TP4 Altergrad

Statement: Answers to tasks 4, 5, 11 are in the notebook

`visualize.ipynb`

as well as the results of the codes.
It was a pleasure doing this lab, I ve learned so much. I thought it was better for me to really give deep thoughts to my project and answers. I then accept being late on that project as far as this work has allowed me to go deep in the subject. And it's much more beneficial for me than rushing the project without really having learned anything from it, even if it means not submitting it on time.
I was short on time and rethinking about my organization between internship interviews other projects for MVA etc... can be challenging but I believe it's in my capacity to do better about time constraints for the next project. For that project I preferred to work consciously and deeply more than badly and I chose that way even if it's costing in terms of grading.

1.)

The formula for the number of edges in a complete graph K_n is:

$$E = \frac{n(n-1)}{2}$$

For $n = 100$:

$$E = \frac{100 \times 99}{2} = 4950$$

The formula for the number of edges in a complete bipartite graph $K_{m,n}$ is:

$$E = m \times n$$

For $m = 50$ and $n = 50$:

$$E = 50 \times 50 = 2500$$

Total Edges:

$$E_{\text{total}} = 4950 + 2500 = 7450$$

Triangles in K_{100}:
The formula for the number of triangles in a complete graph K_n is:

$$T = \binom{n}{3} = \frac{n(n-1)(n-2)}{6}$$

For $n = 100$:

$$T = \frac{100 \times 99 \times 98}{6} = 161700$$

Triangles in $K_{50,50}$:
A complete bipartite graph has no triangles because there are no edges within each partition set.
Total Triangles:

$$T_{\text{total}} = 161700 + 0 = 161700$$

2.)

First Graph

The number of edges is 13 in total **First Cluster: Green**

1. the number of edges within that community is $lc = 6$
2.
   - 1 has degree 3
   - 2 has degree 2
   - 3 has degree 3
   - 4 has degree 3

- 5 has degree 2

The sum is 13

Second Cluster: Blue

1. the number of edges within that community is $lc = 6$
2. - 6 has degree 4
   - 7 has degree 3
   - 8 has degree 3
   - 9 has degree 3

   The sum is 13 too

So

$$Q = \sum_{c=1}^{2} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right] \tag{1}$$
$$= 6/13 - (13/2 \times 13)^2 + 6/13 - (13/2 \times 13)^2 \qquad = 0.42307692307692313 \tag{2}$$

That suggest a good correlation between those two populations.

Second Graph

The number of edges is 13 in total **First Cluster: Green**

1. the number of edges within that community is $lc = 2$
2. - 1 has degree 3
   - 2 has degree 2
   - 9 has degree 3
   - 8 has degree 3

   The sum is 11

Second Cluster: Blue

1. the number of edges within that community is $lc = 4$
2. - 3 has degree 3
   - 4 has degree 3
   - 5 has degree 2
   - 6 has degree 4
   - 7 has degree 3

   The sum is 15

So

$$Q = \sum_{c=1}^{2} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right] \tag{3}$$
$$= 2/13 - (11/2 \times 13)^2 + 4/13 - (15/2 \times 13)^2 \qquad = -0.05029585798816555 \tag{4}$$

That suggest a weak correlation between those two populations. **3.)** Let P_n denote a path graph on n vertices and C_n denote a cycle graph on n vertices. We are tasked to calculate the shortest path kernel for the pairs (C_4, C_4), (C_4, P_4), and (P_4, P_4).

The shortest path kernel maps graphs into a feature space, where each feature corresponds to a shortest path distance and its value is equal to the frequency of that distance in the graph.

Feature Map $\phi(G)$

1. Feature Map for C_4: C_4 is a cycle graph with 4 vertices connected in a circular way. The shortest path distances between nodes are:

- Distance 1: 4 paths (each node connected to its neighbor).
- Distance 2: 2 paths (opposite nodes on the cycle).

Thus, the feature map for C_4 is:

$$\phi(C_4) = [4, 2, 0, \dots, 0]$$

2. Feature Map for P_4: P_4 is a path graph with 4 vertices connected in a line. The shortest path distances between nodes are:

- Distance 1: 3 paths (each edge in the path).
- Distance 2: 2 paths (from one node to another skipping one intermediate node).
- Distance 3: 1 path (from one end of the path to the other end).

Thus, the feature map for P_4 is:

$$\phi(P_4) = [3, 2, 1, 0, \dots, 0]$$

Shortest Path Kernel Calculation

The shortest path kernel $k(G, G')$ between two graphs G and G' is computed as the dot product of their feature maps:

$$k(G, G') = \phi(G) \cdot \phi(G') = \sum_{i=1}^{d} \phi(G)[i] \cdot \phi(G')[i]$$

1. Kernel (C_4, C_4):

$$k(C_4, C_4) = \phi(C_4) \cdot \phi(C_4) = (4 \cdot 4) + (2 \cdot 2) + 0 = 16 + 4 + 0 = 20$$

2. Kernel (C_4, P_4):

$$k(C_4, P_4) = \phi(C_4) \cdot \phi(P_4) = (4 \cdot 3) + (2 \cdot 2) + 0 = 12 + 4 + 0 = 16$$

3. Kernel (P_4, P_4):

$$k(P_4, P_4) = \phi(P_4) \cdot \phi(P_4) = (3 \cdot 3) + (2 \cdot 2) + (1 \cdot 1) + 0 = 9 + 4 + 1 = 14$$

Conclusion

The shortest path kernel values are:

- $k(C_4, C_4) = 20$
- $k(C_4, P_4) = 16$
- $k(P_4, P_4) = 14$

4.)

The graphlet kernel $k(G, G')$ is defined as:

$$k(G, G') = f_G^\top f_{G'}$$

where:

- f_G and $f_{G'}$ are feature vectors, with each entry corresponding to the frequency of a specific graphlet in graphs G and G', respectively.

If $k(G, G') = 0$, this implies that the graphs G and G' do not share any graphlets in common. In other words, the feature vectors f_G and $f_{G'}$ are orthogonal. For every graphlet type i, the frequencies $f_G[i]$ and $f_{G'}[i]$ cannot both be non-zero. Let's recall that this condition occurs when the sets of graphlets present in G and G' are disjoint. For example:

- If G contains only triangles (graphlet G_1) and
- G' contains only paths of length 2 (graphlet G_2),

there will be no overlap in their graphlet feature vectors, leading to $k(G, G') = 0$.

Example of Two Graphs:

- Let G be a **triangle graph** (C_3): This graph consists of 3 nodes and 3 edges forming a cycle. The only graphlet in G is the triangle G_1, so:

$$f_G = [1, 0, 0, 0].$$

- Let G' be a **path graph of 3 nodes** (P_3): This graph consists of 3 nodes and 2 edges forming a path. The only graphlet in G' is the path G_2, so:

$$f_{G'} = [0, 1, 0, 0].$$

For these graphs:

$$k(G, G') = f_G^\top f_{G'} = [1, 0, 0, 0] \cdot [0, 1, 0, 0] = 0.$$

Thus, the kernel value is zero, as the two graphs have no graphlets in common.