Plan for holiday--me

1. Mining useful information from the data set
2. Observe the given dataset
3. Summary the information like how many people a person met in a place, whether the direction can be predicted from the given dataset, how many people he meets in one place, how long he stays in one place, and use some statistical method to represent the distribution of the time period of people stays in one place.
4. Identify techniques can be applied to this dataset

Identify techniques in “trajectory data mining: an overview” that can be used on this dataset, determine whether each technique is appropriate for this data set.

1. Identify methods analyzing the results step 2 gives

Read more papers and find methods to analyze the information identified from dataset.

Plan for holiday--Nic

-- write a description of the dataset; what do each of the columns represent; if there are any you don't understand, ask me for more info.

-- get the dataset loaded into a suitable environment (eg R, Python, etc).

-- create a github/bitbucket repository to store your scripts for processing the data and performing analyses. However, please do \*not\* store the data in the repository, just keep that on your local computer.

-- write a list of possible things you can measure about this data set.

-- choose one of these and work out how to calculate it

-- think about how to visualise the result of your analysis -- what sorts of plots, tables, etc can you produce; what is the best way to communicate the information you have obtained.

eg, for 2:

-- write out a list of techniques, and a brief description

-- for each technique, identify the requirements that apply to the data (eg, scale, resolution, etc). Determine if any techniques will be obviously invalid (nb: you may not have enough info about the data at this point, we will probably need to discuss this).

-- for any techniques that look like they are likely to be useful, these are the ones to start reading further about.

Summary of the discussion at week 1

1. Figure out patterns based on the small dataset.

Like: what movement patterns do people with different age or gender have in one day and how long they stay in one place. I just got an idea after meeting. Maybe I can ignore the latitudes and longitude data at first, rather mark a place as home, work or café or something like that. And then write the sequence like “home(2h)-work(16h)-café(1h)-home(5h)”. And do machine learning on this data set. And draw conclusion like people who are at age from 12-20 get pattern like “home(2h)-work(16h)-café(1h)-home(5h)”(just an example). Am I in the right track?

Before trying to identify techniques to work this out, before next meeting, I’ll try whether I can draw points on a map to roughly observe patterns or by do handwriting.

1. Use what we figure out based on the small data set to analyze the bigger population.

**Report for week 1**

I cleaned the data set using Python, removed the content included in parentheses (attached: day\_in\_the\_life\_drop\_suffix.csv), and merged all the locations a person stays in one day to a single sequence, and only remained the column Age and Gender (attached: day\_in\_the\_life\_after\_processing.csv). There are only 83 pieces data after processing. Without considering about when a person is arriving to a location and how long it stays, I put the data set into Weka and do supervised machine learning: Naïve Bayes, decision tree and random forest with cross-validation. Probably due to the small amount of data, all the algorithms don’t give good results.

A picture containing text

Description automatically generated

Figure 1: Result for Naïve Bayes

A screenshot of a cell phone

Description automatically generated

Figure 2: Result for decision tree

A screenshot of a cell phone

Description automatically generated

Figure 3: Result for random forest

After doing supervised machine learning, I also try two unsupervised machine learning algorithms: cluster (result is in attached file: result\_for\_cluster) and perceptron. As imagined, the result is also not satisfied. Because these kinds of algorithms need huge amount of data to train.

A screenshot of a cell phone

Description automatically generated

Figure 4: Result for Multilayer perceptron

Then, I change every age to age interval (day\_in\_the\_life\_drop\_suffix\_age\_processed.csv). And I was trying to extract some information about trajectory patterns with different age and gender by putting the data set into R and draw a graphic (Figure 1 and Figure 2). X axis represents different trajectory patterns while Y axis is the count of how many persons have a particular pattern with different age interval. As you can see, there are so many different movement patterns. For female, pattern ”Home-Private Transport- Private Transport-other- Private Transport-other” have the most frequency. There are two others in this pattern. We don’t know information about this “other”, because people are unwilling to tell, maybe this is the reason why this pattern is with the highest frequency.

A screenshot of a cell phone

Description automatically generated

Figure 1: Trajectory pattern for female

A picture containing clock, microwave

Description automatically generated

Figure 2: Trajectory pattern for male

All of the code and documents and result file attached here will also be uploaded to Github.

<https://github.com/liyi19950329/Research-Project.git>

**Report for week 2**

This week, I applied Neural Network (NN) on this data set. Firstly, I use one-hidden-layer NN to train. For convenience to calculate, I use number to represent every pattern (label here) from 0-65 (which I think it is not correct, I will improve later). The parameters are: epoch is 1000; loss function is A close up of a logo

Description automatically generated ; activate function is sigmoid function: A close up of a clock

Description automatically generated; two features are age and gender (Figure 1). The result is somehow weird. The value of MSE are very big every epoch, and stops at 1122.458 after epoch 60 and no change. After analyzing, maybe it is not reasonable to use 0-65 to represent different patterns, and the method I use is often used to deal with binary classification problem, so it is not suitable to apply on multi-class classification problem.

A picture containing computer, player

Description automatically generated

Figure 1:one-hidden-layer NN

Improvement

1. To solve the problem that the result of loss function is large, I do data normalization on feature age and label, that is, changing the age to a decimal between (0,1), as well as label. In this way, results of loss function every epoch is like Figure 2

A picture containing keyboard, computer, man

Description automatically generatedA picture containing text, keyboard, computer

Description automatically generatedA picture containing keyboard, computer

Description automatically generated

Figure 2: value of MSE every epoch

But there is still a problem. After training the model, I applied this model on three instances to predict their label. The results are all 0.485 (Figure 3).

A close up of a sign

Description automatically generated

Figure 3: results of three instances

1. Apply multi-class classification NN on this data set.

I’m searching on the Google, and trying to figure out how it works. I only pick two features to make prediction, but there are 65 classes in total. The number of features are too small to applied multi-class classification NN as well as SVM.

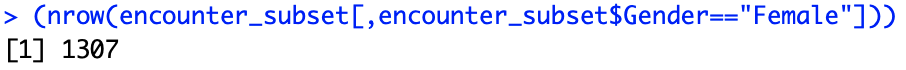
Nic’s suggestions

1) Data Exploration:

ML algorithms are tools for answering questions about data. Before we can use them effectively, we need to know what our question is, and we need to have a good understanding of our data.

I suggest spending a little bit more time first just exploring the data. It's also good to be systematic about this: for each of the fields, what is the distribution of values? For example:

How many unique subjects in this dataset?



How many males / females are in the data set?

A picture containing bird

Description automatically generated

What is the age distribution?

A screenshot of a cell phone

Description automatically generated

As we can see in this picture, most participants are at age around 55-65.

What is the distribution of the number of locations visited by each study participant?

A screenshot of a cell phone

Description automatically generated

To get the distribution of the number of places that a participant have been to, I calculate the number of different locationID with same subject number. Because on locationID represents a unique location. Most of the participants have been to 10 places. Very small proportion participants go to more than 20 locations.

What is the distribution of number of different \*types\* of locations visited by each study participant?

A screenshot of a cell phone

Description automatically generated

3 is of the largest proportion, which means, most participants are at 3 different locations in a day

What is the distribution of times spent at locations? at different location types?

A screenshot of a cell phone

Description automatically generated

THEN you can look at interactions:

How does the number of locations visited vary by age? by sex?

etc.

(Note that a lot of this is reported in the Rolls et al 2015 paper; you should still do this yourself though, and confirm that you obtain the same results!)

2) Characterising trajectories

I think you are onto the right idea in looking at sequences of location types. Though, rather than just concatenating the location types together, it might be more manageable to first convert them to a one or two letter abbreviation, so you end up with a more manageable string.

However, as you note, there are currently too many different types of sequence. Is there a sensible way you could think to reduce this? For example, perhaps you could strip out the "transport" locations, since these are only about movement between locations? Or you could only include locations where the time spent at that location was above some minimum duration? Or something else - I'm sure there are other possibilities.

Perhaps you could even think (at this stage) about just considering the location types as a set, so each individual would be characterised by the set of location types they visited (ignoring the order) This is obviously no longer a "trajectory" but might be a useful starting point.

3) Clustering trajectories

I think your ML approaches are treating each of the ~60 sequences as a unique dependent variable (or outcome), which you are then trying to predict on the basis of independent variables such as age and sex.

Before we go this far, I am interested to know how we might cluster individuals based on their trajectories. Can we identify subsets of trajectories (however we have ended up defining them) that are more similar to one another than they are to trajectories from a different subset?

I've attached a paper that did a similar thing for a much smaller data set (and again, not considering trajectories). Have a look in particular at Figure 4 (and the text around it) where they break the data set up into clusters based on the amount of time that people spend at particular locations. What techniques can you think of learning these clusters (in an unsupervised fashion)?

It's also worth reading about how this paper defines "person contact hours" (ie, amount of time a person spends making contact with other people, in different types of locations). With the data set you currently have, you could calculate the amount of time a person spends in different type of locations. Once you have the social contact information to add in too, you could extend this to PCH.

Again though, this is throwing away the trajectory information, which I think is worth keeping for your project.

4) Other random questions

Given all of these trajectories, can you think of a way that, given N location types, you could predict what location type N+1 is going to be?

(can you think of others?)

5) Data

I've attached the full data set for the trajectory data. However, please don't just throw this into an ML algorithm hoping that more data will improve prediction. We need to understand what it is we're trying to predict better first!

It still doesn't include social contact data, but I want to focus on one thing at a time first.

Hi Yi,

Thank you - yes, feeling much better after a couple of days.

I've had a look through your reports, and it's great that you've started to work with the data; however, I'm going to suggest you pull back a bit and answer some of the more basic questions we discussed before throwing it into machine learning algorithms.

1) Data Exploration:

ML algorithms are tools for answering questions about data. Before we can use them effectively, we need to know what our question is, and we need to have a good understanding of our data.

I suggest spending a little bit more time first just exploring the data. It's also good to be systematic about this: for each of the fields, what is the distribution of values? For example:

How many males / females are in the data set?

What is the age distribution?

What is the distribution of the number of locations visited by each study participant?

What is the distribution of number of different \*types\* of locations visited by each study participant?

What is the distribution of times spent at locations? at different location types?

THEN you can look at interactions:

How does the number of locations visited vary by age? by sex?

etc.

(Note that a lot of this is reported in the Rolls et al 2015 paper; you should still do this yourself though, and confirm that you obtain the same results!)

2) Characterising trajectories

I think you are onto the right idea in looking at sequences of location types. Though, rather than just concatenating the location types together, it might be more manageable to first convert them to a one or two letter abbreviation, so you end up with a more manageable string.

However, as you note, there are currently too many different types of sequence. Is there a sensible way you could think to reduce this? For example, perhaps you could strip out the "transport" locations, since these are only about movement between locations? Or you could only include locations where the time spent at that location was above some minimum duration? Or something else - I'm sure there are other possibilities.

Perhaps you could even think (at this stage) about just considering the location types as a set, so each individual would be characterised by the set of location types they visited (ignoring the order) This is obviously no longer a "trajectory" but might be a useful starting point.

3) Clustering trajectories

I think your ML approaches are treating each of the ~60 sequences as a unique dependent variable (or outcome), which you are then trying to predict on the basis of independent variables such as age and sex.

Before we go this far, I am interested to know how we might cluster individuals based on their trajectories. Can we identify subsets of trajectories (however we have ended up defining them) that are more similar to one another than they are to trajectories from a different subset?

I've attached a paper that did a similar thing for a much smaller data set (and again, not considering trajectories). Have a look in particular at Figure 4 (and the text around it) where they break the data set up into clusters based on the amount of time that people spend at particular locations. What techniques can you think of learning these clusters (in an unsupervised fashion)?

It's also worth reading about how this paper defines "person contact hours" (ie, amount of time a person spends making contact with other people, in different types of locations). With the data set you currently have, you could calculate the amount of time a person spends in different type of locations. Once you have the social contact information to add in too, you could extend this to PCH.

Again though, this is throwing away the trajectory information, which I think is worth keeping for your project.

4) Other random questions

Given all of these trajectories, can you think of a way that, given N location types, you could predict what location type N+1 is going to be?

(can you think of others?)

5) Data

I've attached the full data set for the trajectory data. However, please don't just throw this into an ML algorithm hoping that more data will improve prediction. We need to understand what it is we're trying to predict better first!

It still doesn't include social contact data, but I want to focus on one thing at a time first.