CS:4980 Homework 3, Spring 2023

Released: Mar 28, 2023; Due: 11:59 pm , Apr 14, 2023

Reminders:

1. Out of 100 points. 3 Questions. Contains 4 pages.
2. If you use Late days, mark how many you are using (out of maximum 4 available) at the top of your answer PDF.
3. There could be more than one correct answer. We shall accept them all.
4. Whenever you are making an assumption, please state it clearly.
5. You will submit a solution pdf `LASTNAME.pdf` containing your answers and the plots as well as a zip file `LASTNAME.zip` that contains your code and any output files.
6. Please type your answers either in LaTeXdocument or in a separate file like a Word document and then convert it into a pdf file. Only drawings may be hand-drawn, as long as they are neat and legible.
7. Additionally, you will submit one zip file `LASTNAME.zip` that contains your code and any results files. Code and results for each question should be contained in a separate sub-directory (Eg: `Q1`) and there should be a `README.txt` file for each sub-directory explaining any packages to install, command to run the code files and location of the expected output. Please follow the naming convention **strictly**.
8. If a question asks you to submit code please enter the file path (Eg: `Q1/Q-1.3.1.py`) in the solution pdf.
9. If needed, you can download all the datasets needed for this homework from ICON. Information about the datasets will be given in the `README.txt` file.

1. (30 points) Submodularity

Q 1.1 (7 points) Prove that the coverage function is a monotone submodular function. Recall that the coverage function takes a collection of sets $\{S_i\}_{i=1}^n$ and outputs the size of their union $|S_1 \cup S_2 \cup \cdots \cup S_n|$.

Solution:

Q 1.2 (8 points) Let f be a monotone submodular function. Show that for any two sets S and T: $f(T) - f(S) \leq \sum_{e \in T} (f(S+e) - f(S))$.

Solution:

Q 1.3 (8 points) Let f be a monotone submodular function. Show that the following function is also submodular: $h(A) = \min(f(A), f(V/A))$ where V is the universal set of all elements.

Solution:

Q 1.4 (7 points) Let f be a monotone submodular function and let g be a concave function. Show that the following function is also submodular: $h(A) = g(f(A))$.

Solution:

2. (40 points) Sensors for detection in Network Model

We are going to empirically look at various strategies of selecting social network sensors to detect epidemic outbreaks. We will use the graph in `facebook.txt` [1] which contains a small subset of friendship network of a Facebook group. We will use SI model with $\beta = 0.005$ to simulate the infection.

Q 2.1 (6 points) As a warmup, simulate the SI model for upto $T = 100$ time-steps for 100 runs and plot the average fraction of S,I vs t curve. Select 4 nodes at random to be infected at $t = 0$. Also plot the average number of daily infected people for $t = 0, \ldots, 200$. Report the average time t^* at which the maximum daily infections is maximum.

Hint: Use the implementation of SI model in `sis_model.py` file. If you use another language you can convert the logic of the code in the file.

`Note:`

Solution:

Q 2.2 (15 points) Since the graphs are large, it is not always feasible to keep track of the states of all nodes in practice. Therefore, we will select a smaller subset of nodes to track during the epidemic which we call as sensors. We will implement three strategies for sensor selection:

- `RANDOM`: We choose k nodes uniformly at from the graph
- `FRIENDS`: We choose k nodes uniformly at random and for each of them we select a random friend. We will use these friends as the sensors.
- `CENTRAL`: We select the top k nodes with largest *eigenvector centrality* [2]. You can use functions like `nx.eigenvector_centrality_numpy`.

Set $k = 100$ for all strategies. The file `rand_nodes.npy` contains a random list of nodes. Select the first k nodes from the list to implement the `RANDOM` strategy. For `FRIENDS` strategy, use the k selected nodes from `RANDOM` strategy to randomly select their friends. Simulate the SI model for $T = 100$ steps and note the fraction $\tilde{I}(t)$ of the sensors that are infected at each time-step for each strategy. Note that sensors can also be infected at $t = 0$.

Plot average $\tilde{I}(t)$ vs t for all three strategies averaged over 20 runs. Also plot the average number of daily infections $\tilde{I}_d(t) = \tilde{I}(t) - \tilde{I}(t-1)$ over time.

Solution:

Q 2.3 (6 points) Report the peak time $\tilde{t}^*$ and peak daily infection for all 3 strategies (the time t where $\tilde{I}_d(t)$ is maximum is peak time).

The value $t^* - \tilde{t}^*$ is the *lead time* i.e., the difference in time between detection of epidemic peak among sensors and time when it peaks in entire population. Report the lead-time for all 3 strategies.

[1] taken from `https://snap.stanford.edu/data/ego-Facebook.html`

[2] `https://en.wikipedia.org/wiki/Eigenvector_centrality`

Solution:

Q 2.4 (5 points) Compare the lead-time of various strategies and explain difference in lead-time of various strategies.

Solution:

Q 2.5 (8 points) Now repeat Q 2.2 for $k = 50$ and $k = 500$. For each strategy submit a $\tilde{I}(t)$ vs t plot comparing different values of k. How does the lead time change with value of k for each strategy?

Solution:

3. (30 points) Flu Surveillance using Google Symptoms Data

We will study the efficacy of Google Symptoms Data [3] as a source of Flu surveillance. For this question we will use the data from the CDC about the ILI burden. We measure the usefulness of a signal using the correlation with ILI signals collected by CDC [4]. Specifically we will look at 2018-19 season (datasets can be downloaded from canvas).

Typically, epidemiologists focus on specific weeks of a year where the flu is prevalent called a *flu season*. A flu season starts at week 40 of a given year and ends at week 20 of the next year. We enumerate the weeks of a Flu season as *Epiweeks*.

For example, the 2018-19 season starts at week 40 of 2018 (Epiweek 1) and ends at week 20 of year 2019 (Epiweek 33). Since an year can have 52 or 53 weeks, a flu season can have 33 or 34 Epiweeks. The data for ILI is uploaded on canvas as `ILINet.csv`. In this question, we refer to `% WEIGHTED ILI` of the csv file as ILI values.

Q 3.1 (2 points) We will start by considering the state of Iowa. Extract the `% Unweighted ILI` from `ILINet_states.csv`. Plot the weekly `Unweighted ILI` of Iowa over 2018-19 flu season.

Solution:

Q 3.2 (10 points) CDC lists various symptoms for Flu [5]. We will consider the following symptoms: `Fever, Low-grade fever, Cough, Sore throat, Headache, Fatigue, Muscle weakness`.

Extract the Symptoms trends for each of these sympotms from the files `2018_symptoms_dataset.csv` and `2019_symptoms_dataset.csv` over the weeks of 2018-19 seasons for Iowa. Submit a single plot showing the trends of all the symptoms over 2018-19 seasons with x-axis showing the Epiweeks and y-axis the symptom trend values.

Solution:

Q 3.3 (3 points) We will use Pearson Correlation Coefficient (PCC) [6] to measure how correlated each of symptom's trend is to ILI. Evaluate PCC for each of the symptoms with

[3] `https://pair-code.github.io/covid19_symptom_dataset/`

[4] `https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`

[5] `https://www.cdc.gov/flu/symptoms/symptoms.htm`

[6] `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

ILI for 2018-19 season in Iowa. You may use functions like `scipy.stats.pearsonr` to evaluate PCC. Also, submit the code.

Solution:

Q 3.4 (10 points) Repeat Q3.1 and Q3.2 for following states: `California, Texas, New York, Alaska, Georgia` and `Mississippi`. For each state including Iowa, also report the symptom with the highest PCC along with the value of PCC. Can you think of any reasons for the differences in PCC across these states?

Solution:

Q 3.5 (5 points) There maybe some lead-time between reported ILI and the symptoms trend. Here, we define the lead-time t_s of a symptom s to be the value of t' that maximizes the PCC between ILI time-series from week $t'+1$ to T and time-series of symptom s from 1 to $T-t'$ where $T = 52$ week. Intuitively, we measure the delay between symptoms signal and ILI signal.

For the most correlated symptoms you calculated in Q3.3 for each of 7 states, find the lead-time for the respective symptom. Submit the code.

Solution: