DS102 - Project Part 2

Data science is a collaborative activity. While you may talk with others about the project, we ask that you write your solutions individually. If you do discuss the project with others please include their name in your submission.
You must submit a typed up pdf on Gradescope for the project. We will not accept handwritten submissions. You can optionally use the provided LaTeX template found at https://github.com/ds-102/fa19/tree/master/project/project02.tex to type up your project.
Due by: 11:59pm, Tuesday 10th December, 2019

This is part 2 (of 2) of the project. The setting is the following:

Setup The company is impressed with your work analyzing the public data sets on bike-share usage, and they've hired you full time. Additionally, they've revealed some information on what product they're making exactly: self-driving scooters.

Since the company is still quite small, your job description covers diverse tasks. Specifically, your first two projects deal with (1) designing an adaptive strategy for recruiting users for your scooter share program, and (2) keeping your users' data safe and private. Instructions on each of the projects are given below.

Instructions Like in part 1, we've provided an outline, including a set of prompts and subsections for you to fill out. Please **do not alter the section numbers**. If you wish to copy the outline into a different text editor (google docs, word), that's fine, but please keep the section and subsection structure consistent.

Unlike in part 1 of the project, **all of the prompts are mandatory**.

Remember that this is an independent project. While you can talk to others, your analysis (especially for open ended questions) should be from your own ideas. Please submit your filled-in report as a pdf on gradescope. You must also email any code you use in a zip file to karlk@berkeley.edu with files clearly named to match each section, and subject line "DS102 - project02".

Grading The major components of your grade are:

- **Content**: in each part of the assignment there are structured questions for you to fill out.
- **Completeness**: For all parts of the report, carefully and completely describe what you did. As a rule of thumb, a reader who has taken DS100 but not DS102 should be able to reproduce your analysis without referring to any of your code.

- **Creativity**: For open ended problems, we expect to see a careful or creative integration of topics from class. Document why you're approaching questions with certain techniques, and if things don't work out as expected, that's ok!
- **Professionalism and Readability**: We expect full sentences with correct grammar and spelling. All axes in all figures should be labeled, with units when applicable.

1 Bandits

Your company is considering raising awareness of your scooter-share program by sending a person out around your city to hand out promotional flyers to people passing by and answer questions about your company. The company would like to know what parts of the city to send this person out to. They have a candidate list of intersections, but the idea is that the best intersection to recruit users at could be *learned* over time. There is only one recruiter to hand out flyers, and you can send them to only one intersection every day. The company would like you to (a) formulate this as a multi-armed bandits data-acquisition problem, (b) research the feasibility of using the UCB algorithm for placing the flyer-handouts in the city each day by simulating such an approach on the data sets on bikeshares from your previous project, and (c) return this report summarizing your findings and recommendations.

1.1 Formalizing the problem as a multi-armed bandits problem

First, formalize the problem as a multi-armed bandits problem to maximize the number of flyers given out at the street corner, over some number of days for which the promotion will take place.

Be sure to address at least the following questions in describing the setup you've chosen:

- What are the arms? What are the rewards? Would you model the rewards as sub-gaussian or as bounded?
- What is the time horizon?
- What are the modeling assumptions?
- Which of your assumptions do you expect to be reasonable? Which might you not expect to hold? Why? Which assumptions can you test by gathering data?
- What notion of *regret* are you considering? In particular, which strategy (possibly with perfect knowledge) would you compare your adaptive strategy to?

1.2 Simulate UCB strategy using past data

Here you will simulate the effectiveness of using UCB to determine the best locations with past data, using the datsets `dc.csv`, `chicago.csv`, and `ny.csv`.[1]

In order to fit the simulation with the data given, we'll consider a formulation of the problem that might be different from the bandits problem you've defined above. For the purposes of the simulation, define the arms to be the locations of the bikeshare stations in each city. Also define rewards to be the total number of riders who use the bikeshare station you've chosen for that day (either by starting a ride there or ending a ride there). You want to count the number of rides, so that double-counting the same person is okay if they go to the same station twice.

[1] The NY data is the biggest, so you might want to debug your code first with the DC or Chicago data set.

The goal is to use this data to simulate how a UCB algorithm would perform in this setting. To do so, you will sample observed rides from the dataset. For each day in the dataset, your algorithm will choose a station at which to hand out flyers. In addition to handing out the flyers, your employee will record the number of people only who took rides to or from that specific station. Using that number, the UCB algorithm will update the estimated mean of daily bikeshare users at that location. Therefore, to simulate running the UCB algorithm, you should only sample the value for the location that your UCB algorithm has chosen for each day.

Before simulating the UCB algorithm, take the following steps to make the data sets ammenable to the simulation:

1. Convert each data set to a data set where for each day, you have access to the total counts of bike trips (started + ending) at each station.

2. Next, keep only the entries corresponding to the stations with the 10 highest numbers of trips over the entire time span of the data set. This corresponds to the event that you know which 10 stations are the busiest, but you don't know what their daily traffic rates look like.

1.2.1 Implementation and Results

In this section of the write-up, include the following:

- Describe your simulation procedure at a level of detail such that someone who has taken DS100, but not DS102, could recreate your analysis without reading your code. Make sure to address the following:
  - Are you assuming sub-Gaussianity or boundedness of the rewards? How are you instantiating the associated parameters?
  - How do the means and upper confidence bounds get updated at every step of the algorithm? How do you choose the widths associated with the upper confidence bounds?
- For each of the three cities, plot the following:
  - Regret of the UCB algorithm over time. Define regret with respect to the best strategy that visits the same location every day.
  - The number of total pulls of each arm over time, with each arm labeled by its true mean (average daily usage).
  - Estimated means and upper confidence bounds for each arm over time, with each arm labeled by its true mean (average daily usage).

1.2.2 Discussion

Now, comment on the following aspects of your simulation methodology, before commenting on the results and applicability of your results in the next section.

- Do the results of your simulation change over different orderings of which arms to pull first and in what order?
- Consider the notion of regret that we've defined. Can you think of an adaptive strategy that could possibly beat the no-regret strategy of always sending your employee to the location with the highest average number of rides, in expectation? *(hint: consider the structures in this data that you have found in previous analyses).*

1.3 Takeaways

In summarizing your findings, discuss the following, as well as any other insights you have to relay to the marketing team. Your response for each bullet point should be between 2-4 sentences.

- Discuss the applicability of your simulation to the problem you posed in Section 1.1. In what ways is this simulation suitable for answering whether it is worth investing in the promotional program in your city, at arbitrary street corners? In which ways is the simulation experiment lacking toward answering this question?
- Discuss the applicability of the multi-armed bandits formulation to the task at hand in Section 1.1. Are any traditional assumptions (such as independence of the rewards for each pull of the arms), violated, and how?
- Ultimately, would you recommend using UCB to adaptively place the person handing out promotional flyers? Would you suggest modifying the algorithm to take into account any structure in the problem that your simulations exposed? Would you recommend a different approach altogether to raise awareness of the company? Your recommendation should be backed up by your simulation findings, as well as your knowledge about multi-armed bandits problems in general.

2 Privacy Concerns

Given the usefulness of the bikeshare datasets your company decided to release a dataset of scooter rentals that they have aggregated over the past year within Berkeley. The dataset can be found at https://github.com/ds-102/fa19/tree/master/project/berkeley.csv and has a very similar format to the chicago.csv, ny.csv, and dc.csv datasets. It includes the zip code at which the scooter was rented, and the zip code it was dropped off at, the year and birth month of the renter, and the sex of the renter. The public version also includes the exact GPS coordinates of scooter pickups and dropoffs but we exclude it here for simplicity.

Unfortunately a separate dataset was leaked which includes the full name of all the users, their month and year of birth, their sex, and the zip code of their address. You can find the

leaked file at https://github.com/ds-102/fa19/tree/master/project/leaked.csv. You've been tasked with investigating the severity of this leak.

2.1 Exploratory Analysis

You'll start by conducting a very quick exploratory analysis to get familiar with the berkeley.csv and leaked.csv datasets.

- Plot the number of females and the number of males in leaked.csv.
- Plot the distribution of birth months in leaked.csv.
- Plot the distribution of birth years in leaked.csv.
- Are each of these three attributes uniformly distributed? Are the distributions similar to the datasets you explored in part 1 of the project?

2.2 Simple Proof of Concept

In this section you will investigate whether you can perform a linkage attack given berkeley.csv and leaked.csv. We'll start with a simpler attack that only uses a subset of the leaked data

- Isolate the users from leaked.csv that can be uniquely identified based only on year/month of birth and their sex. We will call this subset "identifiable users" for the rest of this section.
- How many users can be isolated based on just those three attributes?
- Plot the number of females and the number of males in the set of identifiable users.
- Does the distribution you just plotted roughly match the corresponding distribution you plotted in section 3.1? Explain why the distribution matches or does not match.
- Plot the distribution of birth months for identifiable users.
- Does the distribution you just plotted roughly match the corresponding distribution you plotted in section 3.1? Explain why the distribution matches or does not match.
- For each identifiable user, write a script that will extract the scooter rentals they have made from berkeley.csv.

2.3 A More Elaborate Attack

You would like to leverage the information you gained from Section 3.1 to show that you can identify even more user rentals from berkeley.csv. In this section you will conduct a more elaborate linkage attack that makes use of the zip code information to demonstrate this.

Assume that users will tend to rent scooters with start and end location with the same zip code as their address. Users will rent a bike from another (uniformly sampled) zip code with probability p_1, similarly users will end a rental at a different zip code from their address with probability p_2. In other words the distribution on whether a user will start or end a rental at a zip code different from their address is given by $Bern(p_1)$ and $Bern(p_2)$ respectively.

- Given the assumptions above, estimate the parameters p_1 and p_2 using the trips made by the set of identifiable users. Describe how you went about estimating those parameters.
- Generate 95% confidence intervals around both p_1 and p_2 and describe how you went about generating those confidence intervals.
- Assuming berkeley.csv had actually included the address of each user, isolate the users that can be uniquely identified based on all features in leaked.csv. We will call this set the "theoretically identifiable users" for the rest of this section.
- How many theoretically identifiable users are there?
- Assume that your estimated values of p_1 and p_2 are actually equal to the true value. Implement and describe an algorithm that, given a specific trip, will return the most likely user to have generated this trip. If there are multiple most likely users randomly select a single user.

2.4 Takeaways

In summarizing your findings discuss at least the following, as well as any other insight you have

- Summarize your findings in at most three sentences.
- Make a recommendation as to what should be done with the already released data.
- Suggest a better way to release future datasets, if at all.