CS230 Winter 2017-2018
Inferring Educational Preferences through Deep Learning

Zhaolei Shi
zshi2@stanford.edu

Problem Statement

Social science methods have traditionally relied on a batch of statistical methods that have evolved separately from machine learning methods. As such, non-linear deep learning methods may develop new insights for long-standing social science questions or offer new avenues to frame social science inquiries. In this project, I explore the application of deep learning to preference predictions in the choice of schools and discuss the relevance of the results for social science research.

Motivation

Although this project investigates preference modeling in the education context, the application of preference modeling are numerous. For example, online retailers such as Amazon or Alibaba would be able to better manage their supply chain by producing a precise model of individual user preferences and how they evolve over time.

Dataset

The data comes from administrative records of a large school district in the US. The dataset includes both student and school yearly characteristics and student choices of schools. The dataset spans the years 2005 to 2015. $110,528$ student submitted choices over this 11 year period, the choices were among 149 schools in the district.
To create negative outcomes in this dataset, for each student-year combination, I filtered out the schools not on the list of choices for that student in that particular year and randomly sampled 10 (or 3) non-choice schools for each choice school listed.

Statistic	N	Mean	St. Dev.	Min	Max
schoolyear	486,432	2,009.804	3.065	2,005	2,015
wht	461,869	0.127	0.333	0	1
blk	461,869	0.117	0.321	0	1
asn	461,869	0.438	0.496	0	1
gifted	486,432	0.189	0.391	0	1
mothered	234,343	2.636	1.091	1	4
fathered	203,503	2.626	1.099	1	4
currentgpa	224,221	2.781	1.042	0.000	4.000
latitude	398,591	37.748	0.030	37.383	38.158
longitude	398,591	−122.437	0.037	−122.532	−122.064

Table 1: Select Student Background

Models

1. *Categorical* is a simple baseline model using feed forward network to accomplish multi-class classification (among the 149 candidate schools).
2. *Stacked* is a feed forward network taking in a joined vector of student and school features to make a binary classification.
3. *Siamese* is a Siamese-like network where two networks are created separately for student and school features. Each network feeds into two hidden layer of equal size. The z for the output layers is therefore $z = \sum_j w_j |f(x_{stu})_j - f'(x_{sch})_j| + b$

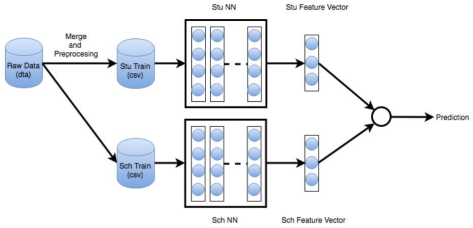

Figure 1: SiameseLike Structure

Results

Model Class	Accuracy	Precision	Recall	F1 Score
categorical	0.287	-	-	-
stacked	0.943	0.726	0.597	0.678
Siamese	0.942	0.689	0.667	0.678

Table 2: Baseline Performance of Different Models

Data Set	Accuracy	Precision	Recall	F1 Score	Network Structure	Feature Vector Size
Top Choice (10 Aug)	0.944	0.677	0.733	0.704	Stu: [875-263-454-152-576-20] Sch: [292-411-20]	20
Top Choice (3 Aug)	0.907	0.779	0.893	0.832	Stu: [123-50-432-20] Sch: [251-175-20]	20
Top 2 Choice (3 Aug)	0.910	0.812	0.848	0.829	Stu: [143-68-110-63-20] Sch: [59-326-818-20]	20
Top 3 Choice (3 Aug)	0.907	0.791	0.862	0.825	Stu: [519-268-72-164-20] Sch: [720-656-20]	20
Any Choice (3 Aug)	0.904	0.788	0.849	0.817	Stu: [143-68-110-63-20] Sch: [59-326-818-20]	20
Higher Choice (3 Aug)	0.906	0.775	0.882	0.825	Stu: [121-133-20] Sch: [251-130-20]	20

Table 3: Best Performing Siamese Models

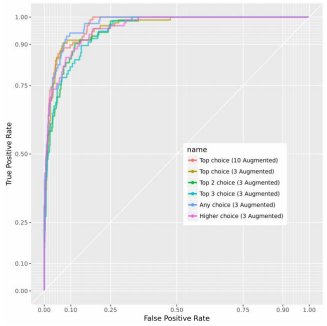

Table 4: ROC Curves for The Best Siamese Models by Dataset

Future Work

Evaluation of the trained models on preference related statistics wi be an important next step. Counter-factual analysis of preferenc changes due to exogenous variations in choice set would also be nice complement.