Patrick Yeh (psy2107)
Sameer Saxena (ss6167)

Project 1 README

Submitted files:

- proj3.py
- get_dataset.py
- squirrels.csv
- INTEGRATED-DATASET.csv
- example-run.txt
- README
- example-run_without_position_time.txt
- INTEGRATED-DATASET_without_position_time.csv

How to run program:

1. Make sure all Python files are in the directory
2. If you need to generate INTEGRATED-DATASET.csv, make sure squirrels.csv is downloaded and run '**python3 get_dataset.py**'
3. Run '**python3 proj3.py INTEGRATED-DATASET.csv [support] [confidence]**'

Dependencies:

The program uses the following libraries: sys, itertools, csv.

Description of dataset:

I chose to use the 2018 Central Park Squirrels dataset from the NYC Open Data site. In particular, I filtered out entries where the fields for 'Age' or 'Primary Fur Color' were left blank. (This preprocessing step was to simplify some of the cleaning.) I then transformed the downloaded dataset into a market basket interpretation of each tuple using the get_dataset.py file. This created an INTEGRATED-DATASET.csv that for each row had a tuple and the attributes that the tuple had. For example, if 'chasing' was set to true, then the corresponding entry in the CSV would have 'chasing' appended to it. After some trials, I also decided to exclude the Primary Fur Color and Age attributes since a super-majority (i.e. 80-90%) of the tuples shared a single value for each (i.e. 'Gray' or 'Adult'). As such, these attributes weren't so useful, and they were discarded so that they wouldn't

dominate the results. Additionally, I transformed the hectare field into a North/South and East/West categorization (taking the median values of the ranges 1-42 and A-I in the N-S, E-W arrangement) in hopes that location data would be useful attributes to examine, and then concatenated them together to get a quadrant of Central Park (i.e. North East, South West, etc.). Moreover, for the output_without_position_time.txt transcript file (which isolates out the position/time attributes to purely focus on the relationships between behaviors), I added a line in the get_dataset.py file which can be commented in/out in order to create this INTEGRATED-DATASET.csv, which does result in other interesting results. For what made the Squirrels census so compelling to me, there are both personal and impersonal reasons. Primarily, because I am a NYC resident living close to Central Park, and because I am fond of squirrels, I was interested in looking at this dataset. In particular, could I gain any information about squirrels and their behavior based on location or time? Perhaps for an avid squirrel enthusiast, they might find joy in finding certain patterns in behavior such as 'quokking,' a term I recently learned about thanks to this project. They could also find pleasure in learning where and when certain types of squirrels frequent. Moreover, perhaps this approach to squirrel census data can be extended to other environmental projects, so there is a sense of civic duty perhaps in choosing this dataset. It is quite fitting, after all, that Earth Day is the due date. Nevertheless, I believe that my choice of dataset has a compelling motive, out of respect for both the environment and the authors' wishes to learn more about squirrels.

Internal Design Overview:

The A Priori algorithm lies at the heart of our project. In particular, much of the design and implementation is based on the descriptions and optimizations discussed in both lecture as well as the original paper (see sources used). The high level flow is quite straightforward. We read and process the CSV, and after initializing the individual market baskets and screening out ones that do not satisfy the minimum support, we run the A Priori algorithm. In essence, we iteratively expand on the size of our market baskets, growing at each iteration by 1 and using libraries such as 'itertools' to go through each possible set we

wish to examine. We then make a decision on whether to include a constructed set in our next list or not (i.e. checking if it satisfies the minimum thresholds specified in the initial command line call) by using the a priori assumption to calculate our new metrics. Once we have ended the iterative process and found our sets, we then create our rules/implications by once again iterating through the sets and isolating an item/attribute to be the right hand side of the implication. We examine if the confidence of that implication is satisfactory, and if so we will then map the string form of the implication to its corresponding metrics. Finally, once all the rules have been figured out, we will use these mappings, sort them by decreasing order of confidence, and then print out/output the results to the user.

Example run:
'python3 proj3.py INTEGRATED-DATASET.csv 0.01 0.7'

From the results, in the example-run.txt transcript, we see that squirrels are primarily recorded away from the South West region. Perhaps this corresponds to the true distribution of squirrels in the area (i.e. there might be less trees or a lack of sheltered areas near there), or perhaps there is some sampling bias (i.e. people were less likely to frequent those areas, for whatever reason, and thus those observations reflect this tendency). Moreover, in terms of behavior, 'indifferent' or 'foraging' squirrels seem quite common. This coheres with our general understanding of squirrels. In most encounters, they do seem to be indifferent and searching for food. On the other hand, 'approaching' was a field that is notably not of high enough support, which strongly correlates with our understanding that squirrels are averse to interactions with humans (to the dismay of those who would like to see 'friendly' squirrels). In terms of implications, we find that in terms of finding 'friendly' squirrels that approach the observers, the findings may show some promise. In particular, the rules [Approaches,Eating,North West] => [PM] and [Approaches,Indifferent] => [Foraging] hint that it is better to do so in the evening and while the squirrel is foraging for food. In particular, the foraging aspect makes sense since the approaching behavior of squirrels occurs when they are seeking

food from humans. On the other hand, a squirrel will make kuks (a specific kind of noise) if they feel alarmed or excited. From the extracted rules, [Kuks,Runs from,South East] => [Climbing], [Kuks,Tail twitches] => [Foraging], and [Kuks,North West] => [AM] hint that squirrels may feel less at ease in the morning while foraging or climbing. Thankfully, although somewhat surprisingly, none of the rules pertained to quaas or moans, other noises used by squirrels to indicate the presence of predators. Perhaps this is due to the fact that Central Park may be a relatively safe place for squirrels, without many natural predators. Finally, I also ran A Priori (using the same support and confidence of 0.01 and 0.7) over an integrated dataset that excluded position and time information (in order to isolate purely behaviors that relate with each other). To repeat this process and generate the correct INTEGRATED-DATASET.csv file, one needs to go into the get_dataset.py file and comment out the line corresponding to position/time (following the comments). The information is stored in the output_without_position_time.txt file, and I got the following rules for tail twitching: [Approaches,Chasing,Running] => [Tail twitches], [Chasing,Foraging,Kuks] => [Tail twitches], and [Approaches,Foraging,Running] => [Tail twitches]. Tail twitches are a sign of squirrel curiosity, so it makes sense that squirrels that are approaching humans show signs of curiosity. Additionally, foraging and running both seem to correspond with tail twitching, which could perhaps indicate behaviors where squirrels are more observant or curious of their surroundings. Likewise, kuks corresponds with our understanding of the behavior in that squirrels may be more curious when excited or alarmed.

Sources used:

1. Class materials
2. Section 2.1 of Agrawal and Srikant, VLDB 1994