Discrete choice and machine learning

Two complementary methodologies

Michel Bierlaire

Mathematical Modeling of Behavior

EPFL

Discrete choice and machine learning

Motivation

- The success stories obtained by machine learning methods in many disciplines have motivated several researchers to apply them on choice data.
- In this lecture, we discuss the differences between the two approaches. We also identify a couple of pitfalls that have to be avoided.

Machine learning

Supervised learning / classification

- Population partitioned into J classes.
- Training set: features + true class.
- Classifier: given the features, predicts the class.
- Types: logistic regression, support vector machine, decision trees, neural networks, etc.

Machine learning

Supervised learning / classification

- Population partitioned into J classes.
- Training set: features + true class.
- Classifier: given the features, predicts the class.
- Types: logistic regression, support vector machine, decision trees, neural networks, etc.

Discrete choice

- Choice set with J alternatives.
- Training set: explanatory variables + choice.
- Choice model: given the variables, predicts the choice.
- Types: logit, nested logit, mixtures, etc.

Advantages of machine learning

Model specification

- DC: hand-crafted, simple (linear utility, few interactions).
- ML: complex functional form, data driven.

Model development

- DC: incremental, trial-and-errors.
- ML: systematic.

Model selection

- DC: statistical theory.
- ML: out-of-sample validation.

Pitfalls of machine learning for choice data

Probability that an item n belongs to a class i

Choice models

Probability is used in applications.

Discrete classification

- Typically, class with highest probability is selected.
- "Optimal decision rule"

Severe aggregation bias

Example: classify 1000 items in two classes.

Simple data generation process

51% class 1 / 49% class 2

Perfect ML model

After projection: always predicts class 1

Total number of items in class 1

- In reality: 510
- Predicted: 1000
- Even if the model is acceptable at a disaggregate level, it suffers from a strong aggregation bias.

Aggregation bias increases with the number of classes

Example: classify N items in K classes.

Data generation process

$\frac{1+\varepsilon}{K}$ class 1 / $\frac{K-1-\varepsilon}{K(K-1)}$ class i

Perfect ML model

After projection: always predicts class 1

Total number of items in class 1

- In reality: $N\frac{1+\varepsilon}{K}$
- Predicted: N
- The problem increases with the number of classes.

Machine learning for choice data

Probabilistic classification methods

- Bayes classifiers,
- logistic regression,
- multi-layer perceptrons,
- etc.

Recommendation

Use the probabilistic output for aggregation. Do not project on $0/1$.

Panel data

Out-of-sample validation

- Data randomly split between training and validation set.
- Validation set: not seen during estimation.
- If the data is panel, selection must be based on individuals n and not on observations nt.
- If split on observations, strong correlation between validation set and training set.
- Superiority of ML over DC in prediction has been amplified in the literature based on this mistake.

[Hillel et al., 2021], [Hillel, 2020]

Discrete choice

Important aspects to keep in mind

- Extrapolation.
- Policy analysis.
- Sampling.
- Interpretability.

Notations

Variables

- Explanatory: x
- Choice: i

Models

- Data generating process: $p(i, x)$.
- Discrete choice: theory: $P(i|x;\theta)$
- Machine learning: data-driven:

$$Q(i|x) \approx \frac{p(i, x)}{p(x)}$$

Extrapolation

Objective

Predict i for values of x outside the training data.

Machine learning

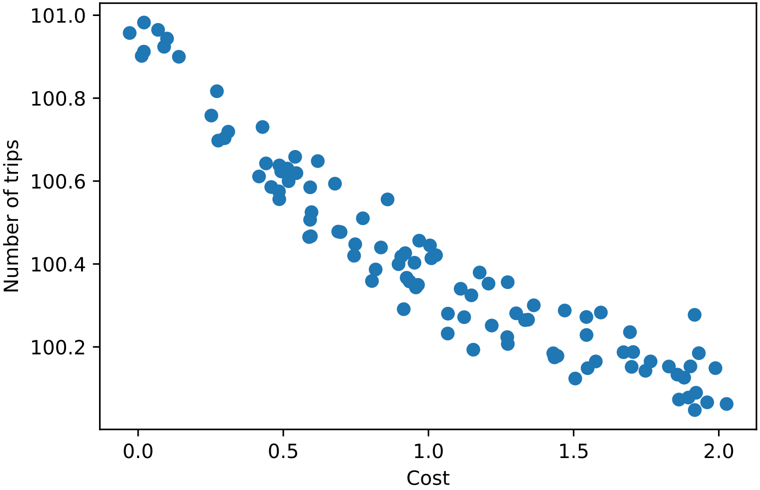
- Focus on goodness of fit.
- Usually good for interpolation.
- May be poor in extrapolation.

Illustration

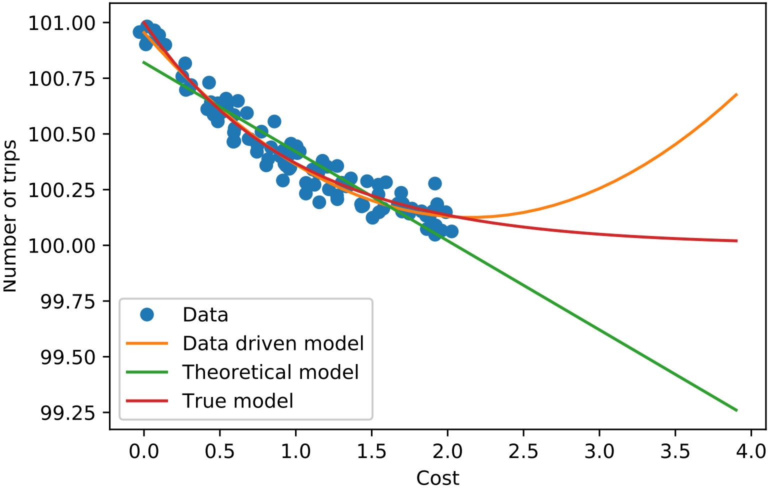
Setup

- Regression (for the sake of illustration).
- For a detailed discrete choice example, see the video Survival of the fittest... or not [youtu.be/_w7RxZIUBqI]
- Predict number of trips vs. cost of public transportation.
- Data generation process: $t = 100 + \exp(-c)$.
- Theoretical model: $t = \theta_1 + \theta_2 c$.
- Theoretical argument: t decreases when c increases: $\theta_2 < 0$.

Data



Models



101.00
100.75
100.50
100.25
100.00
99.75
99.50
99.25
Number of trips
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Cost
Data
Data driven model
Theoretical model
True model

Extrapolation

H. Varian, Berkeley — Chief Economist at Google

Naive empiricism can only predict what has happened in the past. It is the theory — the underlying model — that allows us to extrapolate.

[Varian, 1993]

What-if analysis

Policy analysis

- Impact of new taxes, subsidies.
- Impact of investments that improve the level of service or quality.
- Impact of restrictions, constraints.
- Impact of modifications of the choice set.

What-if analysis

Discrete choice

- Model: $P(i|x;\theta)$.
- Assumption: causal effect is behavioral, and stable over time.
- x can be manipulated at will, within the scope of validity of the model.

Machine learning

- Model: $p(i,x)/p(x)$
- Assumption: captures the correlation in the data.
- Used to predict with x drawn from $p(x)$.
- Does not work if $p(x)$ is modified.

Representativity of the sample

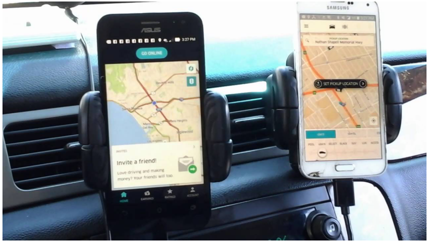
Machine Learning: data processing

- Data is the universe.
- Representativity is assumed.
- Main argument: the size of the data set is very large.
- In general, the research question is posterior to the data collection.

Discrete choice: inference

- A population is identified.
- A research question is proposed.
- Data collection strategies are designed.
- Potential sampling biases are corrected.

Potential implications



Classification

- Bias of the parameters.
- Not necessarily an issue for interpolation.
- May amplify the extrapolation problem.

Aggregation

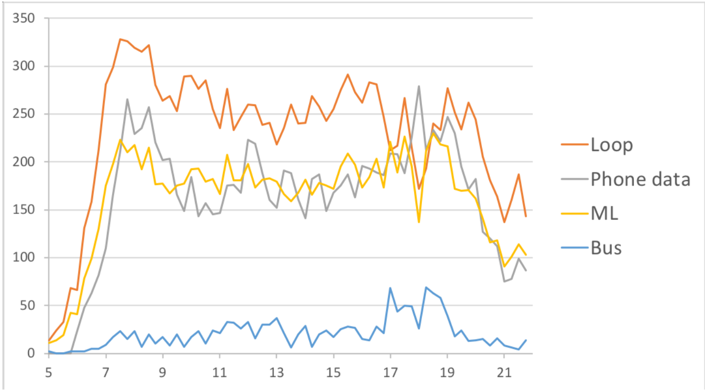
- Counting.
- Aggregation biases may be severe.

Example

City of Geneva

- Data for March 2, 2017.
- Phone data: boundary flows, between adjacent zones.
- ML: results of the ML learning algorithm of the phone company.
- Compared with loop detectors: flows of cars

Results



Source: Montesinos Ferrer, Lamotte, Geroliminis

Interpretation

Beyond prediction

- Models are used to derive quantitative indicators.
- Examples: willingness to pay, social welfare, elasticities, etc.
- Requires a theory, a structure.

Sanity check

- Comparison with published results.
- Communication with practitioners (trust).
- Open the box.

Indicators with ML models

Elasticities

- Point elasticity: OK $\partial Q(i|x)/\partial x_k \cdot x_k/Q(i|x)$.
- Arc elasticity: same issue as "what-if analysis".

Willingness-to-pay and social welfare

- Concepts based on utility theory.
- Requires a combination of ML and utility theory [Sifringer et al., 2018]

Summary

	Discrete choice	Machine Learning
Driven by	theory	data
Captures	causality	correlation
Focus on	interpretability	accuracy
Model:	choice model	classifier
Model selection:	hypothesis testing	out-of-sample validation
Main usage:	what-if analysis	interpolation

Research trends

ML for choice data

- [Brathwaite et al., 2017]
- [Wong et al., 2018]
- [Semenova, 2018]
- [Lee et al., 2018]
- [Hillel et al., 2021]

Using ML strengths in discrete choice

- [Sifringer et al., 2018]
- [Lederrey et al., 2021]

Assistance to specification

- [Hillel et al., 2019]
- [Aboutaleb, 2019]
- [Aboutaleb et al., 2020]

Bibliography I

Aboutaleb, Y. M. (2019).
Learning structure in nested logit models.
Master's thesis, Massachusetts Institute of Technology.

Aboutaleb, Y. M., Danaf, M., Xie, Y., and Ben-Akiva, M. (2020).
Sparse covariance estimation in logit mixture models.
Technical Report arXiv:2001.05034, arXiv.

Brathwaite, T., Vij, A., and Walker, J. L. (2017).
Machine learning meets microeconomics: The case of decision trees and discrete choice.
Technical Report arXiv:1711.04826, arXiv.

Bibliography II

Hillel, T. (2020).
New perspectives on the performance of machine learning classification algorithms for mode-choice prediction.
Technical Report TRANSP-OR 20200704, Transport and Mobility Laboratory - School of Architecture, Civil and Environmental Engineering, EPFL.

Hillel, T., Bierlaire, M., Elshafie, M., and Jin, Y. (2021).
A systematic review of machine learning classification methodologies for modelling passenger mode choice.
Journal of Choice Modelling, 38(100221).

Bibliography III

Hillel, T., Ying, J., Elshafie, M. Z. E. B., and Bierlaire, M. (2019).
Weak teachers: Assisted specification of discrete choice models using ensemble learning.
In Proceedings of the 8th Symposium of the European Association for Research in Transportation (HEART).

Lederrey, G., Lurkin, V., Hillel, T., and Bierlaire, M. (2021).
Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms.
Journal of Choice Modelling, 38(100226).

Bibliography IV

Lee, D., Derrible, S., and Pereira, F. C. (2018).
Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling.
Transportation Research Record, 2672(49):101–112.

Semenova, V. (2018).
Machine learning for dynamic discrete choice.
Technical Report arXiv:1808.02569, arXiv.

Sifringer, B., Lurkin, V., and Alahi, A. (2018).
Enhancing discrete choice models with neural networks.
In Proceedings of the 18th Swiss Transport Research Conference.

Bibliography V

Varian, H. R. (1993).
What use is economic theory?
Working paper 93-14, Department of Economics, University of Michigan, Ann Arbor, Michigan 48109-1220.

Wong, M., Farooq, B., and Bilodeau, G.-A. (2018).
Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling.
Journal of Choice Modelling, 29:152–168.