CPSC 310: Homework 2

Due Feb. 21st, 2020 at 11:59pm through Canvas Upload

Suppose that a change is made to Yale infrastructure so that (a) you must swipe into every building at all times, and (b) a central database is kept of all building or dining-hall entrance events logged by the IoT locks or registers. You've been tasked with using data mining and machine learning, together with this new dataset in `door_data.csv`, to help the university staff better prepare for traffic in dining halls.

The dataset[1] is provided as a comma-separated values file with the following fields:

- Student ID: (16 digit number)
- Day: (sequential day of collection, you'll receive around 28 days of data)
- Day of the Week: (number from 0-6, representing Sun-Sat)
- Time of Day: (minute represented by number from 0-1439)
- Building: (numerical identifier for row in `building_codes.csv` file)
- Dining Hall: (boolean represented as 1 or 0, indicates whether student swiped into the dining hall)

You are also provided with a `readme.txt` file with some additional information helpful for parsing both `door_data.csv` and `building_codes.csv`. You are welcome to use any language/software package to implement your program, but the `scikit-learn` library with Python is recommended and has ample documentation. Make sure to document clearly across your code which problem it is used for; you are welcome to use separate scripts as well.

Part A)

1. Your task for the day is to help Yale Dining fulfill demand at the lunch rush, ensuring that no dining halls run out of food before the end of lunch by routing extra ingredients to busy dining halls. The goal is to take early semester data and project out demand to make the dining halls more efficient for the remainder of the semester. To do this, you must train a decision-tree classifier that, given the college, day of the week, swipes by 11:45am, swipes by 12:00pm, and swipes by 12:15pm, predicts whether it'll be a low-traffic day ($< 95\%$ mean traffic for the college), a high-traffic day ($> 105\%$ mean traffic), or a normal day ($\geq 95\%$ and $\leq 105\%$ mean traffic). Do this for Silliman College and Branford College. What is the prediction for:
   - Branford with 78 swipes by 11:45am, 131 swipes by 12:00pm, and 232 swipes by 12:15pm on a Monday?
   - Silliman with 90 swipes by 11:45am, 171 swipes by 12:00pm, and 230 swipes by 12:15pm on a Sunday?

[1] Disclaimer: the provided dataset for this assignment is composed of randomly generated fake data.

For each, include a representation of the decision structure of the tree for it, and explain the corresponding decision. Does the tree logic match behavior you'd expect of dining hall use? You may find the `scikit-learn` documentation and the content at `https://christophm.github.io/interpretable-ml-book/tree.html` helpful introductions to decision trees, the latter in particular for interpreting the behavior of your decision trees. (Hint: You'll need to process the data before training. To get the decision structure with `scikit-learn`, you can use graphviz as is shown in the documentation.)

Part B) After completing this initial task, your manager returns with a new project. Your company wants to pitch Yale on a tool that will, given a specific student identifier, uncover "behaviors of concern" and autogenerate a notification to students of available mental health and wellness resources. Given a list of students (provided in `student_list.txt`), write code that will complete the following tasks.

1. Find all students on the list who skip more than 7 brunch/lunch or dinner meals in a week at least twice over the 28-day period. In a `readme.txt` file, explain your approach and reasoning used to make this determination.
2. Find all students on the list who appear to skip all classes and academic activities for a week. In the readme, explain the approach and reasoning you used to make this determination.
3. Find all students on the list who swipe back into a residential college or dorm between 3:00am and 5:00am on at least three non-weekend nights (i.e., not Friday or Saturday night) over the 28-day period. In the readme, explain the approach and reasoning you used to make this determination.
4. Find all students who, on at least three occasions, do not appear to leave a residential college or swipe into a dining hall for a full calendar day. In the readme, explain the approach and reasoning you used to make this determination.
5. Finally, for all students that fulfill all of these criteria, generate an automated email message to the student and a parent describing the concerning behavior and explaining what resources are available from University mental health and wellness services to assist the student. Your email must mention the student by "name" (just use the numeric identifier), and must also use the student's preferred pronouns as specified in the `ug_database.csv` file.

As you complete the task, you grow concerned about whether the project is reasonable. In your `readme.txt` file, list and explain at least two objections to this project. One of your objections must be technical (e.g., with respect to data quality), but the other(s) may be technical, legal, or moral.

Rubric

1. (8pts):
   - Data preprocessing (2pts).
   - Decision tree learning (2pts).
   - Branford correct answer (1pt).
   - Silliman correct answer (1pt).
   - Written explanation of decision and tree structure (2pts).
2. (12pts):
   - List of students who skip meals (1pt) and explanation how they were found (1pt).

– List of students who skip classes (1pt) and explanation how they were found (1pt).
– List of students who are out late (1pt) and explanation how they were found (1pt).
– List of students who stay in (1pt) and explanation how they were found (1pt).
– Email generation for students who fulfill all criteria (2pts).
– Written objections to your manager's project (2pt).