A.K

6+6+5+5+4 = 26

CS5600 Data Mining Exam - 1, Fall 2025

(Closed Book, 1.5 hrs, 35 marks)

Name Pandrangi Adilya Sriram ID EE22BTECH11039

Instructions:

- Return the question paper along with the answers written in the empty spaces provided.
- Use the provided extra sheets to do the rough work and not on the question paper.
- Be precise and concise in the answers. If you make any additional assumptions, write them clearly.
- The marks for each question is given in square brackets in bold font.

6

Q1 [6*1 = 6 marks] Mark true/false for the following questions. Correct answer: +1 & Wrong answer: -0.5

[TRUE/FALSE] In the PageRank algorithm, we fix the irreducible and aperiodic property of the graph just by adding a link from each page to every page. False (we add a probability that it can go to any page & not follow a link)

[TRUE/FALSE] Minhashing and LSH can only be used with text documents. False

[TRUE/FALSE] Let's define the candidate column pairs in LSH as those that hash to the same bucket for ≥ 2 bands. This will reduce the number of false positives. True

[TRUE/FALSE] The above change will also reduce the number of false negatives in LSH. ~~False~~ False

[TRUE/FALSE] If the support of itemset {A,B} is high, then the confidence of association rule A → B must be high. False

[TRUE/FALSE] In the multiple minimum support model, superset of a non-frequent itemset may be frequent. True

Q2 [5*2 = 10 marks] Pick the correct choice(s). If more than one option is correct, marks will be awarded only if, only and all the correct options have been marked. No negative marking.

2

1. For which of the following pattern mining algorithm, we always use F(k-1) (frequent itemsets of size k-1) to compute F(k) (frequent itemsets of size k):
   a. Apriori based class association rule (CAR) mining algorithm
   b. GSP algorithm for sequential patterns
   c. MS-Apriori algorithm
   d. Apriori algorithm

2

2. Consider two documents with Jaccard similarity of 0.8. Using LSH with b=20 bands and r=2 rows per band, what is the approximate probability that this pair will be identified as a candidate pair?
   a. 0.98 b. 0.81 c. Very close to 1 d. 0.19

3. Which of the following features you would use for a content based document recommendation system
   a. Top-K stop words with highest TF.IDF score
   b. Top-K non-stop words with highest TF.IDF score
   c. Top-K shingles with highest frequency
   d. The author of the document

d, b, d ✗

Consider the citation matrix L and its transpose shown below. Let rows 1, 2, 3, 4, 5 in L represent the papers A, B, C, D, and E respectively. Answer the following two questions based on this matrix:

$$L = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad L^T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

4. Which of the following pairs of papers have greater than or equal to 1 co-citation

{b}

a. (A, E) b. (A, D) c. (A, B) d. (E, C)

5. [A, D] Which of the following paper(s) have the least number of citations (the number of received citations)

2

a. A b. B c. C d. E

Q3 [5*2 = 10 marks] Very short answer/fill in the blanks.

[6marks] Assume a rating matrix of 100*1000 (i.e., 100 users and 1000 items). For each of the following how many free parameters needs to be determined using gradient descent:

2 Latent factorization with 10 latent factors and no regularization. 11,000

Latent factorization with 10 latent factors and regularization of only user-movie interactions. 111,000

Latent factorization with 10 latent factors and regularization of biases of users and items, and user-movie interactions.

122,000

[2 marks] Given two item sequences S_1 and S_2, consider an alternative approach to join the two sequences: remove one item from the last elements of both S_1 and S_2 to check whether the resulting sequences are identical. Using this join approach list the candidate(s) that would be generated by joining the sequence <{a}, {b}, {c}> and <{a}, {b}, {d}>

2

① <{a}, {b}, {c, d}>

② <{a}, {b}, {c}, {d}> ③ <{a}, {b}, {d}, {c}>

[2 marks] Consider four items 1, 2, 3 and 4 in a data set with minimum item support:

MIS(1) = 10%, MIS(2) = 20%, MIS(3) = 5%, MIS(4) = 6%

Assume our data set has 100 transactions. The first pass gives us the following support counts:

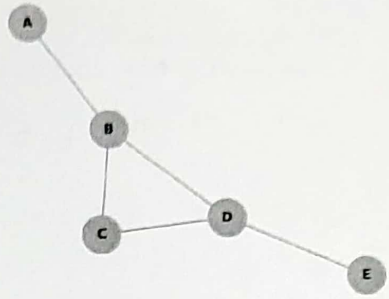
{3}.count= 3, {4}.count= 3, {1}.count= 9, {2}.count= 25

1

L = {1, 2} and F_1 = {2}

(5)

Q4 [5 marks] Consider the following unweighted and undirected graph. Compute the betweenness centrality of each vertex.



Vertex A :

∃ No vertexes i, j s.t shortest path b/w i, j passes through A

∴ Betweenness centrality = 0

Vertex E :

Similarly, betweenness centrality = 0

vertex B :

Shortest path between A, D ~~must~~ passes through B in 1/1 case

" A, C " 1/1 case

" A, E " 1/1 case

∴ Betweenness centrality (normalized) $= \frac{\left(\frac{1}{1}+\frac{1}{1}+\frac{1}{1}\right)}{\frac{(n-1)(n-2)}{2}} = \frac{3}{4\times 3/2} = \frac{1}{2}$

~~(Total~~ Shortest path b/w C, D, and C, E and D, E do not pass thr

Vertex C :

Pair.	No of shortest paths	No of shortest paths through
A, B	1	0
A, D	1	0
A, E	1	0
B, D	1	0
B, E	1	0
D, E	1	0

∴ Betweenness centrality = 0

Vertex D :

By symmetry with B, betweenness centrality = $\frac{1}{2}$

Vertex	Betweenness centrality
A	0
B	1/2
C	0
D	1/2
E	0

Q5 [4 marks] There is a 100-length signature and we need to perform LSH to distinguish similar pairs. Consider 2 possibilities: (b=20, r=5) and (b=5, r=20). State where one should use parameters (b=20, r=5) vs parameters (b=20, r=5)? Explain your answer with the help of the S-curve.

Let Jaccard similarity of any two given colums be s. (colums C_1 & C_2)

~~In case of b = 20, r = 5~~

$\Rightarrow$ The probability of ~~any~~ given one row being ~~similar~~ same $= s$

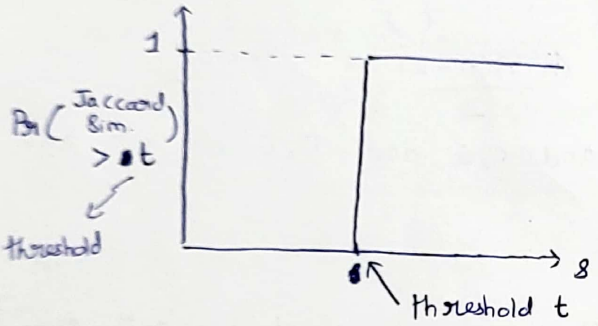
$\Rightarrow$ " r rows in a band being same $= s^r$

$\Rightarrow$ " band being different $= 1 - s^r$

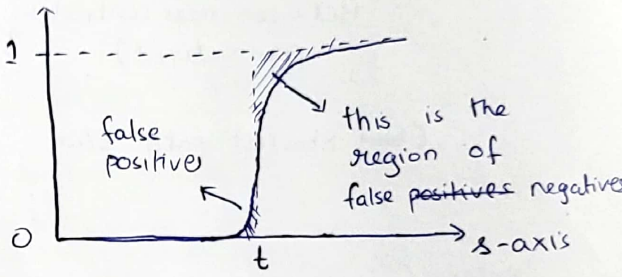
$\Rightarrow$ " all b bands being different $= (1 - s^r)^b$

$\Rightarrow$ " atleast one band is identical $= 1 - (1 - s^r)^b$

In case where atleast one band is identical, we can use LSH to hash those two columns to same bucket in atleast one case. (as they have high probability of being similar). Assuming min-hashing was done



Ideally



In practice
$1 - (1 - s^r)^b$

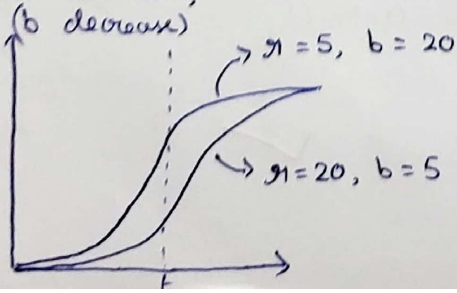
In reality, $1 - (1 - s^r)^b$ goes to the threshold function only for $r, b \to \infty$. Thus, with some probability, at Jaccard sim $< t$ can appear (false +ve) and Jaccard sim $< t$ ea may not appear (false negative)

By controlling r and b s.t $rb = 100$, we get a tradeoff b/w FP and FN.

In cases where it is desired that we want more FP than FN (as we can easily verify FPs but not FNs), we choose $r = 5$, $b = 20$.

In cases where we want more FN, then choose $r = 20$, $b = 5$. ✓

Reason : As r increases, Pr(FP) decreases and curve moves downwards (b decrease)



eg. Consider $s = 0.5$

Then at $t = 0.5$,

Pr(same bucket) $= 1 - (1 - s^r)$

When $r = 20$, $b = 5$, $= 0.0000$

$r = 5$, $b = 20$, $= 0.47$ ✓