UIUC-CS512 “**Data Mining: Principles and Algorithms**” (Spring 2021)

First Midterm Exam

(Monday, Mar. 15, 2021, 100 marks)

IMPORTANT Notes

- Please provide brief explanations of your answers.
- Please sign the honor code in page 2. Your exam will not be graded unless the above agreement is signed. Please attach the signed honor code to your answer sheet.
- You can either (1) type in your answers in Latex/Word and submit your answer sheet in pdf, or (2) provide hand-written answers and submit a scanned version (pdf). For (2), Please make sure that the scanned version is clear and recognizable. Otherwise, you might loose points.

:

Name: NetID: Score:

1	2	3	4	5	6	Total

CS512 Spring 2021 Exam Honor Code

I understand that the rules of the CS512 first mid-term exam in Spring 2021. That is, (1) the exams are “open book”, and (2) I am not allowed to confer with other people about the questions or solutions to the exam (to either give or receive aid).

I have neither given nor received inappropriate aid during this exam.

I understand that my exam will not be graded unless the above agreement is signed.

NetID (print):

Name (print):

Name (signed):

Date:

1 Short Questions [10 points]

[True or False] For the following questions, answer true or false. If your answer is true, provide a brief justification. Otherwise, give a short explanation or a toy counterexample.

(a) (1 pt) If we train a classifier with few training samples, then the classifier is less likely to overfit.

(b) (1 pt) Self-training does not require labelled data.

(c) (1 pt) Transfer learning is effective when the source and target tasks share many common grounds.

(d) (1 pt) The results for two runs of K-Means algorithm on the same dataset are expected to be the same regardless of different initialization.

(e) (1 pt) In a transaction database, if the absolute support of the itemset X is larger than the absolute support of the itemset Y, then the relative support of the itemset X must be larger than the relative support of the itemset Y.

[Short Answers] For the following questions, give a short answer of few sentences

(a) (1 pt) Explain why Random Walk with Restart (RWR) is a good measure for node proximity in terms of catching information from both short and long distance?

(b) (1 pt) In frequent graph pattern mining, what is the major computational cost for pattern-growth approach? And list one solution for this challenge.

(c) (1 pt) What is the key difference between active learning and transductive learning?

(d) (1 pt) After the training for an SVM classifier is done, how the testing/evaluation is performed given a new data sample $\mathbf{x}'$?

(e) (1 pt) In Gaussian Mixture model, how to alleviate the problem of local optima?

2 Frequent Pattern Mining [20 pts]

(a) Given the database of five transactions in Table 1, let $min\_support = 60\%$ and $min\_confidence = 80\%$.

- (3 pts) Find all frequent itemsets.
- (2 pts) List all ***strong*** association rules (which satisfy both minimum support and minimum confidence) matching the following meta-rule,

$$\forall\ x \in transactions, x.item_1 \wedge x.item_2 \Rightarrow x.item_3$$

TID	Items
1	$\{A, B, C, D, E, F\}$
2	$\{H, B, C, D, E, F\}$
3	$\{A, I, D, E\}$
4	$\{A, X, M, D, F\}$
5	$\{M, B, B, D, T, E\}$

Table 1: Transaction database.

(b) Given the sequence database in Table 1.

SID	Sequence
1	$\langle b(bde)(be)gj(em)\rangle$
2	$\langle (bg)e(de)j(bj)\rangle$
3	$\langle (jm)(bdg)gmed\rangle$
4	$\langle ep(am)edeg\rangle$
5	$\langle eg(bem)(edj)bd\rangle$

Table 2: Sequence Database.

- (2 pts) Construct the projected database for prefix $\langle b\rangle$ and $\langle j\rangle$.
- (3 pts) Compute the support for the following sequential patterns, (1) $\langle bej\rangle$, (2) $\langle mdg\rangle$ and (3) $\langle mbj\rangle$.

(c) Given the graphs in Figure 1 (*Remarks:* the edges '=' and '–' are different).

- (5 pts) Let $min\_support = 0.6$, find all frequent patterns with size $k > 2$.
- (5 pts) We first define the *confidence* for graph association rules as follows, where $\mathcal{S}$ and $\mathcal{S}'$ are two graph patterns,

$$confidence(\mathcal{S} \Rightarrow \mathcal{S}') = \begin{cases} \frac{support(\mathcal{S}')}{support(\mathcal{S})}, & \text{if } \ \mathcal{S} \subset \mathcal{S}' \\ 0, & \text{otherwise} \end{cases}$$

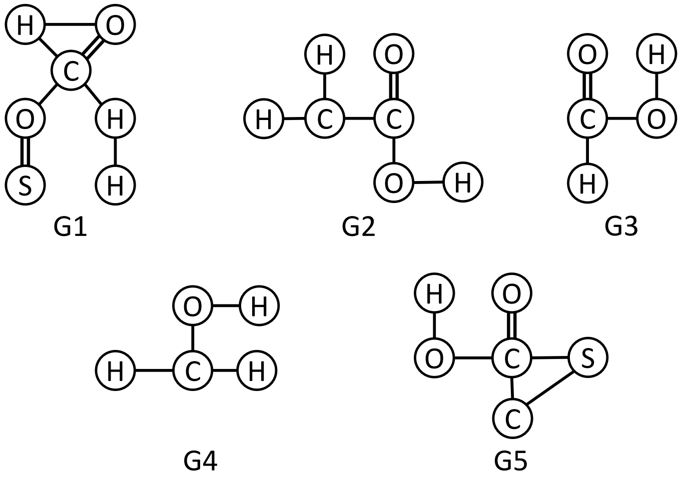

Figure 1: Input graphs for Problem 2-(c)

Let $min\_confidence = 0.7$ and $min\_support = 0.6$. List **one *strong*** association rule (which satisfies both minimum support and minimum confidence) and the corresponding confidence.

3 SVM [15 points]

Given a set of 1-D data samples as shown in Figure 2, three of them are positive data points $\{x = 2, x = 3, x = 4\}$, and two of them are negative data points $\{x = 1, x = 5\}$.

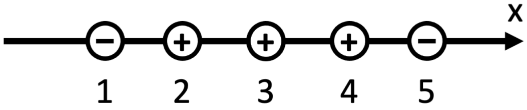

Figure 2: Training samples for SVM.

(a) [3 pts] If there is a mapping $\mathrm{R}^1 \to \mathrm{R}^2$ by which the mapped positive and negative data samples can be linearly separable? Plot the data points after mapping in 2-d space.

(b) [3 pts] If we implement a hard-margin linear SVM on the mapped data samples from problem (a), what is the decision boundary? Draw it on your figure and mark the corresponding support vector(s).

(c) [4 pts] For the feature mapping, what is the corresponding kernel $K(x_1, x_2)$?

(d) [5 pts] We change the data points into positive data points $\{x = 2, x = 3, x = 4, x = 6\}$, and negative data points $\{x = 1, x = 5\}$. Does there exist a mapping $\mathrm{R}^1 \to \mathrm{R}^2$ such that the mapped positive and negative data samples can be linearly separable? Plot the data points after mapping in 2-D space.

4 Random Walk with Restart [13 points]

Given an unweighted undirected network G as Figure 3 shows.

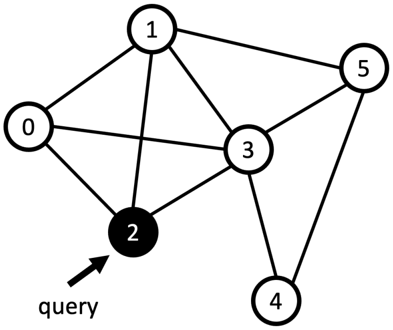

Figure 3: Network G with 6 nodes.

(a) [4 pts] What is the adjacency matrix $\mathbf{A}$ of the network G?

(b) [4 pts] If we adopt row normalization, what is the normalized adjacency matrix $\tilde{\mathbf{A}}$?

(c) [5 pts] Given the node 2 as the query node, based on the RWR formula: $\mathbf{r} = c\tilde{\mathbf{A}}\mathbf{r} + (1 - c)\mathbf{e}$, try to identify which node is the most relevant one to node 2 (except node 2 itself). Here, $\tilde{\mathbf{A}}$ is the row normalized adjacency matrix, $c = 0.9$. Initialize $\mathbf{e} = [0, 0, 1, 0, 0, 0]^\top$ and $\mathbf{r} = [1, 1, 1, 1, 1, 1]^\top$. Report the results of the first 3 iterations (no need to report the final converged results and round your result to 2 decimal places).

5 K-Means [17 points]

Given five data samples: $A = (-1, 2), B = (1, 1), C = (-3, 1), D = (0, -3), E = (2, -2)$.

(a) [3 pts] What is the Manhattan distance between data point A and E?

(b) [5 pts] If we run K-Means on the given data samples with Manhattan distance as the measure. $K = 2$ and the initial centroids are data D and data E. Report the results from the first 3 iterations with the position of centroids and the membership description of every data point. (e.g., in the first iteration, `centroid 1` : $(0, -3)$, `centroid 2` : $(2, -2)$, A is the member of cluster 1/2, ..., D is the member of cluster 1, E is the member of cluster 2.) [*Remarks.* If the distance from a data point to two centroids is the same, you should report 'Cannot decide the membership of data point X' and report the previous intermediate results]

(c) [4 pts] Has the K-means converged in the first 3 iterations on the provided data? Why?

(d) [5 pts] If we run K-Means on the given data samples with Manhattan distance as the measure. Let $K = 2$ and we randomly choose the initial cluster centroids (any initialized centroids including but not limited to the given A,B,C,D,E data points) and run K-Means with sufficient numbers of iterations, how many different possible clustering results are there? [*Remarks.* (1) No need to consider the case where a cluster is empty, and (2) no need to consider the case where the distance from a data point to two centroids is the same.]

6 2-Way Spectral Graph Partitioning [25 points]

Given the adjacency matrix $\mathbf{A}$ of an undirected unweighted graph G. $\mathbf{A}[i,j] = 1$ indicates that node i and node j are connected. Otherwise, $\mathbf{A}[i,j] = 0$. Based on the MinCut algorithm, we try to partition the graph G into two clusters based on the membership vector $\mathbf{q} \in \{-1, 1\}^n$, where n is the number of nodes in graph G. *Remark:* $\mathbf{A}[i,j]$ denotes the entry of matrix $\mathbf{A}$ at the i-th row and the j-th column; $\mathbf{q}[i]$ denotes the i-th entry of vector $\mathbf{q}$; $\mathbf{A}^T$ denotes the transpose of matrix $\mathbf{A}$.

(a) [5 pts] The loss function of MinCut can be formualted as:

$$J = \frac{1}{4}\sum_{i,j}(\mathbf{q}[i] - \mathbf{q}[j])^2\mathbf{A}[i,j]$$
$$s.t.\ \mathbf{q} \in \{-1, 1\}^n$$

Explain with your own words that why MinCut problem can be formulated by this loss function.

(b) [5 pts] Prove that the loss function of MinCut $J = \frac{1}{4}\sum_{i,j}(\mathbf{q}[i] - \mathbf{q}[j])^2\mathbf{A}[i,j]$ can be rewritten as $J = \frac{1}{2}\mathbf{q}^T(\mathbf{D} - \mathbf{A})\mathbf{q}$. $\mathbf{D}$ is a diagonal degree matrix whose entries $\mathbf{D}[i,i] = \sum_j \mathbf{A}[i,j]$

(c) [5 pts] The loss function of MinCut can be written in the matrix form as follows:

$$J = \frac{1}{2}\mathbf{q}^T(\mathbf{D} - \mathbf{A})\mathbf{q}$$
$$s.t.\ \sum_i(\mathbf{q}[i])^2 = n$$

Explain with your own words that why we need a relaxed constraint: $\sum_i(\mathbf{q}[i])^2 = n$.

(d) [5 pts] The solution $\mathbf{q}$ of the above loss function is the eigenvector corresponds to the second minimum eigenvalue. Explain with your own words that why we cannot select the eigenvector corresponds to the minimum eigenvalue to do the 2-way partition.

(e) [5 pts] If we change the loss function of MinCut by setting the constraint from n into $10n$ as,

$$J = \frac{1}{2}\mathbf{q}^T(\mathbf{D} - \mathbf{A})\mathbf{q}$$
$$s.t.\ \sum_i(\mathbf{q}[i])^2 = 10n$$

how will that affect the optimal solution q and the partition results? Explain with your own words.