Name	Saurabh Zinjad
ASU ID Number	1229549687

CSE 472: Social Media Mining

Homework I - Linear Algebra, Graph Essentials, Network Measures

Prof. Huan Liu
Due at 2023, September 7^{th}, 11:59 PM

This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Gradescope**. This is a fillable PDF and you are able to type into answer boxes provided for each question.

1. **[Linear Algebra]** Consider 2-dimensional data points of [-1, -2], [1, 0], [-1, 1], [2, 0], [4, 1].

   (a) Arrange the data points in ascending order based on their length and gather them together in the following matrix. Let's assume $[X]_{2\times5}$ is that matrix. Fill the following matrix. [*Hint:* The length of the vector $[x, y]$ is $\sqrt{x^2+y^2}$.

$$X = \begin{pmatrix} 1 & -1 & 2 & -1 & 4 \\ 0 & 1 & 0 & -2 & 1 \end{pmatrix}_{2\times5}$$

   (b) What is the point showing the center of these points? [*Hint:* Calculate the mean of the values in each dimension].

$$\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix}_{2\times1}$$

(c) Calculate $Y = (X - \mu)(X - \mu)^T$ in which X^T is the transpose of X. To calculate $(X - \mu)$, easily subtract the μ from all the data points.

$$Y = \begin{pmatrix} 18 & 5 \\ 5 & 6 \end{pmatrix}_{2\times 2}$$

(d) Solve $|Y - \lambda I| = 0$ to extract the values of λ. $|\cdot|$ is the determinant and I is the identity matrix. λ values are called eigenvalues.

$\lambda_1 =$ 19.81 , $\lambda_2 =$ 4.19

(e) Calculate the corresponding eigenvector to the **largest** eigenvalue (assuming the eigenvector has norm 1).

$$v = \begin{pmatrix} 0.94 \\ 0.34 \end{pmatrix}_{2\times 1}$$

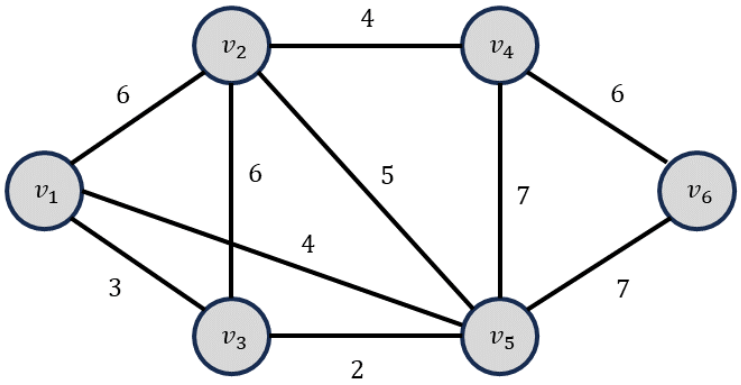
(f) Compute $\hat{X} = v^T X$.

$$\hat{X} = \begin{pmatrix} 0.94 & -0.6 & 1.88 & -1.62 & 4.1 \end{pmatrix}_{1\times 5}$$

Congratulations you performed Principle Component Analysis (PCA) procedure, a well-known dimensionality reduction method in machine learning. In other words, you projected your 2-dimesional data into 1-dimensional one such that you preserve the variance as much as possible (i.e. the least information has been lost).

2. **[Graph Algorithms]**

(a) Imagine a thriving social media platform called "ConnectWorld", where millions of users from around the globe come together to form a virtual community. Within ConnectWorld, friendships flourish, connections are made, and ideas are shared. In this scenario, each user within ConnectWorld represents a node in a graph, and the edges between them symbolize the effort that required to maintain the friendship. By applying **Prim**'s algorithm to the ConnectWorld graph below with starting **node** v_3, find the most cost-effective way to create a network that connects all users while minimizing the total effort required to maintain those connections. In the following table, at each step, write down the chosen edge and calculate the cumulative weight up to that step. Represent the edge between nodes v_i and v_j using the notation $v_i - v_j$. **Use as many steps as needed to solve this problem.**



Step	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Edge (start from v_3)	v3-v5	v3-v1	v5-v2	v2-v4	v4-v6			
Total Weights	2	5	10	14	20			

(b) Under what circumstances does the Prim's algorithm find multiple minimum spanning trees (MST) in a graph?

Prim's algorithm find multiple minimum spanning trees when:
1. Two or more edges have the same weights.
2. When the graph has disconnected components.

(c) In the space below, draw a simple example of a directed graph with negative-weight edges for which Dijkstra's algorithm produces incorrect answers.

src_node -> to_node, weight
---
1 -> 2, 8
1 -> 3, 6
2 -> 3, -5

Now, when finding the shortest path from node 1 to 3, Dijkstra's algorithm produces the answer '6,' which is incorrect.

The correct answer is '3', obtained by traversing from 1 -> 2 -> 3, with a combined weight of 8 + (-5) = 3.

Hence, Dijkstra's algorithm fails for negative weights.

(d) Argue whether "Algorithm 1" below always produces the shortest paths from one source node to others for graphs that have negative weights but do not have negative cycles.

No, Algorithm 1 does not always produce the shortest paths from one source node to others for graphs that have negative weights.

This is because Algorithm 1 attempts to modify the standard Dijkstra's algorithm for graphs with negative weights by subtracting a constant value (C) from all edges before running the actual Dijkstra's algorithm. If the original graph has very large negative weights, subtracting a constant might still result in negative edge values, which will still lead to incorrect results.

While it may work for some cases, it is not a guaranteed solution for handling negative weights.

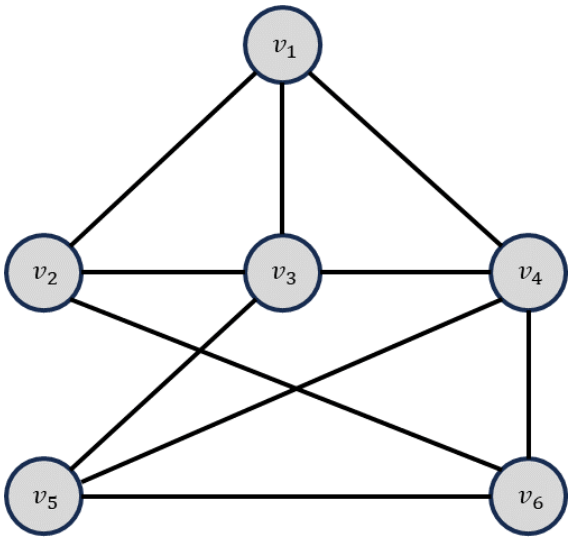
Algorithm 1: Dijkstra Algorithm for graphs with negative weights.

Input : Adjacency Matrix M, Source node s.
Output: Shortest Path from s to other nodes.

$C \leftarrow$ Find minimum weight in M
for all i and j:
    $M[i, j] \leftarrow M[i, j] - C$
return Dijkstra(M, s) *// use the original Dijkstra algorithm to find the shortest paths*

3. [**Network Measures**] Based on the following network answer the questions,

(a) Fill the adjacency matrix.



	v_1	v_2	v_3	v_4	v_5	v_6
v_1	0	1	1	1	0	0
v_2	1	0	1	0	0	1
v_3	1	1	0	1	1	0
v_4	1	0	1	0	1	1
v_5	0	0	1	1	0	1
v_6	0	1	0	1	1	0

(b) Calculate the "Betweenness Centrality" (normalized) values, "Closeness Centrality" and "Katz Centrality" values with $\alpha = 0.25$ and $\beta = 0.15$ (you can use Matlab or other mathematical software to calculate the eigenvalues).

	Betweenness Centrality	**Closeness Centrality**	**Katz Centrality**
v_1	0.03	0.714	0.92
v_3	0.13	0.833	1.11
v_6	0.083	0.714	0.88

(c) Is the above alpha value a good choice for Katz centrality? Why?

The given alpha value of 0.25 is a good choice for Katz centrality because it is less than the reciprocal of the largest eigenvalue of the adjacency matrix.

According to the Katz centrality definition, the value of alpha should be less than 1/lambda_max, where lambda_max is the largest eigenvalue of the adjacency matrix. Therefore, to determine if the given alpha value of 0.25 is a good choice for Katz centrality, we need to calculate the largest eigenvalue of the adjacency matrix and compare it with 4, which is the reciprocal of 0.25.

we can calculate the largest eigenvalue using python as follows:

```python
```

(d) Discuss what would happen to Katz Centrality if we set $\alpha = 0$?

if you decrease the value of alpha, you are reducing the influence of direct connections. This might be useful if you want to focus more on the importance of nodes in the context of their indirect connections.
Placing alpha as zero, basically we are not giving importance to influence of direct connections.

(e) Calculate the local clustering coefficient for nodes v_1, v_3, v_5, and v_6.

Local Clustering Coefficient for:
Node 1 - 0.6667
Node 3 - 0.5000
Node 5 - 0.6667
Node 6 - 0.3333

(f) Compute the similarity between nodes v_3 and v_5 using both Jaccard and Cosine similarities.

Jaccard Similarity between v3 and v5: 0.1667
Cosine Similarity between v3 and v5: 0.2887

Good Luck