Machine Learning Engineer Nanodegree

Capstone Proposal

Jesse Gallegos
January 26, 2017

Proposal

Domain Background

This project investigates a Speed Dating dataset from Kaggle. The dataset was compiled by two professors from Columbia University's Business School. Information including demographics and preferences of attraction were collected from participants during weeknights in the years 2002 through 2004. Moreover, the study aimed at reproducing a business model that Speed Dating services use and chose similar environments to hold these events (e.g. popular bars near universities).

During the events, daters had a four minute "first date" with every other participant of the opposite sex. At the end of each four minute "date", participants were asked if he, or she, would date that person again. Participants were also asked to rate his, or her, date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

In addition, the dataset includes a questionnaire that gathered data from participants at different points in the process. These fields include: demographics, dating habits, self-perception. Other data include what features of attraction participants think people find valuable in a mate. These preferences of attraction were asked at the start of the event, half way through the event, a day after the event and three weeks after the event.

Finally, the dataset yielded a paper published by the same professors mentioned above, that is, Ray Fisman and Sheena Iyengar (Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment). The rest of project will explore the mentioned dataset.

Problem Statement

One finding to verify with the papers written by the Columbia professors is that men value physical attractiveness of a partner more than women do. Another is that women place a greater weight on intelligence than men do. Finally, this project will try to do two things:

1) The first task at hand is to investigate the types of clusters that males and females fall into, respectively. These clusters will be based off interests that males and females state in the questionnaire at the start of the event. The goal is to see if, for instance, taking one male from a night clubbing cluster, pertaining to men, matches with a female from the night clubbing cluster pertaining to women.
2) The other task is to generalize a model to predict if one male and one female match with each other.

The metric to evaluate the model will be f1score. This metric measures how the model makes a correct decision. f1score also takes into consideration the number of times that one male and one female do not match. f1score, hence, is a good choice as it is a harmonic mean of the ratio of people who match and do not match.

Datasets and Inputs

The dataset file, Speed Dating Data, has 8378 data points with 195 different features. There were 551 people involved in the dating event with the population being divided into 274 women and 277 men. For the sake of brevity, the reader is invited to consult the Data Key file included with the dataset. The Key file gives the exact number of participants for each of the 21 'wave' events. Moreover, the Data Key file gives a thorough overview for what the allowed values for each field are.

The first attributes that will be explored at the start of the project include 'iid', 'gender', 'wave', 'attr', 'intel', 'fun', 'amb', and 'shar'. 'iid' uniquely identifies each participant in the event and is useful for querying unique members. 'gender' helps to query by gender. 'wave' number also helps in querying in what wave an individual participated. The features of interest are: 'attr', physical attractiveness; 'intel', intelligence; 'fun', how fun a person is; 'amb', ambition; and 'shar', shared interests. Each of these features of attractiveness are rated in one of two ways:

1) ratings based off a scale ranging from 1 to 10.
2) daters are asked to distribute 100pts to the six features of attraction in the manner he or she feels is most important.

Other things to note: features of attraction with extensions "1_1" relate to the question asked at some point during the event; the first number pertains to the question number and the second number refers to the time of the event (e.g. 'attr1_1' is question 1 asked about 'attr' at the start, '1', of the event). Likewise "1_s" refers to question '1' asked halfway, 's', through the event. "1_2" refers to question '1' asked the day after, '2', the event. "1_3" refers to question '1' asked three weeks, '3', after the event. These extensions are important to see, if after various dates, daters change their outlook on stated preferences of attractions with respect to the beginning, middle, day after, or three weeks after the dating event.

Solution Statement

To verify that men value physical attraction, I will be taking a subset of the original dataset. The subset in mind identifies each person uniquely by 'iid'. Based off this subset, my approach will be to make two more subsets: one for males and another for females. At this point, data will need to cleaned in the sense that missing data points will need to be filled, dropped, or corrected if not within the Data Key file's allowed range for each feature. Once cleaning is done, I will normalize the features of attraction being investigated. The features of attraction will be divided into four categories:

1. how individuals rated relevant features of attraction at the start of the event.
2. how individuals rated relevant features of attraction halfway through the event.
3. how individuals rated relevant features of attraction a day after the event.
4. how individuals rated relevant features of attraction three weeks after the event.

Afterwards, a histogram will be created to investigate how preferences of attractions change over the course of the speed dating event. Basic statistics, or frequency charts, will aid in deciding whether there is an overwhelming preference for a feature.

The clustering aspect of the solution will involve using the mentioned subsets formed above. I will be using K-means clustering with a Euclidean metric to determine distances between points. Moreover, my solution will use several values for the number of clusters to find how many centroids are optimal. Afterwards, I will develop a model to determine matches between clusters and use f1score to evaluate error.

Benchmark Model

The benchmark model that my project will be compared against is a compilation of two linear models. Fisman and Iyengar proposed several models to determine decisions of participants based off what daters believed important. In fact, looking through the attached article. "Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment," shows that the two authors created new features by using old features and defined a decision variable as a weighed mean of new features. The formulae that will be investigated from the author's papers are:

1) the Decision function dependent on attractiveness, ambition, and intelligence.
2) the Decision function dependent on new features formed from ratings received from a dater's partner and self rating of the dater.

Finally, the author summarized results for both these Decision function in tables with various statistics: AVG, STD, R2, and so on. The assumption that the authors made in their paper is that people when dating have "straightforward behavior," meaning that people date people he or she likes more.

Evaluation Metrics

The evaluation metric that is useful in this project is Mean Absolute Error (MAE). MAE will measure the error between what I am doing in this project against something the published results by Fisman and Iyengar. This evaluation metric will compare the total sum of decisions against the number of decisions of 'yes' decisions the professors predicted. What I can also do is compare the basic statistics with the same metric. Finally, the two professors have their results standardized and making cross comparisons is feasible.

Project Design

As mentioned in the solution section, this project will start off by taking subsets of the dataset and perform data clean up. A quick look at the data set reveals that some of the data does not fall within the specified ranges laid out in the Key file. An example includes entry fields based on the Likert scale from 1 to 10. Some fields have violated these ranges and need to be fixed to conform to the range. Numbers below 1 will be raised to 1; number greater than 10 will be lowered to 10.

The dataset had other issues. The instance in mind is where daters were asked to distribute 100 points across each features of attraction in question. Some respondents accidentally distributed more than 100 points across the six features of attraction. This is

no problem because these distributions are relative to how the these points were assigned. Normalizing these responses will rescale everything to lie between 0 and 1. Once this subset is cleaned, I will use univariate analysis to see how features of attraction evolve over time.

The second task at hand will be to create a model that will learn from the dataset and attempt to make matches. What I want to explore in this section is if people clustered by sex match with people in the same cluster for the opposite sex. An example would be a male who lands in the 'clubbing' cluster matches with a female from the 'clubbing' cluster. A scatter matrix plot will help to identify correlation between features. This will help to reduce the number of dimensions and, perhaps, reduce the problem so that the clustering is visually interpretable. The clustering algorithm of choice will be K-means and the number of clusters will be chosen with respect to the silhouette score.

Constructing the matching algorithm will require me to choose what features to include. Choosing the features for this algorithm will depend on the structure of the mentioned clustering above. The dataset includes ratings that dates give each other and will be incorporated in the algorithm. At the moment, I will need to start the project and allow creativity to dictate the direction of the project.