Assignment 3

CS5824/ECE5424 – Fall 2020

Out: Oct 12, 2020
Due: Oct 22, 2020 (11:59pm)

For all questions, submit both PDF and .ipynb file with all of your code and output. Both Colab and Jupyter are allowed. Please submit both files separately on Canvas, no need to submit a zip file.

Late submissions incur a 0.5% penalty for every rounded up hour past the deadline for the first 24 hours and 0.6% penalty for every rounded up hour past the deadline for the second 24 hours. For example, an assignment submitted 5 hours and 15 min late will receive a penalty of ceiling(5.25) * 0.5% = 3%. A grade of zero will be given on the third late day.

Be sure to include your name and student number with your assignment.

1. [**20 pts**] Given the citibike data represented as a graph $\mathcal{G}$ consisting of the set of nodes V representing bike stations, and the set of edges E representing connections between two stations (start station) and (end station). Use the data file $citibike.csv$ for the network data. To observe the data, use *networkx* package.

   Compute and print out the graph statistics. Each questions is worth **4 pts**.

   (a) The number of nodes (V) in the network.

   (b) The number of edges (E) connected between the nodes in the graph.

   (c) The maximum, minimum and average degree of the nodes in the network.

   (d) The number of nodes with more than 5 edges.

   (e) What is the average path length of the edges connecting the stations.

2. [**40 pts**] Use the same network data as question 1. Compute and visualize the statistics associated with the network. Each questions is worth **10 pts**.

   (a) Define centrality, betweenness centrality and closeness centrality. Plot the three metrics for the given network.

   (b) Visualize the network by choosing 20 random stations and plotting the degree distribution of each station.

   (c) Plot the data based on the Geo-locations of the stations (i.e Latitude and Longitude) and discuss your observations from the plot.

   (d) Provide qualitative analysis that you observed on the network data after you study properties of graphs such as degree distribution, connectivity, centrality measures and clustering coefficient.

3. [**40 pts**] Given cancer patients data in the form of a network that is represented in the form of a graph, $G = (V, E)$. Each patient is represented as a node and the nodes are connected by edges. The patients are labeled as malign(1) or Benin(2). The network data is included in two files. The file $'Edges.txt'$ contains edge information. Each edge connects two nodes (start node, end node). The first column of the data file is the start node and the second column is the end node. Assume each weight $W(i, j) = 0.5$. The file

labelednodes.txt provides information about the labeled nodes. Assume all other nodes are unlabeled. Implement a **probabilistic relational classifier** to predict the node labels using 5 iterations.

(a) predict labels of unlabeled nodes.

(b) compute prediction accuracy by comparing predicted labels to true labels provided in the file: *groundtruth.txt*.

Hint: Please follow Lecture Relational Learning (Part 2), Slide 30 to implement the classifier.