Project Proposal

CSCI 374

Jon Carl

November 10, 2023

1. Background In the age of unlimited access to information, learning about new ideas has become increasingly accessible. Yet, an increase in the accessibility of information has also led to a growing difficulty to make decisions based upon that information. A prime example of this problem is choosing a restaurant to dine at. Stating the problem a bit more formally: given a collection of preferences for n individuals, we want to choose the restaurant that maximizes satisfaction.

2. Hypotheses and Goals The goal of this project is twofold. First, I intend to provide suitable restaurant recommendations to any individual whose preferences provide some mechanism for matching their own preferences to the features of a restaurant. Second, I hope to produce group predictions that maximize the joint rankings of restaurants from each individual in a group.

Using machine learning to investigate this problem will be useful in two ways. Firstly, creating predictions for restaurant rankings is helpful for individuals who struggle to make decisions about what restaurant they want to eat at. In a sense, the machine learning model takes away decision fatigue some individuals may face. More importantly, though, this kind of problem generalizes to a broader problem. In particular, this project models a utility maximization problem where a set of preferences for n individuals yields some utility depending on their preferences. Moreover, the project produces an interesting philosophical decision to be made when making predictions for a group of people. On one hand, some may argue that choosing the welfare-maximizing utility set of preferences for a group is the correct objective function. However, others may argue that choosing an objective function which maximizes total utility yields the best optimum. In this way, thinking about the trade-offs of different objective functions will be challenging, yet rewarding.

3. Sources of Data The data are obtained via UCI Machine Learning Repository.

4. Sources of Complication One source of complication I foresee is that the data are not IID. In particular, the observations in the data that stem from the same consumer are not independent. Since the preferences of a consumer likely impacts *all* of their rankings, we cannot assume that each observation in the dataset is IID. In the analysis, I'll need to ensure that the model assumes independence across batches of data where a single batch corresponds to a single user. We can assume that each batch is IID though, assuming that a single consumer's preferences has no correlation with any other consumer's preferences.

5. Rough plan of Analysis I plan on using using random forests and clustering algorithms for predicting restaurants rankings. Seeing that a set of preferences may yield similar outcomes (in terms of restaurants ranked), clustering seems like a good option for finding similarities among restaurants who share features that align with an individual's preferences. Moreover, seeing that random forests handle multi-class classification tasks, predicting a ranking (where each ranking is its own "class") for a consumer seems like another good option for predicting an individual's ranking for restaurants.