Deep Learning Models for Restaurant Choice

CS 230 Final Report

Evan Munro

December 16, 2018

Introduction

The application of deep learning in the literature on economic choice is still in its infancy. This is largely due to two reasons: First, datasets on personal choice have historically been difficult to collect and limited in size, which makes training a resilient neural network challenging. Second, economists are reluctant to move away from conditional logit based methods with interpretable parameter estimates.

I address the first barrier by using a novel dataset from the Athey Lab, where I am a PhD student, which has a $>$ 10GB dataset on restaurant choice derived from mobile location data. This dataset is large and rich enough to train a deep neural network properly. The primary goal of the project is to test how the prediction of more complex deep learning approaches compares to the conditional logit model that economists are accustomed to using. A secondary goal is to compare neural network approaches to state of the art Bayesian matrix factorization approaches.

There were a variety of challenges addressed in this project. The data is large and very sparse. There are thousands of restaurants in the dataset, but each user appears infrequently and each restaurant is chosen infrequently. There is also no standard for neural networks in the literature on consumer choice. So, I construct a variety of architectures to explore which is the most promising area for future development.

I have constructed as simple single layer neural network that estimates conditional logit as a baseline, which achieves a training and test accuracy of approximately 3.3%. This is not bad, considering users choose between over 4,000 restaurants for each lunch session. I extend this simple baseline model in three different ways: 1) adding non-linear features to capture complex interactions within the features of each restaurant 2) adding encoder layers to add user and restaurant specific embeddings and their interactions to the feature space 3) adding dependency on nearby restaurants using an RNN framework to capture more complex interactions between available restaurants. The best resulting model achieves just under 10% accuracy, which is nearly a 3-fold improvement from the baseline model.

Literature Review

The data for the project is the same data that was used in [1], who estimate a probabilistic graphical model and use the parameter estimates from the model to explain consumer preference for travel time and to investigate effects of restaurant closings and openings.

There is a small literature on neural networks and discrete choice modeling. [4] recognizes the link between multinomial logit and neural networks and builds a single layer neural network that adds density compared to the multinomial logit single layer network. They find an increase in the log-likelihood for a Swissmetro travel choice survey dataset. [3] generalizes multinomial logit in a restricted boltzmann machine framework. [2] uses a RNN framework to predict consumer choice of airline itineraries; the very wide choice set of airline itineraries is comparable to the restaurant choice setting.

I expect that companies doing discrete choice prediction for advertising and recommendation systems may also have developed related deep learning systems for similar problems, but much of that information may be private.

Dataset

The dataset is a 10GB tab-separated file that contains data for 78,524 choices sessions from 8,552 users choosing between over 4,921 restaurants for lunch in the Bay Area. The time period covered is January 2, 2017 to October 10, 2017. The features in the dataset include unique identifiers for

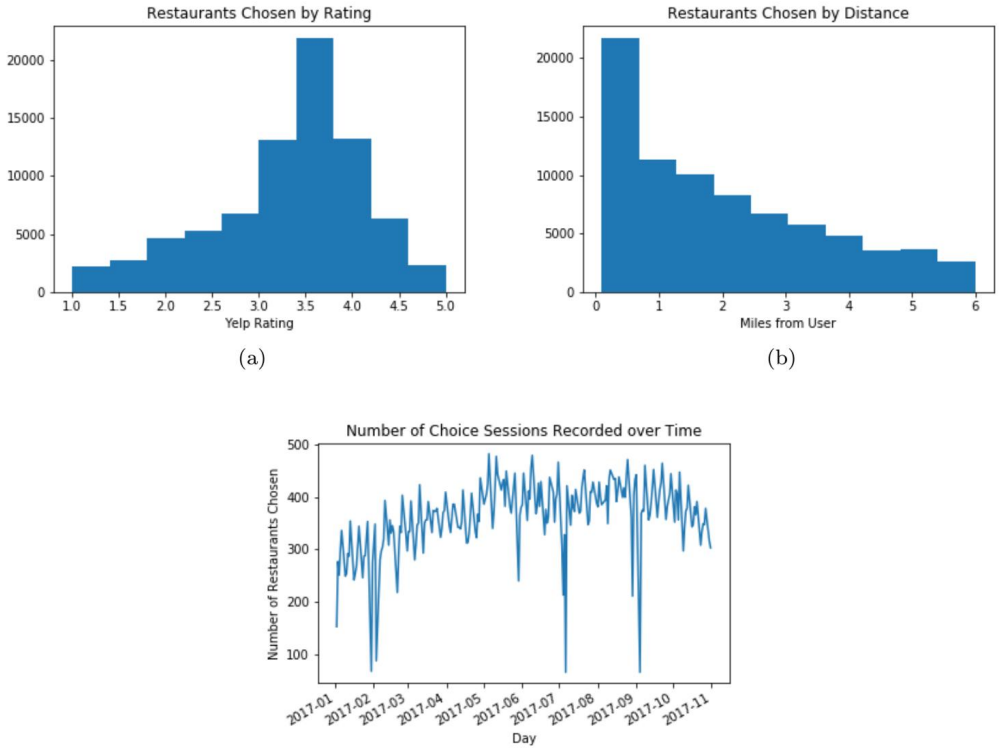


Figure 1: Plots of Chosen Restaurants

each user and restaurants, the distance of the consumer to every restaurant within 6 miles of them before lunch, as well as Yelp data on each restaurants about the average rating, price range and the type of food served. The outcome variable is which restaurant out of the restaurants within 6 miles of them that the consumer chose for lunch. The choice sessions are derived from users' anonymized location data. In Figure 1, I show the histogram of choices by distance and ratings; consumers prefer closer restaurants, and restaurants with higher ratings, as expected. I also include a plot of the number of choices recorded per day over the period, which fluctuates around a reasonably constant value, apart from noticeable downward spikes on Labor day and Fourth of July, when many restaurants are closed.

I have ordered the dataset by date. I assigned the first 95% of choices sessions to the training set, the next 5% to the validation set, and the final 5% to the test set. I have transformed the distance to log(distance), so the coefficient on the distance term in the logit model can be interpreted as an elasticity.

There a variety of challenges with this dataset. Each consumer theoretically can choose any of the 4,921 restaurants in the Bay Area for lunch, but the dataset only includes the distances of restaurants within 6 miles of them. Neural network approaches expect a fixed input size for non-temporal data, so I have chosen to use a masking approach. I input a fixed $4,921 \times n_x$ input of restaurant choices for each choice session, where only those restaurants that were within 6 miles of the user contain non-zero data.

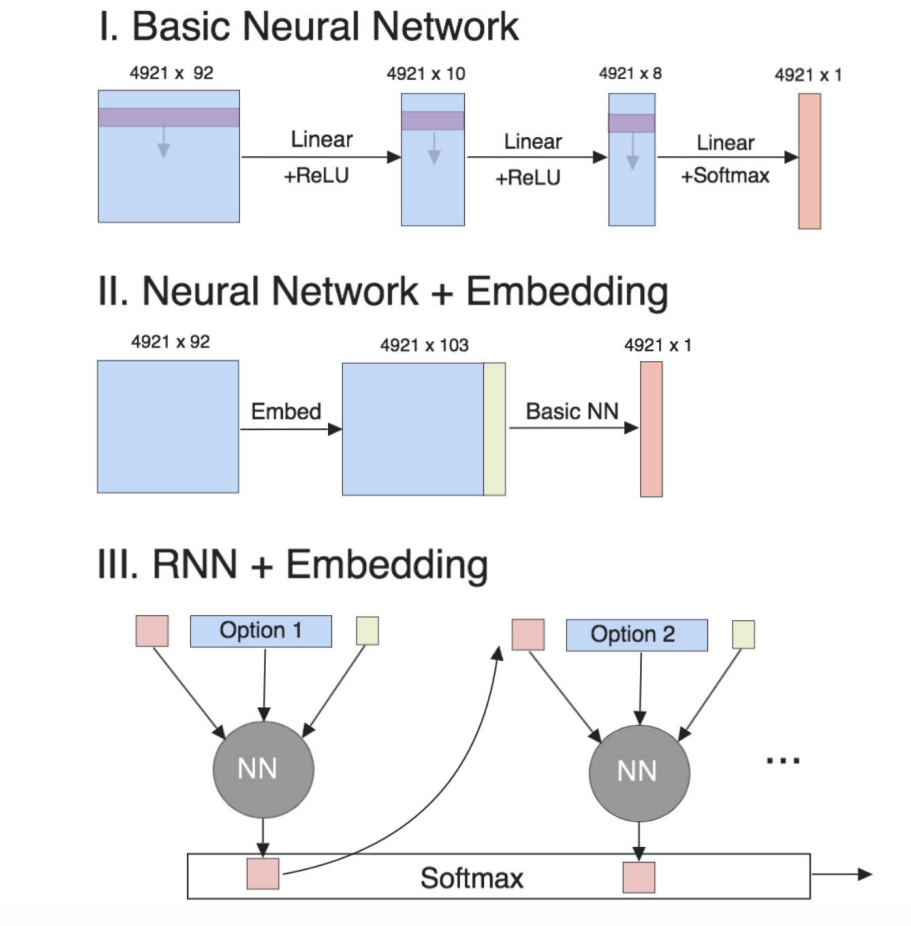
I predict the probability that a consumer i chooses restaurant j in choice session t, $Pr(Y_{it} = j|X_{it})$ as a function of the restaurants' characteristics, distance from the user, as well as other nearby restaurants' characteristics and distances.

Models and Estimation

The architecture of the three non-baseline models that I estimate on the restaurant data for this project are in Figure 2. I have estimated the models in PyTorch on the Stanford Sherlock Computing Cluster either on CPU or GPU. I have written all the code for this project myself. In this section, $s = 1, \dots S$ are the S choice sessions, $i = 1, \dots, N$ are the N users, and $j = 1, \dots, J$ are the $4,921$ restaurants. The first three models have a $4,921 \times n_x$ input layer, and all models have a final softmax layer that ouputs the probability of choosing each of the $J = 4,921$ restaurants.

$n_x = 92$ for each model and includes features such as log distance from user to the restaurant, the type of restaurant, its rating and its price. To update the parameters of each model, I use momentum gradient descent with $\beta = 0.9$. Within each class of model, I used the dev set to select the hyperparameters of the model.

Figure 2: Neural Network Architectures



Baseline Model: Conditional Logit

A conditional logit model is a one layer neural network, with softmax activation and cross entropy loss. The network is constructed with a $J \times n_x$ input, a single $n_x \times 1$ linear layer with softmax activation, and a $J \times 1$ output, when there are J choices and n_x features for each choice.

Adding Non-Linearities: Neural Network

For the first neural network model, I construct a shallow neural network. I add two hidden layers with ReLU activation before the output + softmax layer. The first takes as input the n_x input features and outputs 10 activations using 10 linear+reLU units, the next contains 8 hidden units, and the last maps the 8 outputs from the previous layer to an output of size 1. Softmax activation normalizes the scores for each restaurant to determine which has the highest predicted probability.

Adding Restaurant and User Embeddings

For the second neural network model, I augmented the provided features with a 5-dimensional user and 5-dimensional restaurant feature vector. I also included a feature which was the dot product of the user and restaurant feature vector. This allows the neural network to estimate user-specific, and restaurant-specific unobserved attributes that influence the probability of a user choosing a restaurant. The estimation of the embeddings for users and restaurants is done in the first layer of the neural network. Embedding layers are green in the Figure 2.

Model	Layers	Train Loss	Train Accuracy	Test Loss	Test Accuracy
Cond. Logit	1	5.67	3.10%	5.63	3.30%
NN	3	5.42	4.70%	5.40	4.75%
NN + Embed	4	4.99	8.94%	5.04	8.95%
RNN + Embed	4	6.18	3.55%	6.20	3.60%

Table 1: Model Performance Summary

Adding Interactions Between Restaurants

The previous three models allow increasingly complex functional forms for the probability that a user selects a restaurant, given the user's and the restaurant's characteristics. However, the characteristics of other restaurants impact the probability of selecting a restaurant only in the final softmax layer. To add interactions between restaurants, I modify the input data so that each choice session is not fixed in size, so restaurants that are not within 6 miles of the user are not considered. In addition, I ordered each choice session by ascending distance to the user. The model assumes that the user considers each restaurant, starting with the closest restaurant. The activation from the previous restaurant is passed as an additional feature to a three layer neural network, along with restaurant embeddings, user embeddings and restaurant-user characteristics

This allows the characteristics of other restaurants in the choice set to influence the probability that the current restaurant is chosen in more complex ways than through the final-layer softmax normalization.

Results and Discussion

I have recorded train and test cross-entropy loss and accuracy for the baseline model and the four neural network models estimated. The conditional logit model has a test accuracy of 3.3%. It manages to find a relationship in the data that results in far better accuracy than random selection would. Furthermore, the coefficients that the conditional logit have signs and magnitudes that correspond to intuition; for example, having a missing Yelp rating results in a decrease in probability of selecting that restaurant, and a 1 percent increase in distance from a user results in a 1.2 percent decrease in the probability of selecting the correct restaurant. But it is limited by its strict functional form. The neural network model allows a more flexible form of mapping from restaurant features to probability of selection and improves the accuracy by 50%.

Only a few characteristics about restaurants and users are observed in the dataset. It is highly likely that there are systematic features of users and restaurants that are not observed in the dataset, but can be estimated to improve model performance. As a result, the embedding layer added to the basic neural network results in the best performance. This layer assumes there are 5-dimensional feature vectors for users and restaurants that are unobserved. These embedding vectors and their dot product are added to the feature matrix for each choice session. The latent variable approach combined with a neural network results in the best test accuracy of 9%.

The RNN model has the highest test and training loss of any model, although the test accuracy does beat the conditional logit model. More work is needed to use an RNN approach effectively for this problem.

For most models, test error is actually lower than training error. This is because the test set contains choices that are later in the time frame of the data than the training set. So, the model improves over time, even out of sample, due to repetition in choice behavior that is learned through the embeddings and other neural network parameters.

Discussion and Future Work

The target for this project was to beat 20% test accuracy, which is the state of the art test error on this sort of problem achieved by the Travel Time Factorization Model (TTFM) in [1]. The best neural network model only achieved about 50% of the accuracy of the TTFM.

One limit to performance of the models was computational speed. A significant time investment was required to come up with the masking approach for dealing with missing choices and to find model architectures that could capture some complexity in the choice process but had enough parameter sharing to be feasibly estimated. Due to the limited time available to train and tune the models, I trained each model for only four epochs. The input for each choice session contains over 400,000 integers, so training is slow, especially for the RNN, which doesn't take advantage of vectorization. The RNN took approximately 7 hours per epoch on the CPU and the most complex

neural network was about half of that. My GPU access on Sherlock was limited - in the future training on multiple GPUs would be required to quickly estimate some of the models considered.

The largest barrier to better performance was the limited number of times restaurants and users were observed. The best model, with restaurant and user embeddings, has over 60,000 parameters that need to be trained just in the embedding layer. Some restaurants and users were observed infrequently, meaning the updates for their feature vectors may have only happened once or twice in the training process. One way to address this challenge is using Bayesian methods. The TTFM is a Bayesian matrix factorization approach to predicting restaurant choice. Adding the right prior distribution to the embeddings would allow updates of restaurants and users seen frequently to also influence the updates of similar restaurants and users seen infrequently. This approach, combined with the flexible non-linear functional form of neural networks, has the potential to beat the TTFM model and will be pursued in a follow-up project.

Conclusion

Relying on conditional logit models for choice prediction in problems where there are large training datasets available results in a significant cost to out of sample accuracy. Though neural networks do not provide a direct estimate of elasticity with respect to distance or restaurant type, the approaches are several times more effective than the baseline for predicting choice, even in the high-dimensional setting of Bay Area restaurant prediction. A Bayesian neural network with embeddings is the most promising avenue for future research in consumer choice prediction.

A Code Repository Location

Link: https://github.com/evanmunro/restaurant_choice

References

[1] S. Athey, D. Blei, R. Donnelly, F. Ruiz, and T. Schmidt. Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data. *AEA Papers and Proceedings Vol 108*, pages 1–36, 2018.

[2] A. Mottini and R. Acuna-Agost. Deep Choice Model Using Pointer Networks for Airline Itinerary Prediction. *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[3] M. Otsuka and Takayuki Osogami. A Deep Choice Model. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 850–856, 2016.

[4] B. Sifringer, V. Lurkin, and A. Alahi. Enhancing Discrete Choice Models with Neural Networks. *18th Swiss Transport Research Conference*, 2018.