Machine Learning with Graphs: Homework III - Spectral Graph Theory, Non-linear Embedding, Label Propagation

Due: 03/12

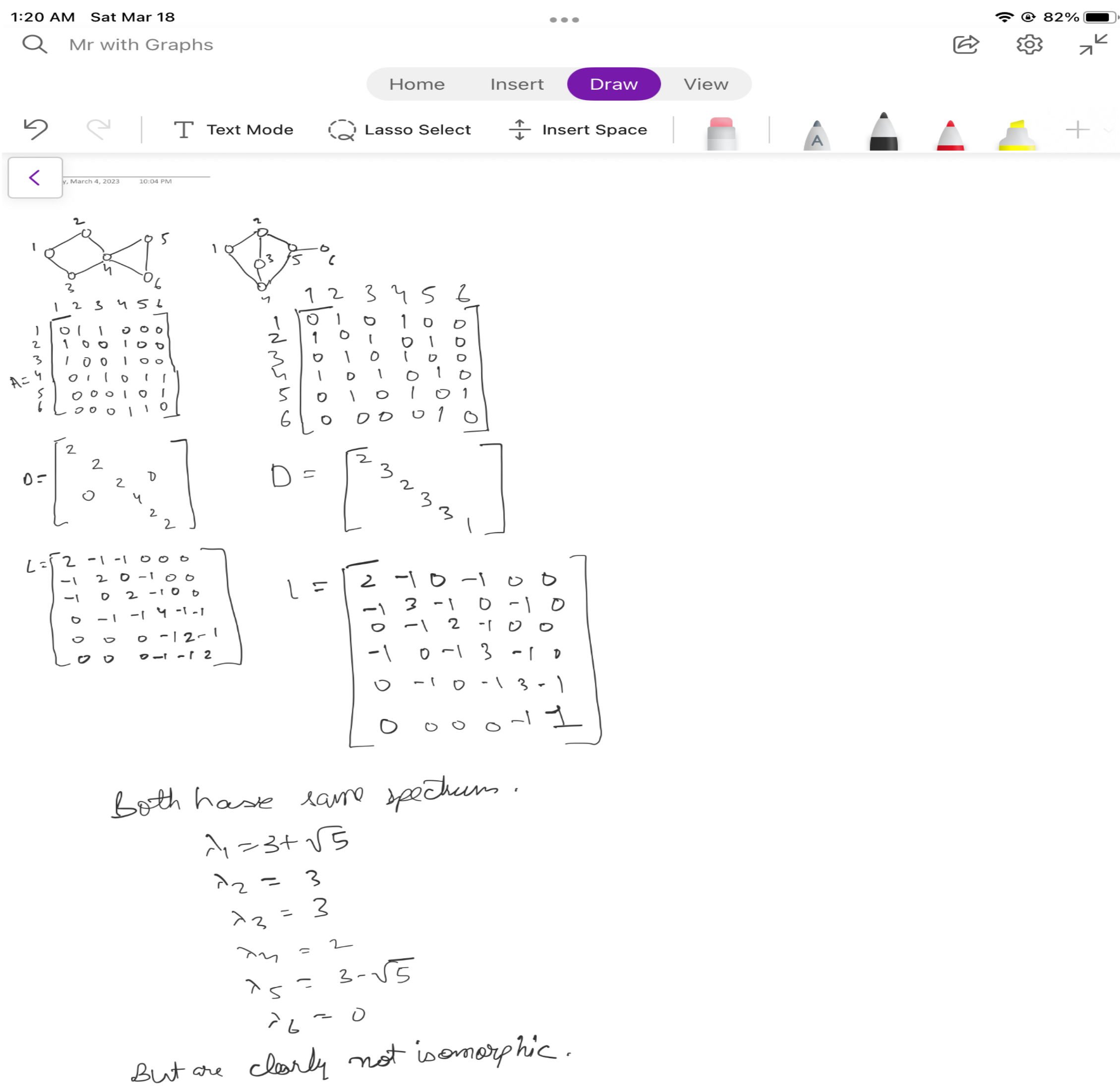
Spring 2023

You can choose between problems 3 or 4.

Problem 1

Two graphs, $G(V, E)$ and $G'(V', E')$, are *isomorphic* if there is a *bijection* $f : V \to V'$ such that $(u, v) \in E$ iff $(f(u), f(v)) \in E'$. Do isomorphic graphs share the same eigenvectors? And the eigenvalues? Propose an isomorphism checking algorithm based on the spectrum of the Laplacian matrix. Is this algorithm always correct?

Ans: If two graphs are isomorphic then their Adjacency matrix and degree matrix will be the same, hence their laplacian matrix will also be the same. Even though the laplacian is the same, it is not necessary that they will share the eigen vectors as eigen vectors are not unique. The eigen values will be the same.
The algorithm that I propose is as follows: If the laplacian of two graphs have different eigen values, then we can be sure that the graphs are not isomorphic. However, when the eigen values are the same, we cannot say with conviction that the graphs are isomporphic, but my algorithm will return true in this case. The algorithm is not always correct, the eigen values of different matrices can also be the same, so even if two graphs are not isomorphic they can have the same eigen values. But if two graphs are isomorphic then they will definitely have the same eigen values. So if the eigen values are not the same, we can be sure that graphs are not isomorphic. An example of a case where two graphs are not isomorphic but still have the same eigen values is shown below.



$$A = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 1 & 0 & 1 & 1 \\ 5 & 0 & 0 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 0 \end{array} \qquad \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 1 & 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$D = \begin{bmatrix} 2 & & & & & \\ & 2 & & & & 0 \\ & & 2 & & & \\ & & & 4 & & \\ & 0 & & & 2 & \\ & & & & & 2 \end{bmatrix} \qquad D = \begin{bmatrix} 2 & & & & & \\ & 3 & & & & \\ & & 2 & & & \\ & & & 3 & & \\ & & & & 3 & \\ & & & & & 1 \end{bmatrix}$$

$$L = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ -1 & 0 & 2 & -1 & 0 & 0 \\ 0 & -1 & -1 & 4 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix} \qquad L = \begin{bmatrix} 2 & -1 & 0 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ -1 & 0 & -1 & 3 & -1 & 0 \\ 0 & -1 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

Both have same spectrum.

$\lambda_1 = 3+\sqrt{5}$
$\lambda_2 = 3$
$\lambda_3 = 3$
$\lambda_4 = 2$
$\lambda_5 = 3-\sqrt{5}$
$\lambda_6 = 0$

But are clearly not isomorphic.

Problem 2

Consider the problem of selecting training vertices for a label propagation algorithm (similar to the problem in homework 1). You can consider any of the label propagation approaches described in the class. Propose an algorithm for selecting training vertices. You should justify your algorithm with either a (strong) theoretical argument or with experiments. You are free to make adjustments to the problem—e.g. assuming you have access to all vertex labels and want to select a subset or that only the graph structure is available. If you opt for a theoretical argument, you can make reasonable assumptions about the relationship between attributes and the graph structure. If you opt for an empirical argument, find a (labeled) dataset and compare your approach against an alternative that picks training vertices at random.

Code Link: Link to code

Problem 4 (optional)

The Cora dataset[1] is a citation network where vertices represent papers, edges represent citations, and vertex attributes are binary vectors indicating the absence/presence of words in the papers. Your goal is to apply spectral clustering to group the papers based on attributes (words) and the citation network (links). Show results comparing the clusters found using a clustering evaluation metric (e.g. normalized rand-score). Which type of information is more useful? How would you combine them? Because the graph is disconnected, you can use its largest connected component. Hints: (1) sklearn has an implementation of spectral clustering that works for attributes and a graph (adjacency matrix) and (2) vary the number of clusters predicted and use *nearest-neighbors* for the clustering based on attributes (it is faster).

Code Link: Link to code

[1] `https://linqs.org/datasets/`