Homework #6 (Take Home Final)
Due May 10, 2021, 11:59pm

Problem 6.1.

Consider the graph X with 10 nodes shown on the figure above.

1. Create the adjacency matrix A for the graph X.
2. Using the powers of the adjacency matrix, calculate the number of closed triangles on X. Explain how you obtained the result.
3. Calculate the Laplacian matrix of X, and calculate its eigenvalues and eigenvectors.
4. Use the eigenvectors corresponding to the two lowest eigenvalues (different from 0) as the x,y coordinates to re-plot the graph.

Problem 6.2.

Use the CovidNYT database on the SciServer to perform the following statistical analysis from a Jupyter notebook:

(1) Using the Census data in the database, write and execute SQL queries to select the 200 Counties in the US with the highest and lowest population densities. Extract the total population in both categories. You can use the CasJobs of the SciServer to explore the databases interactively, and to debug your queries.

(2) The StatsC table contains the cumulative counts of both the Covid19 infections and deaths for each county in the US, for each day. Using the queries from the previous step, extract the maximum of the cumulative infections and deaths for the top and bottom 200 counties (in population density). This should give a total of 4 numbers (corresponding to Apr 18, 2021).

(3) Using the population counts, calculate the values normalized to a population of 100,000 and determine if there is a statistically significant difference between the highest and lowest population densities. Try to estimate the difference between the values in terms of standard deviations, both for the infection counts and for the deaths.
Problem 6.3.

The dataset for handwritten digits is stored in three csv files in the class data directory. \texttt{vals.csv} contains 1797 vectors. These each represent 8x8 pixel images turned into a 64 element vector. The true label (value) of the digit is in the file \texttt{targ.csv}, and \texttt{names.csv} contains the names of the digits (0..9). The task is to try to find an automated classifier of each image using two techniques, TSNE and PCA.

1. First, reorder the data vectors according to the target values, so that images with the same target values are adjacent to each other.
2. Each of the two methods (TSNE and PCA) should project the data down to 2D, so each image is represented by a 2D point. Before doing the transforms, write a function that creates scatter plots, coloring each point according to its label.
3. Run first the PCA returning 2 eigenvectors, and plot the result.
4. Run TSNE, and plot the result.

Hints:
6.2. Assume that the full counts are Poisson distributed, when you estimate the standard deviation. An example notebook on how a SQL database can be queried from Jupyter is listed below the homework links.
6.3. Use the Scikit-learn package for both operations.