Discrete choice and machine learning: two peas in a pod?

Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

October 17, 2018

Outline

Future mobility

Trends

- Mobility as a service
- Shared mobility
- Demand patterns are more and more complex
- New sources of data

Travel demand

- Traditional methodology: discrete choice
- Emergence of machine learning

IATBR 2018

Session 3E: Machine Learning – Fundamentals
Session 6E: More Machine Learning

Interest from young researchers

My PhD topic is “Understanding Multi-Modal Passenger Behaviour at City Scale.” I have used trip diary data to compare the performance of multiple discrete choice models, including various **multinomial logistic regression models**, **random forests**, **support vector machines** *and* **neural networks**.”

Outline

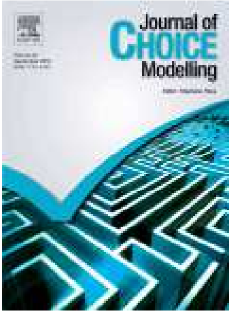
Journal of choice modelling

Issues

- 22 to 29
- 2017 and 2018

Procedure

- Download HTML
- Write Python script to extract words.
- Calculate the occurrences of words.



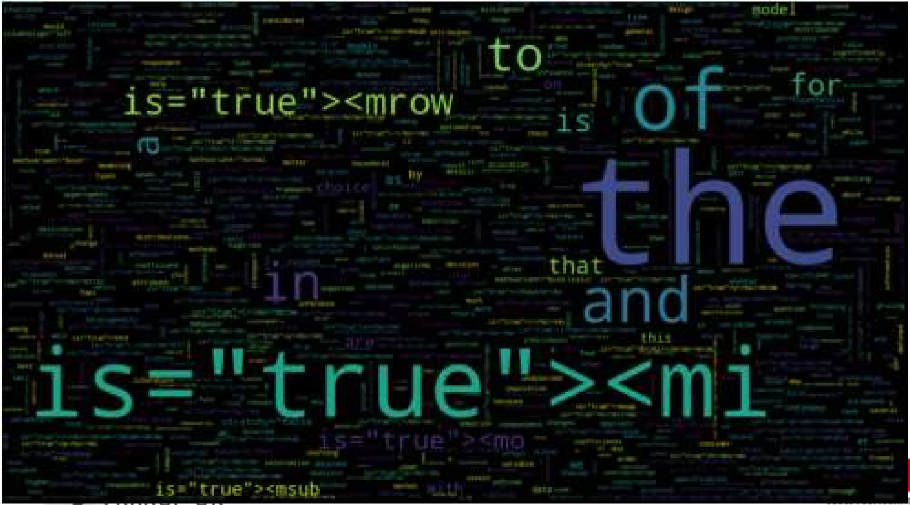
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Some counts

logit	4948
utility	8326
machine	99
learning	1459
statistics	634
pattern	383
classification	0

Visual representation: first attempt



Manual cleaning

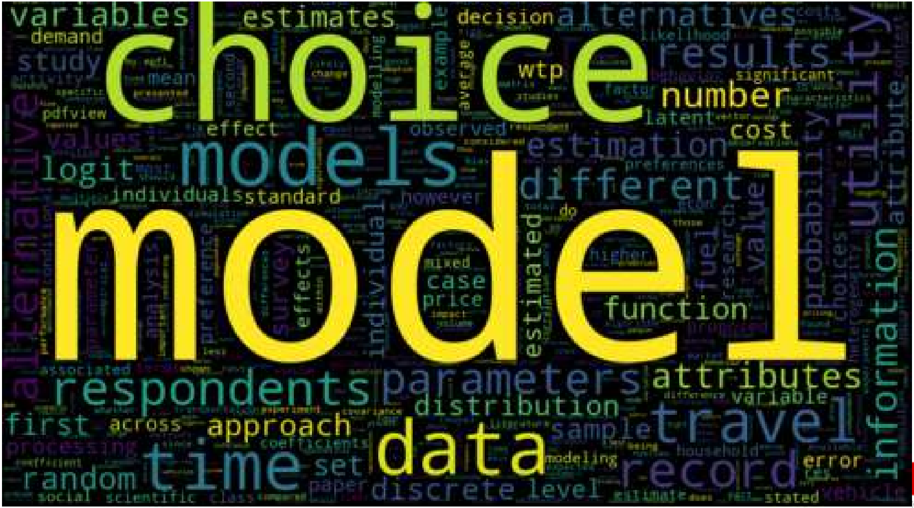
Remove common words

`the is in and of with to for a that are as each et al on by this we be can from it has where such also may pp not all an their one other was than two at only when use table our how new at or they but using both were using if three no more which these have then given into while over used because section based there will about you some many been did between who same would its any among under could`

Remove patterns

Keep only real words — no digits, no special character

Visual representation: second attempt

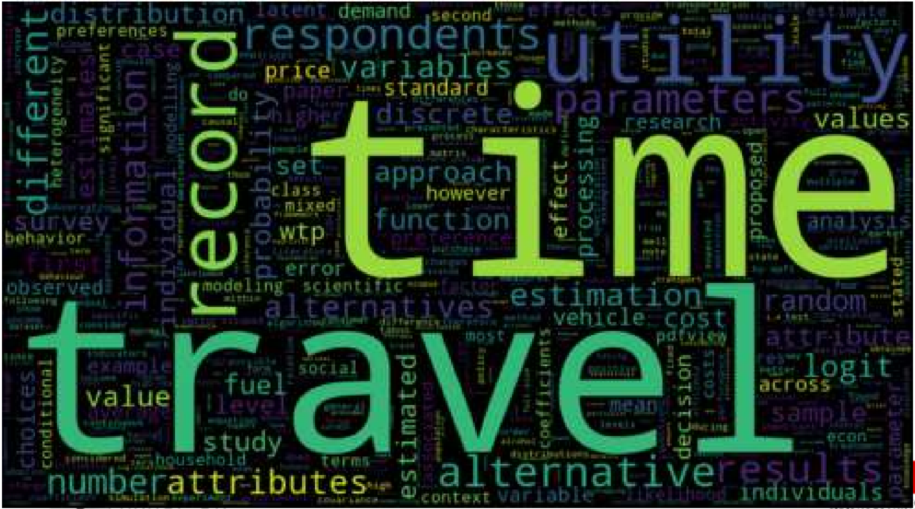


Manual cleaning

Remove obvious words

```
model models choice data
```

Visual representation: third attempt



Machine learning

Manual intervention is common

- “*A great deal of manual work goes into building and training intelligent machine learning algorithms.*” Sascha Schubert, business solutions manager at SAS, May 22, 2017.
- “*Whenever new learning is involved in ML, the human programmer has to intervene and adapt the programming algorithm to make the learning happen.*” Paramita Ghosh, Dataversity.net, April 13, 2017.
- Hyperparameter tuning.
- Learning rate tuning.

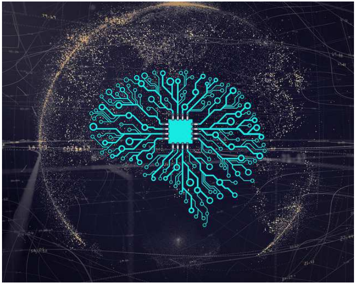
Outline

Definitions

Machine learning is

- an interdisciplinary field
- that uses statistical techniques
- to give computer systems the ability to "learn" from data,
- without being explicitly programmed.

[Wikipedia]

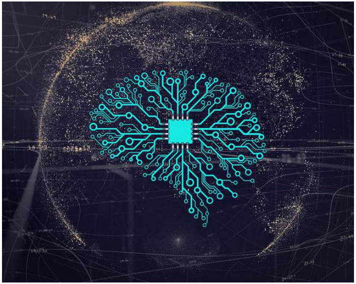


Definitions

Applications of machine learning

- classification
- regression
- clustering
- density estimation
- dimensionality reduction

[Wikipedia]



Discrete choice and classification

Discrete choice from a ML perspective

- dependent variable is discrete
- supervised learning
- logistic regression

Introduction to Discrete Choice Models www.edx.org

Outline

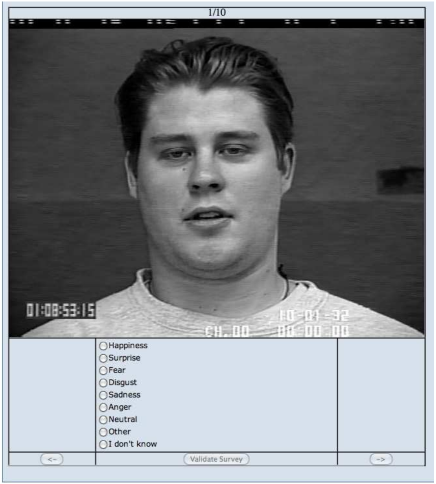
10 years ago: Automatic Facial Expression Recognition

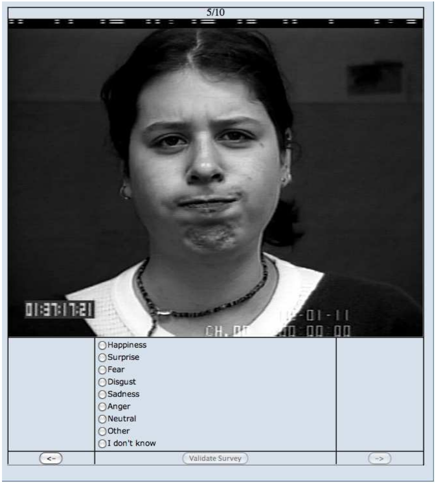
"The face is the most extraordinary communicator, capable of accurately signaling emotion in a bare blink of a second, capable of concealing emotion equally well"

Deborah Blum

Typical machine learning application

Choice experiment

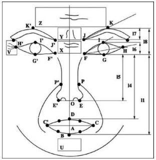


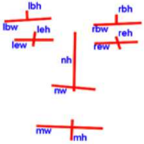


Choice model

[Sorci et al., 2010]

$$V_i = \mathrm{ASC}_i + \underbrace{\sum_k I_{ik}\beta_{ik}^{\mathrm{FACS}}\mathrm{AU}_k} + \underbrace{\sum_h I_{ih}\beta_{ih}^{\mathrm{EDU}}\mathrm{EDU}_h} + \underbrace{\sum_\ell I_{i\ell}\beta_{i\ell}^{\mathrm{AAM}}\mathrm{AAM}}$$





Ingredients

- Facial Action Coding System (FACS) [Ekman and Friesen, 1978]
- Expression Descriptive Units (EDU) [Antonini et al., 2006]
- Active Appearance Model (AAM) [Edwards et al., 1998]

Main conclusions of this work

- Quality of classification similar to neural networks and Bayesian networks.
- Behavioral insights of the discrete choice model.
- Interpretation of the parameters.
- Possibility to exploit know-how in the specification.

Outline

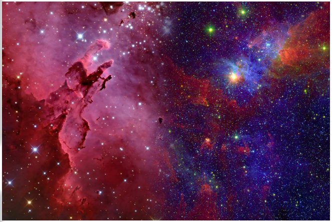
Data universe

Machine Learning: data processing

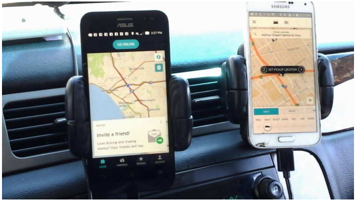
- Dataset is the universe
- Data generation process is usually ignored
- Representativity is assumed
- Main argument: the size of the dataset is very large

Discrete choice: inference

- A population is identified
- Data collection strategies are designed
- Data sets are rebalanced to represent the population



Potential implications



Classification

- Results from statistics: bias of the parameters
- Not necessarily an issue if cross-validation is applied

Aggregation

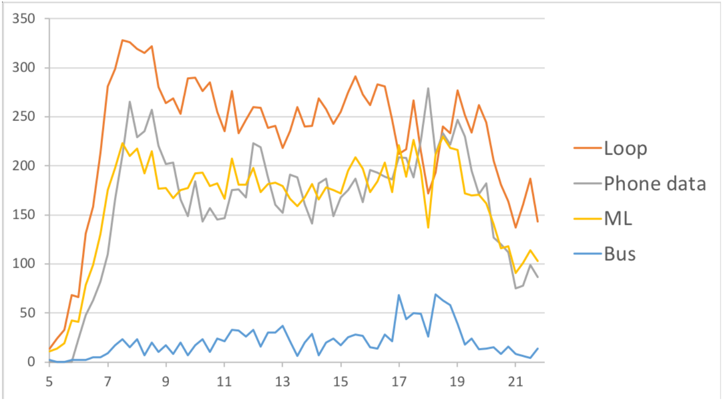
- Counting
- Aggregation biases may be severe

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

Outline

Model output

Probability that an item n belongs to a class i

Choice models

Probability is used in applications

Classification

Class with highest probability is selected

TRANSP-OR

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Severe aggregation bias

Example: classify 1000 items in two classes.

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

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

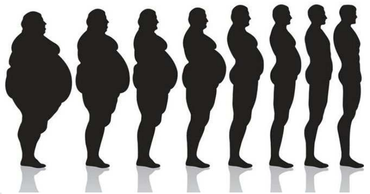
Outline

Loss function/goodness of fit

Formalism

- Class i, item n
- Independent variables/features : $x_n = (x_{in})_{i=1}^J$
- Choice / class: $y_n = (y_{in})_{i=1}^J \in \{0,1\}^J$
- Unknown parameters: $\beta \in \mathbb{R}^K$
- Model: $f(x_n;\beta) \in [0,1]$
- Loss function: finite sums

$$L(\beta) = \sum_{n=1}^{N} L(f(x_n;\beta), y_n)$$



Loss function/goodness of fit

$$L(f(x_n;\beta), y_n) =$$

-Log likelihood / cross entropy

$$-\sum_{i=1}^{J} y_{in} \ln f(x_n;\beta)$$

Square loss

$$\sum_{i=1}^{J} (1 - y_{in} f(x_n;\beta))^2$$

Hinge loss

$$\sum_{i=1}^{J} |1 - y_{in} f(x_n;\beta)|_+$$

Exponential loss

$$\sum_{i=1}^{J} \exp(-\gamma y_{in} f(x_n;\beta))$$

Stochastic gradient descent

Loss function

$$L(\beta) = \sum_{n=1}^{N} L(f(x_n;\beta), y_n)$$

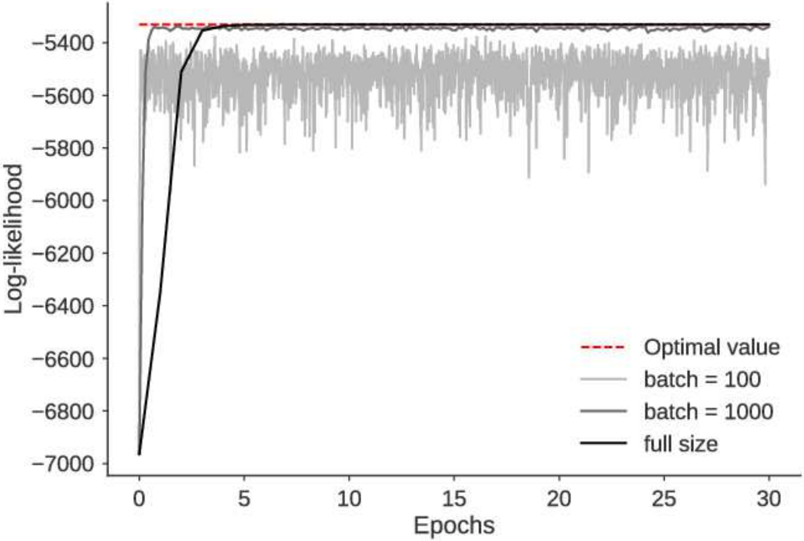
Key ingredient for optimization

Gradient:
$\nabla L(\beta) = \sum_{n\in\{1,\ldots,N\}} \nabla L(f(x_n;\beta), y_n)$

Big data

Approx.: $\sum_{n\in B\subseteq\{1,\ldots,N\}} \nabla L(f(x_n;\beta), y_n)$.

Stochastic gradient on choice data [Lederrey et al., 2018]



Outline

Cross-validation

Main ideas

- How to select the best model?
- It should be the one that predicts best

Example: leave-one-out

$$I_f = \frac{1}{N}\sum_{n=1}^{N} L(f(x_n;\widehat{\beta}_{n-}), y_n)$$

Outline

Summary

	DCM	ML
Manual intervention	Model spec.	Algorithm
Interpretability	Yes	Not quite
Sampling issues	Handled	Mainly Ignored
Model output	Probability	Mostly 0/1
Estimation	standard NL opt.	stochastic gradient
Cross-validation	Mainly ignored	Yes

Conclusions

- Two different communities
- Two different state-of-practice
- Similar objectives

Research agenda

Bring the best from each world

Bibliography I

Antonini, G., Sorci, M., Bierlaire, M., and Thiran, J.-P. (2006).
Discrete choice models for static facial expression recognition.
In Blanc-Talon, J., Philips, W., Popescu, D., and Scheunders, P., editors, *Advanced Concepts for Intelligent Vision Systems*, volume 4179 of *Lecture Notes in Computer Science*, pages 710–721. Springer Berlin / Heidelberg.
ISBN:978-3-540-44630-9.

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998).
Face recognition using active appearance models.
In *European conference on computer vision*, pages 581–595. Springer.

Ekman, P. and Friesen, W. (1978).
Facial coding action system (facs): A technique for the measurement of facial actions.

Bibliography II

Lederrey, G., Lurkin, V., and Bierlaire, M. (2018).
SNM: Stochastic newton method for optimization of discrete choice models.
In *Proceedings of the 21st IEEE International Conference on Intelligent Transportation Systems*, Maui, Hawaii.

Sorci, M., Antonini, G., Cruz, J., Robin, T., Bierlaire, M., and Thiran, J.-P. (2010).
Modelling human perception of static facial expressions.
Image and Vision Computing, 28(5):790–806.