Mini-Project 1 – Part 3

ECE 471

Fall 2024

Zach Larson (zlarson2), Matthew Paul (mjpaul3), Maximo Rojas Pelliccia (mgr9)

Note: small decimal numbers were rounded to 3 significant figures or the best representation.

For complete decimal numbers, refer to the Jupyter notebook.

Task 4

Q1: Parse the provided Carla simulation dataset and calculate the following probabilities for the cut-in scenario (you need to filter our invalid data points before doing the analysis):

a. The probability of accident P(acc=1) across all weather conditions.

   P(acc=1): 0.089989898989899

b. The probability of an accident conditioned on the weather, (acc=1 | weather=?), for each weather condition.

   P(acc = 1 | weather = clear) 0.008130081300813009
   P(acc = 1 | weather = cloudy) 0.04251012145748988
   P(acc = 1 | weather = rain) 0.12474849094567404
   P(acc = 1 | weather = snow) 0.1830985915492958

Task 4

Q2: The baseline simulated dataset contains accident information for snowy conditions and rainy conditions. In California it is sunny 80% of the time, rainy 5% of the time, snowy 2% of the time and cloudy the rest of the time. In Chicago, it is sunny 60% of the time, it rains 15% and it is snowy 20% of the time, and cloudy the rest of the time. Calculate the probability of an accident in the cut-in scenario for California and Chicago, respectively. Clearly state your assumptions and method

- Assumptions:
  - Assuming prior probabilities are accurate
  - Assuming exactly one weather condition can be occurring at any time

- Method
  - Law of Total Probability: P(acc) = P(clear)**P(acc|clear)* + *P(cloudy)**P(acc|cloudy) + P(rain)**P(acc|rain)* + *P(snow)**P(acc|snow)

P(acc = 1) in California: 0.0219297772083937 1
P(acc = 1) in Chicago: 0.06233554680507255 6

Task 4

Q3: In Part 2, Task 3.4, you calculated the AV's probability of an accident per mile for the California DMV dataset. Suppose you want to compare the simulated accident rate with the real dataset accident rate .

a. Unfortunately, the real DMV only has sunny and cloudy weather:
    i. How would you make a reasonable comparison between the probability of an accident of the simulated dataset and the real dataset in this case?

To make a more reasonable comparison, we can remove rain and snow data points from simulated dataset.

    ii. Compare the two probabilities and report your findings

Real dataset accident rate 0.0011511258263112764
Simulated dataset accident rate 0.012935893365724 83
The simulated probability is about 10x higher

b. Suppose that there are **k** cut-in scenarios per one mile, and the probability of having an accident in one cut-in scenario is ***p***. What is the probability to have at least one accident per one mile? State your reasoning in detail

Binomial: P = 1 - (p^0 * (1-p)^k)
We chose to use a binomial distribution to model the k cut-in scenario because for each cut-in, 1 - p is the probability for no accident. Thus, to get the total probability, we raise this value to the k-th power and subtract it from 1.

Task 4

Q3: In Part 2, Task 3.4, you calculated the AV's probability of an accident per mile for the California DMV dataset. Suppose you want to compare the simulated accident rate with the real dataset accident rate .

c. Why do you think the probabilities of accident are so different between the simulation and the rea-world data

One reason for the difference is that the real world probability is calculated with an arbitrary reaction time. It also has the assumption that disengagement + slow reaction ==> accident (always). Also the prior weather probabilities in the simulation probably do not match up with the real world weather probabilities. Finally, the simulation had equal distribution of weather conditions which was not the case for the real world data (~80% cloudy and ~20% clear).

d. What are the pros and cons of the simulation data given the discrepancy between probability of accident in simulation and the real-world data

Pros: Simulation accounts for more varied weather conditions ; Easier to generate simulation data ; We can control other variables when collecting data.

Cons: Not directly translatable to real world scenario ; Inaccurate due to model assumptions that are not present in real world.

Task 4

Q4: What are your suggestions to improve MP1? What difficulties did you encounter? Please make sure each member individually (mentioning the member's name) lists their suggestions for improvements on that problem.

Matthew Paul
- Be more clear in part 1 which parts of the dataset to use in each question.

Zach Larson
- Give recomended resources and libraries.

Máximo Rojas
- Give template at start of MP instead of right before the deadline.