 4 = 6200
The average seasonal sales for 2018 = 26400 / 4 = 6600

d) Here are the steps of forecasting with seasonality. Please follow the steps to fill out the blank cells in the following form.

1. Calculate the average seasonal sales for each year;
2. Calculate each seasonal index (by dividing the actual seasonal sales by the average seasonal sales);
3. Compute the average indexes;
4. Predict the average seasonal sales for the next year (i.e., 2019);

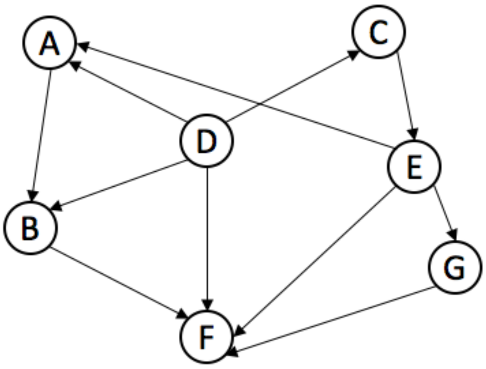
5. Multiple next year’s average seasonal sales by each average seasonal index.

(**Note**, you have already done the step 1) by answering sub-question c) and please put those values into the table. Also, step 4 is already done for you (the red value). You can copy the form to your answer sheet then fill blank cells).

Quarter	2017	Seasonal Index	2018	Seasonal Index	Average Index	2019
Spring	4836	0.78	5412	0.82	0.8	5520
Summer	5890	0.95	6138	0.93	0.94	6486
Fall	6510	1.05	6666	1.01	1.03	7107
Winter	7564	1.22	8184	1.24	1.23	8487
Average	6200		6600			6900

Question 2

a) Degree centrality is the most basic centrality metric. Find the degree centrality of each vertex in the graph and rank the vertices using their degree centrality. Fill out the following table. (**Note**, you can copy the form to your answer sheet then fill blank cells.)



Vertex	Centrality	Ranking
A	3	2
B	3	2
C	2	3
D	4	1
E	4	1
F	4	1
G	2	3

Question 3

Schema reuse is a new trend in creating database schemas by allowing users to copy and adapt existing ones. The motivation behind schema reuse is the slight differences between schemas in the same domain; thus making schema design more efficient. Reusing existing schemas supports reducing not only the effort of creating a new schema but also the heterogeneity between schemas.

Finding related schemas is one of the core problems of schema reuse. You work as a data engineer at Oracle. Oracle has a large repository of schemas. Each database schema has a set of attributes. Some attributes are common among schemas, while others are not. Your task is to support database designers to create new schemas via the schema reuse paradigm. For instance, when a database designer wants to create a new schema, he wants to query the schema repository for references:

- He can start with a few attributes and query the schema repository for hints to finish his design.
- Alternatively, he can complete a schema and query the schema repository to check his design.

Example: we have a repository of schemas:

- *S1: {a1, a3, a7}*
- *S2: {a1, a4, a8}*
- *S3: {a2, a6, a9}*
- *S4: {a1, a5, a10}*

Given a query Q = {a1, a2}, we should rank these schemas as S3 > S1=S2=S4. S3 has the highest rank since attribute a1 occurs frequently in many schemas and thus has less discriminatory power (i.e. the more schemas contain an attribute, less information it provides).

Design an algorithm to find related schemas ranked by their similarity to the query.

a) How do you model the problem (input, output, etc.)? Justify your model.

Using tf-idf weights to normalize each attribute

b) What steps should be involved? Provide a quantitative measure for each step if needed. Justify the design choice for each step.

Step 1: Calculate the term frequency (attributes) in each schemas

Step 2: Calculate the document frequency for each attribute

Step 3: Calculate the inverse document frequency

Step 4: Calculate the tf-idf

c) Apply your approach to the above example and calculate the quantitative results.

Step 1:

Term frequency (tf)

Terms	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
S1	1	0	1	0	0	0	1	0	0	0
S2	1	0	0	1	0	0	0	1	0	0
S3	0	1	0	0	0	1	0	0	1	0
S4	1	0	0	0	1	0	0	0	0	1

Step 2:

Document frequency (df)

Terms	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
df	3	1	1	1	1	1	1	1	1	1

Step 3:

Inverse document frequency (idf)

Terms	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
idf	0.42	2	2	2	2	2	2	2	2	2

Step 4:

Tf-idf

Terms	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
S1	0.42	0	2	0	0	0	2	0	0	0
S2	0.42	0	0	2	0	0	0	2	0	0
S3	0	2	0	0	0	2	0	0	2	0
S4	0.42	0	0	0	2	0	0	0	0	2

Question 4

You work as a software engineer for Netflix, a popular movie streaming service across the globe. The CEO suggests that Netflix provides users with movie recommendations. You are the lead engineer in this project.

1) You start by considering different approaches to build a recommender system. Netflix has many movie titles in their database. However, they only recently started recording user ratings and have very few ratings in their database. In case you need to implement the system before being able to collect a lot of data, which of the following would you consider to be a better recommender system?

a) User-based collaborative filtering
b) Item-based collaborative filtering
c) Content-based collaborative filtering

Give a brief justification of your answer.

I think content-based collaborative filtering is the most suitable in this case because Netflix only recently started recording user ratings and have a very few ratings in their database so we cannot use rating-based collaborative filtering (both user-based and item-based). We can only use content-based collaborative filtering to match the users' interest to the descriptions of the movie.

2) A few years later, Netflix has collected lots of data, and the situation in the previous part of the question no longer applies. You are working in the team that develops an algorithm to decide if two movie shows are similar to each other. To do this for two shows A and B, we compute the Pearson correlation coefficient between the ratings given to show A, and the ratings given to show B.

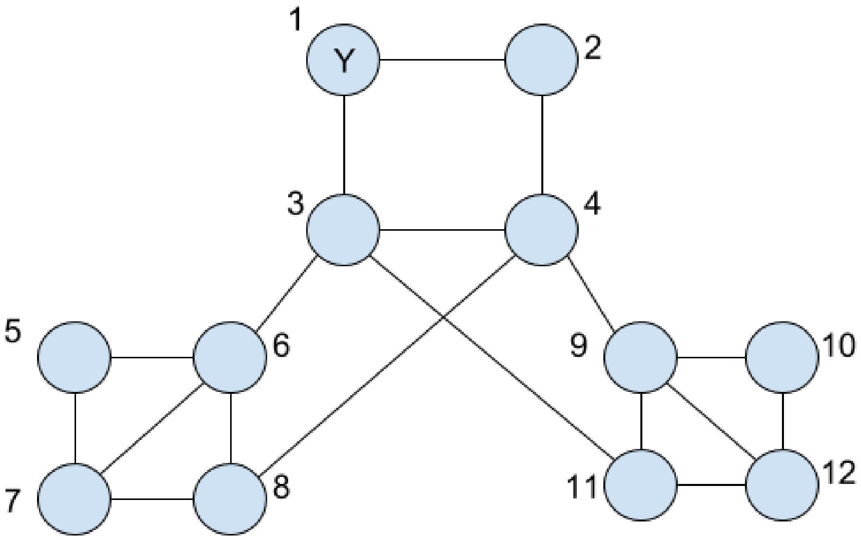
Your friend Thomas works at Netflix, and proposes to use a similar strategy to determine if users are similar to each other, based on whether other uses have followed them or not. Whenever a user u follows another user v, Thomas gives the (user, user) pair a "rating" of r(u,v)=1. Because there are (millions choose 2) pairs of users on Netflix, Thomas does not store any ratings for pairs of users who don't follow each other. Thomas again proposes to compute the Pearson correlation coefficient between two sets of ratings.

Is Thomas's approach a good strategy? Why or why not?

I think Thomas's approach is not a good strategy because to determine if users are similar to each other, we have to compare the ratings of each movies users watch. If we only consider users have followed another user, it will not be correct because there can be a situation like A follows B but both A and B do not have the same ratings of each movies and there is C who does not follow A but has similar ratings of each movie compared to A. Applying Thomas' strategy, C will not be considered because C does not follow A and this is wrong.

Question 5

The parliament has organized a voting scheme for a new bill this summer. You are a strategic advisor in charge of vote forecasting and voter acquisition tactics. You have the following social graph of voters, which is undirected.



The undecided voters will go through a 3-day decision period where they choose a candidate based on the majority of their friends. The decision period works as follows:

1. The graphs are initialized with every voter's initial state as the above figure. (yes (Y), no (N), or undecided)
2. In each day, every undecided voter decides on a vote 'yes' or 'no'. Voters are processed in an increasing order of node ID. For every undecided voter, if the majority of their friends (>=50%) vote 'yes', they now vote 'yes'. Otherwise, they vote 'no'.
3. When processing the updates, use the values from the current day. For example, when update the votes for node 2, you should use the updated votes for nodes 1 and 4 from the current day.
4. There are 3 days of the process described above.
5. On the 4th day, the votes are counted.

a) Perform iterations of the voting process. How many votes each option has?

	1	2	3	4	5	6	7	8	9	10	11	12
Day1	Y	Y	N	N	N	N	N	N	N	N	N	N
Day2	Y	Y	N	N	N	N	N	N	N	N	N	N
Day3	Y	Y	N	N	N	N	N	N	N	N	N	N

b) You have a public relation idea to increase the ‘yes’ voters by organizing a very classy $1000 per plate dinner event. Assume everyone that comes to your dinner is instantly persuaded to vote ‘yes’ regardless of his/her previous decision. This event will happen before the decision period.

Choose a minimum number of voters to invite for dinners such that all the voters in the graph vote ‘yes’. Justify your strategy and compute the voting result.

Ans:
Clearly you can see there are 3 communities: {1,2,3,4}, {5,6,7,8}, {9,10,11,12}.

I will choose from each community one node such as 3,8,12. If we choose like this, we can make all voters vote ‘yes’.

Because 3 & 8 vote ‘yes’, 6 will vote ‘yes’ and this make the community {5,6,7,8} will all vote ‘yes’ because 2 of 4 of this community have already voted ‘yes’.

With the community {1,2,3,4} is likely the same because 4 will vote ‘yes’ due to 3 & 8 and this makes this community all vote ‘yes’.

Similar to the last community, all with vote ‘yes’.

c) You have another idea to increase the ‘yes’ voters by spending $1000 to make any two voters in the network become friends.

Choose a minimum number of connections you want to create such that all the voters in the graph vote ‘yes’. Justify your strategy and compute the voting result.

Ans: I will make 8 & 11 to become friends so that all the voter will vote ‘yes’. Then will only need 3 & 8 to vote ‘yes’ already.

END OF EXAM