Chapter 2

Project

A. Monty Hall

From Wikipedia:

> The Monty Hall problem is a famous problem in probability. The problem is based on a television game show from the United States, Let's Make a Deal. It is named for the host of this show, Monty Hall. In the problem, there are three doors. A car (prize of high value) is behind one door and goats (booby prizes of low value) behind the other two doors. First, the player chooses a door but does not open it. Then the host, who has knowledge of what is behind every door, opens a different door which he is certain has a goat behind it (opening either door with equal chances if the car is behind the player's door). Last, the host lets the player choose whether to keep what is behind the first door or to change choices to the third door (the one the host did not open). The rules of the problem are that the host has to open a door with a goat behind and has to let the player switch. The question is whether changing choices increases the chances of getting the car.

Questions

1. Do you think that changing choices increases the chances of getting the car?
2. Write Matlab code to simulate the problem. Was your guess right?
3. Can you explain the result you get with a mathematical proof?

B. Applied epidemiology

A class of 50 (N_{tot}) students is working hard to get their master degree. However they need some distractions, and each friday night they drink a lot, then randomly hook up (formation of 20 random couples, 10 random students being too drunk for this kind of activities). However one student has mononucleosis, a highly infectious disease: an individual hooking up with one infected individual has $P_{inf} = 1/2$ chances of becoming infected. Luckily this disease is curable, and each week an infected individual has $P_{cur} = 1/10$ chances of recovery, after which we consider he is immunized and can not be infected again.

Modeling choices

- We take time $t = 1$ at initial situation (one infected student, nobody is immunized).
- At each time point we apply first the "recovery" process then the "infection" process.

Questions

1. Run 1000 simulations of the spread of the disease.
2. Draw the expected number of infected individual as a function of (discrete) time, with standard error.
3. At time t1 = 6 weeks, draw the histogram of the number of infected individuals.
4. Draw the histogram of the time t_f at which the class becomes entirely disease free.
5. Same question for the histogram of time t_{half} at which half of the class is immunized. Exclude from this histogram the simulations in which the number of immunized students never reaches half of the class.
6. We now vary the initial number of infected students Ni. Draw the variance of the time t_f as a function of Ni.
7. Taking $Ni = 10$ and $N_{tot} = 500$, make a color plot of the total number of infected students as a function of P_{inf} and P_{cur}.

C. Social interaction network

In this exercise, we will use some data that come from Marcel Salathé's group. They measured contact networks in a high school to be able to predict the spread of infectious diseases and find the best vaccination strategy (you can see the publication: `http://www.pnas.org/content/early/2010/12/08/1009094108.full.pdf`). We will perform simple statistical analyses on these data.

Description of the raw data

Each node represents one person, identified by a unique number between 1 and 789. The folder `nodes` contains one file per node, with one line per interaction with another node. The lines are composed of the number of the other node and the time at which the interaction occurred. For example if the file `node-1` contains one line `2 1500` it means that persons identified by numbers 1 and 2 were interacting at time 1500 (arbitrary unit). All persons have one of the following role: student, teacher, staff, other. The file `roles` indicates the role of each person: 1 for student, 2 for teacher, 3 for staff, 4 for other.

Loading the data

We suggest that you read first the questions so you can choose the best way to store the data. Be careful that if A interacts with B at time t, then B interacts with A, however for some reason this reciprocal connection is not always reported.

Questions

1. Who interacted with the biggest number of other people during the day? What is his role?
2. Is there a statistical difference in total number of interactions during the day between students and teachers?
3. Do students interact statistically more with other students than with teachers?

D. Mortality from influenza

All the data we will use in this exercise come from `http://epidemiology.mit.edu/`. This is a real dataset used by epidemiology researchers, working mainly on cancer but having also other data. We will focus on mortality by influenza and try to produce simple figures from raw data.

Description of the raw data

The file `influenza.xls` contains data about mortality by influenza in United States for European American population between 1900 and 2006. "EAM" refers to European American Males, and "EAF" refers to European American Females. "Raw Data" is the number of deaths from influenza at each time point and in each age category, "Population" is the number of people alive at each time point and in each age category, "1 minus TOT" is 1 minus the mortality by any cause at each time point and in each age category.

Loading the data

We suggest that you first read the whole exercise before choosing which data you will store and how you will store them. Be careful about possibly missing data. You will probably have to make approximations, interpolations, ...

Output

1. A first kind of graph we want to extract from these data would show us, for a given time point (parameter), what is the mortality by influenza (y-axis) for each age category (x-axis).
2. A second kind of graph we want to extract from these data would show us, for a given age category (parameter), what is the mortality by influenza (y-axis) at each time point (x-axis).
3. A third kind of graph we want to extract from these data would show us, for people born in a given year (parameter), what is the mortality by influenza (y-axis) at each age (x-axis).

Plot one graph of each category. Do not forget to label the axes of your plot, add a legend, ...

E. Seam Carving

The following links explain an automated image resizing algorithm called seam carving. Watch the Youtube video, read the Wikipedia page. The third link is the paper by the inventors of this technique which provides more details if you need.

- `http://www.youtube.com/watch?v=6NcIJXTlugc`
- `http://en.wikipedia.org/wiki/Seam_carving`
- `http://perso.crans.org/frenoy/matlab2012/seamcarving.pdf`

1. Try to summarize with your words what is seam carving and what are the interests and weaknesses compared to other resizing techniques.
2. We want you to implement seam carving with Matlab. Here are the steps we suggest:
   - Identify the arguments of the Matlab function.
   - Load the image (*Documentation:* **imread**.).

- Compute the energy matrix.
- Choose and implement an algorithm to find the best seam to remove.
- Perform the removal and update the energy matrix.
- Loop the last two steps.
- Save the image (*Documentation:* **imwrite**.).

3. Test on several images (`http://images.google.com/`).
4. Every improvement or innovation you can do is welcome! For example you can try to use other energy functions, to improve the performance of your program, to add a user interface ...

F. Iterated prisoner dilemma

The prisoner dilemma is a classical game-theory problem in which two players are going to decide at the same time to cooperate or not with the other player. When both choose to cooperate, they get a payoff of 1. When none cooperate, they get -1. When only one cooperate, he gets -2 and the other (that does not cooperate) gets 2.

Briefly said, when both player cooperate it is a win-win situation. However a player that does not cooperate get even more because he "exploits" the other. But if both do not cooperate it becomes a lose-lose situation. What makes the game interesting is that "players" will encounter each other a high number of times, and will remember previous interactions, so they can choose to act according to their "history" with a particular player.

The goal of this exercise is to write the best possible strategy, to win first against a determined number of "bots", then against the strategy of other students. A strategy is a Matlab function that takes the arguments `myID, opponentID, GameHistory` (described below) and returns 1 if you wish to cooperate, 0 if you wish not to cooperate.

- `myID` is your ID (a number that identifies you in the history of interactions). It is assigned randomly at the beginning of the tournament and stays the same.
- `opponentID` is the ID of your opponent (that identifies him in the history of interactions)
- `GameHistory` represents the history of interactions you had with this opponent. Each line represent a past interaction, formed of the number of the "fight", the ID of opponent A, the ID of opponent B, the behaviour of opponent A (1 for cooperate, 0 for not cooperate), the behaviour of opponent B. `GameHistory(j,:)=[IdFight, ID opponent B, ID opponent B, decision A, decision B]` for jth fight. Note that for each line of this history you can be either opponent A or opponent B (this can be different each time).

1. Download the archive `http://perso.crans.org/frenoy/matlab2014/PD.tar.gz` and extract it in your Matlab folder.
2. Write a strategy that respect the specifications described above, and modify `TournamentScript.m` so it adds your strategy to the competition with predefined bots.
3. When you are satisfied with your strategy, send it to me by email, and we will make a competition with other students strategies, past year students strategies, and my strategy, in addition to existing bots.