# Final

New Attempt

**Due** May 4 by 11:40am **Points** 100 **Submitting** a file upload
**File Types** pdf **Available** May 4 at 9:30am - May 4 at 11:40am about 2 hours

**Q1 (20 points)** Given are the following eight transactions on items $\mathcal{V} = \{A, B, C, D, E, F\}$.

| Transaction id | Transaction (set of items) |
|---|---|
| 1 | ABC |
| 2 | BCD |
| 3 | CDE |
| 4 | BC |
| 5 | CD |
| 6 | ABCD |
| 7 | ABD |
| 8 | EF |

The support of an itemset/pattern $S \subseteq \mathcal{V}$ is the number of transactions containing all items of $S$. Let the minimum support be $minSup$ =2. A **frequent pattern** is an itemset whose supports are at least $minSup$. A **closed pattern** is a frequent pattern whose all supersets are less frequent than it. List all **closed patterns** and their corresponding **supports**.

**Q2 (15 points)** Given two data points $x, y \in \mathbf{R}^d$, define the distance between $x$ and $y$ as $dis(x,y) = 1 - \frac{x^T y}{|x|_2 |y|_2}$. Formulate an optimization problem of using such distance function to do K-Means clustering. Design an algorithm to solve this optimization problem. You can assume that the input dataset is $\{x_1, x_2, \dots, x_n\}$ and the initial $k$ centroids are $\{c_1, c_2, \dots, c_k\}$. You also need to show that your algorithm can converge.

**Q3 (10 points)** We want to conduct a survey to learn for a population, what is the percentage of people who like Justin Bieber. To protect the privacy, we adopt the randomized response technique, where one randomly answers "yes" or "no" with probability $p$ and answers truthfully with probability $1-p$. Suppose the ratio of people answering "yes" is $y$ according to our survey. How do we derive $\hat{x}$, an estimation of the real ratio of people who like Justin Bieber in the population?

**Q4 (15 points)** We have the following user rating matrix.

| | a | b | c | d | e | f | g | h |
|---|---|---|---|---|---|---|---|---|
| A | 4 | 5 | | 5 | 1 | | 3 | 2 |
| B | | 3 | 4 | 3 | 1 | 2 | 1 | |
| C | 2 | | 1 | 3 | | 4 | 5 | 3 |

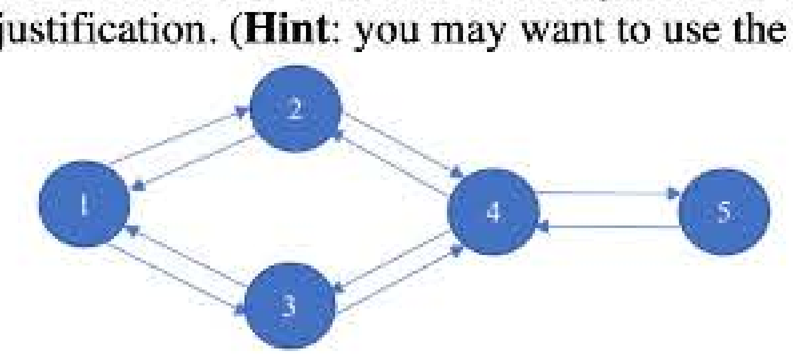
A, B, and C are users. a, b, …, h are items. Compute the following from the rating matrix.
4.1 (**5 points**) Treat the matrix as binary where observed ratings are regarded as 1 and missing ratings are regarded as 0. Compute the Jaccard similarity between each pair of users.
4.2 (**5 points**) Treat missing ratings as 0 and compute the cosine similarity between each pair of users. The cosine similarity between two vectors $x$ and $y$ is $\cos(x,y) = \frac{x^T y}{|x|_2 |y|_2}$.
4.3 (**5 points**) Use the cosine similarity defined in 4.2 to perform user-based collaborative filtering to predict the rating of user A on item c.

**Q5 (10 points)** Suppose the Jaccard similarity between to documents x and y is $s$. If we apply the banding technique where we use (4,3) AND-OR followed by (3,4) OR-AND, how many hash functions do we need? What is the probability that (x,y) is a candidate pair?

**Q6 (10 points)** Suppose we use the Independent Cascade (IC) model to model the influence diffusion in a network. In the IC model, we have all the seed node(s) activated at round 0. For each node $u$ firstly activated at round $t$, at round $t+1$, $u$ tries to activate each of its inactive out-neighbor $v$ (which means there is a directed edge $(u,v)$) with a success probability $pp_{uv}$. Note that if $u$ fails to activate $v$ at round $t+1$, $u$ will not have another chance to activate $v$ in the future rounds. The diffusion ends at a round when we do not have any newly activated nodes. Given a seed set $S$, to calculate the probability $p_v$ that node $v$ is activated in the diffusion started by $S$, one comes up with the following non-linear system.

$$p_v = \begin{cases} 1, & if\ v \in S \\ 1 - \prod_{(u,v)\in E} (1 - p_u * pp_{uv}), & if\ u \notin S \end{cases}$$

The intuition is that if $v$ is not a seed, $p_v$ depends on each of its in-neighbor $u$'s (which means there is an edge $(u,v)$) probability of being activated. Suppose we can solve this non-linear system exactly, which means we can find all the $p_v$ to make all the equations hold. Can we use the solution to this non-linear system to calculate each $p_u$ exactly? Please provide your justification. (**Hint**: you may want to use the following influence graph as an example.)




**Q7 (10 points)** Define the graph $G_n$ to have the $2n$ nodes

$$a_0, a_1, \dots, a_{n-1}, b_0, b_1, \dots, b_{n-1}$$

and the following edges. Each node $a_i$, for $i = 0,1,\dots,n-1$, is connected to the nodes $b_j$ and $b_k$, where $j = 2i \bmod n$ and $k = (2i+1) \bmod n$. For instance, the graph $G_2$ has the following edges $(a_0,b_0), (a_0,b_1), (a_1,b_0), (a_1,b_1)$.
7.1 (**5 points**) Find a perfect matching for $G_6$.
7.2 (**5 points**) Prove that when $n$ is an even number, we always have a perfect matching for $G_n$.

**Q8 (10 points)** For an undirected graph $G = <V, E>$ without loops with nodes $V = \{v_1, \dots, v_n\}$ and edges $E = \{e_1, \dots, e_m\}$ (edges are ordered pairs $e_i = (v_j, v_k)$ indicating a connection to node $v_j$ from node $v_k$, where $j < k$ for ensuring no duplicate edges), the $n \times m$ matrix $B = (b_{ij})$ is defined as

$$b_{ij} := \begin{cases} +1 & \text{if } e_j = (v_i, v_x) \\ 0 & \text{if } v_i \notin e_j \\ -1 & \text{if } e_j = (v_x, v_i) \end{cases}$$

with $v_x$ being an arbitrary node. Let $L$ be the Laplacian matrix of $G$. Show that $L = BB^T$.

Submission

✓ Submitted!
May 4 at 11:35am
Submission Details
Download 56046680.pdf

You may not see all comments right now because the assignment is currently being graded.