Restricted Boltzmann Machine based Multiple Discrete Continuous Model for very large datasets

Melvin Wong, Bilal Farooq
Ryerson University

July 13, 2018

15th International Conference on Travel Behaviour Research

Session 3E: Machine Learning

Outline

- Introduction
- Generative modelling
  - Restricted Boltzmann machine
  - Model estimation
  - Learning algorithm
- Case study: MTLtrajet dataset
- Conclusion

Table of Contents

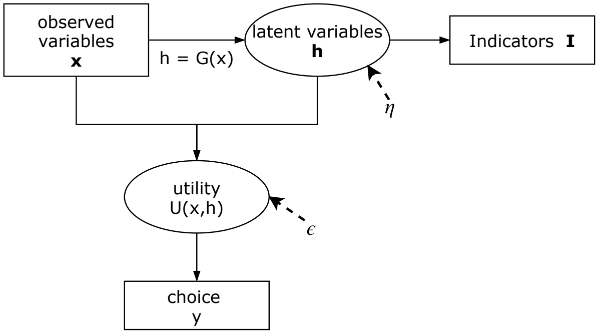
Introduction

Discrete choice models

- Model choice preference y as a function of a set of explanatory variables: $\mathbf{x} = \{x_1, x_2, ..., x_i\}$
- Multinomial logit model
  - $P(y_k = 1 | x_1, x_2, ..., x_i) = e^{V_k} / \sum_{k'} e^{V_{k'}}$
- Estimate parameters $\hat{\theta}$ by max log-likelihood over obs. $\{1, ..., n\}$
  - $\mathcal{L}(\hat{\theta}) = \frac{1}{N} \sum_n \log P_n(y_k)$

Introduction

Structural equation modelling: ICLV model



Introduction

Questions

- Combine various discrete and continuous data types
  - mode choice + trip length + number of trips, etc...
- Joint estimation of probability distributions
  - e.g. $P(y_1, y_2, y_3 | x_1, x_2, ...x_i) \propto P(y_1, y_2, y_3, x_1, x_2, ...x_i)$
- Enhancing behaviour models with latent variables
  - Machine learning algorithm

Ryerson University

LITrans

Table of Contents

Generative behavioural modelling

Background

- Abundance of "Big Data" sources
- Data contains much more latent information than what can be observed
- State-of-the-art in Machine learning applications
  - Hierarchical architectures (Artificial neural networks), recommender systems, collaborative filtering etc...
- Applications in travel behaviour models?

Ryerson University

LITrans

Generative behavioural modelling

Framework

- Describes the generation of data by some **unknown stochastic process**
- Assume some stochastic binary latent variables
- $\mathbf{h} : \{h_1, h_2, ..., h_j\} \in [0, 1] \forall j$
- Given observed data $\mathbf{x}_n : \{x_1, x_2, ..., x_i\}_n$
  - i: explanatory variables and choice variables
- Describe in probabilistic how way latent variables $\mathbf{h}$ could have generated $\mathbf{x}$

Generative behavioural modelling

Restricted Boltzmann Machine

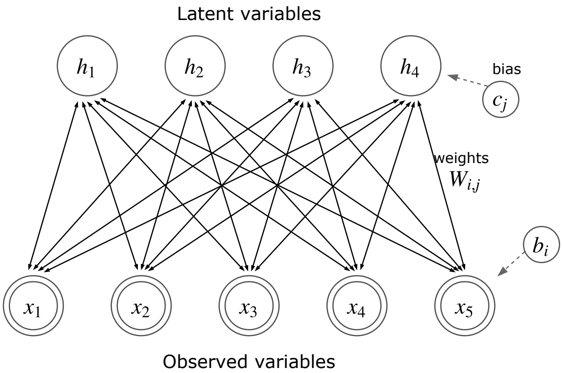
- An energy-based **data-driven** learning model that encodes the joint probability distribution $P(\mathbf{x}, \mathbf{h})$
- Energy $E(x, h)$ is a function that defines a particular configuration of ($\mathbf{x}$,$\mathbf{h}$) pair in the network
- Using a **joint distribution** of observed and latent variables

$$P(\mathbf{x}, \mathbf{h}) = \frac{e^{-E(x,h)}}{\sum_{x,h} e^{-E(x,h)}}$$

$$E(x, h) = -\sum_{i,j} x_i W_{ij} h_j - \sum_{i} b_i x_i - \sum_{j} c_j h_j$$

Generative behavioural modelling

Restricted Boltzmann Machine



Generative behavioural modelling

Energy can be expressed as a sum of terms, or a *Product of Experts*

$$E(x,h) = \sum_i \sum_j f(x_i, h_j)$$

$$P(x,h) \propto \prod_i \prod_j e^{-E(x,h)}$$

Model estimation

The probability of an observed random variable **x** is given as the average over all possible latent variable **h** states:

$$P(\mathbf{x}) = \frac{\sum_h e^{-E(x,h)}}{\sum_{x,h} e^{-E(x,h)}}$$

Maximizing the log likelihood given by

$$\mathcal{L}(\hat{\theta}) = \frac{1}{N} \sum_n \ln P_n(\mathbf{x})$$

Model estimation

Information theory based estimation

- Using the information of the 'true' data distribution to fit the model
- Duality with the **Kullback-Leibler Divergence** D_{KL} measure:

$$
\begin{aligned}
D_{KL}(P_{\theta\{data\}}||P_{\hat{\theta}\{model\}}) &= \sum_{\mathbf{x}} P_{\theta}(\mathbf{x}) \ln \frac{P_{\theta}(\mathbf{x})}{P_{\hat{\theta}}(\mathbf{x})} \\
&= \sum_{\mathbf{x}} P_{\theta}(\mathbf{x}) \ln P_{\theta}(\mathbf{x}) - \sum_{\mathbf{x}} P_{\theta}(\mathbf{x}) \ln P_{\hat{\theta}}(\mathbf{x}) \\
&= \mathcal{H}(x) - \mathcal{L}(\hat{\theta})
\end{aligned}
$$

Log Likelihood = Entropy - KL Divergence

Model estimation

Machine learning optimization

Compute the gradient of $\ln P(\mathbf{x})$:

$$\Delta W_{ij} = \partial \ln P(\mathbf{x}) / \partial W_{ij} = \mathbb{E}_{P_{data}}[x_i h_j] - \mathbb{E}_{P_{model}}[x_i h_j]$$

In practice: difficult to compute $\mathbb{E}_{P_{model}}[x_i h_j]$

- solution: Contrastive Divergence objective function
- approximation of the likelihood function
  - (Carreira-Perpinan and Hinton, 2005)

Model estimation

Machine learning optimization

Generate samples: $\tilde{\mathbf{x}} \sim P(\mathbf{x}|\mathbf{h})$ that resembles the actual observations

- The conditional probability are symmetric
  - $P(\mathbf{h}|\mathbf{x}) = \prod_j P(h_j|\mathbf{x})$
  - $P(\mathbf{x}|\mathbf{h}) = \prod_i P(x_i|\mathbf{h})$
- We can generate (model) different types of data by changing the probability distribution $P(x_i|\mathbf{h})$

Modelling different data types

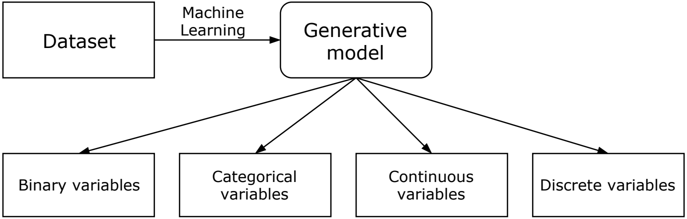
variable type	$x_i \sim P(x_i \mid \mathbf{h})$	distribution
Binary	$\sigma(\sum_j (W_{ij} h_j + c_j))$	Bernoulli
Categorical	$\frac{e^{\sum_j (W_{ij} h_j + c_j)}}{\sum_{x_i} e^{\sum_j (W_{ij} h_j + c_j)}}$	Gumbel
Real (continuous)	$\ln(1 + \exp(\sum_j (W_{ij} h_j + c_j) + \eta_i))$	Normal
	$\mathcal{N}(\sum_j (W_{ij} h_j + c_j), \sigma(\sum_j (W_{ij} h_j + c_j))$	
Real (discrete)	$\sum_r x^r \sim \sigma(\sum_j (W_{ij} h_j + c_j))$	Binomial

$\sigma(x) = \frac{1}{1+e^{-x}}$

Ryerson University

LITrans

Modelling different data types



Model validation

- Accuracy, mean squared-error, cross entropy

Example

Data:= {length, purpose, mode departure time, arrival time}

$$\downarrow$$

{length, purpose, mode departure time} $\rightarrow$ (model) $\rightarrow$ {arrival time}

{purpose, mode} $\rightarrow$ (model) $\rightarrow$ {length, age, departure time, arrival time}

Ryerson University

LITrans

RBM Learning algorithm (I)

Gibbs sampling

- Set initial points for Gibbs sampling:
  - $\mathbf{x}^0 \leftarrow \text{data}$
  - $\mathbf{h}^0 \sim P(\mathbf{h}|\mathbf{x}^0)$
- Compute k-step Gibbs chain
  - $\mathbf{x}^1 \sim P(\mathbf{x}|\mathbf{h}^0)$
  - $\mathbf{h}^1 \sim P(\mathbf{h}|\mathbf{x}^1)$

$$\text{data} \rightarrow \mathbf{x}^0 \rightarrow \mathbf{h}^0 \rightarrow \mathbf{x}^1 \rightarrow \mathbf{h}^1 \ldots \rightarrow \mathbf{x}^\infty \rightarrow \mathbf{h}^\infty$$

RBM Learning algorithm (II)

Learning parameters by stochastic gradient descent

$$\Delta W = \partial \ln P(\mathbf{x}) / \partial W = \mathbb{E}[x_i^0 h_j^0] - \mathbb{E}[x_i^1 h_j^1]$$

$$\begin{aligned} W_{ij}^{\tau} &= W_{ij}^{\tau-1} - \eta \Delta W_{ij}^{\tau} \\ b_i^{\tau} &= b_i^{\tau-1} - \eta \Delta b_i^{\tau} \\ c_j^{\tau} &= c_j^{\tau-1} - \eta \Delta c_j^{\tau} \end{aligned}$$

η : learning rate

Table of Contents

Case study

Data source

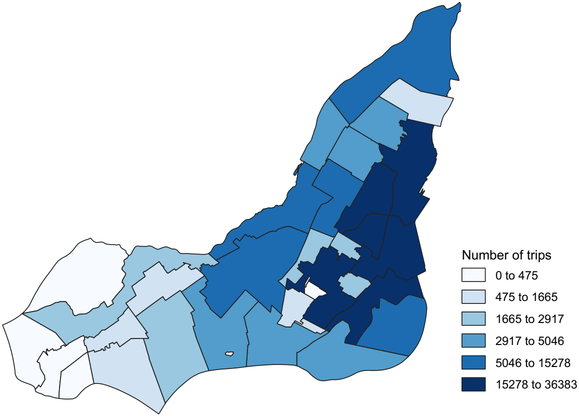
- Open data mobile travel survey app
  - `ville.montreal.qc.ca/mtltrajet/`
  - 293,330 trips
  - features: trip purpose, trip mode, O-D district ID, trip duration, average speed, trip distance (km), O-D departure/arrival time, number of links

Case study

Data processing

- Categorical features (e.g. trip purpose)
  - one-of-k encoding, e.g. $2 \rightarrow [0, 1, 0, 0], \mathcal{C} = 4$
- Cyclic features (e.g. time of day)
  - sin/cos transformation: $t \rightarrow [\sin(2\pi * t/24), \cos(2\pi * t/24)]$

Case study



Ryerson University

LITrans

Case study

Experiment

- Hyperparameters
  - number of latent variables (**h**): 5, 25, 100
  - 100 iterations (SGD + momentum)
  - η: $1e-2$
- Simulation sample size
  - 60497 observations
  - 10 variables $\rightarrow \mathbb{R}^d = 95$
  - Total parameters estimated: 575; 2495; 9695

Case study

Monitoring ML cost

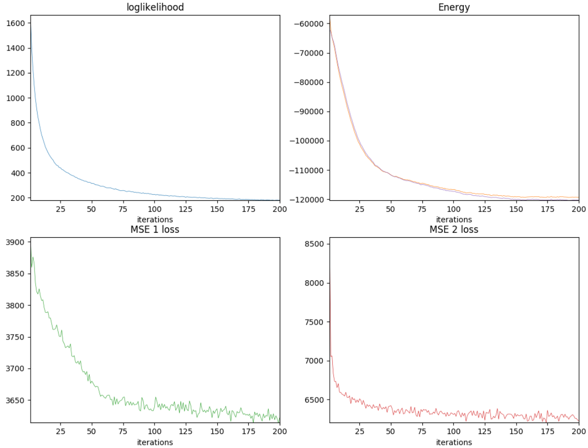
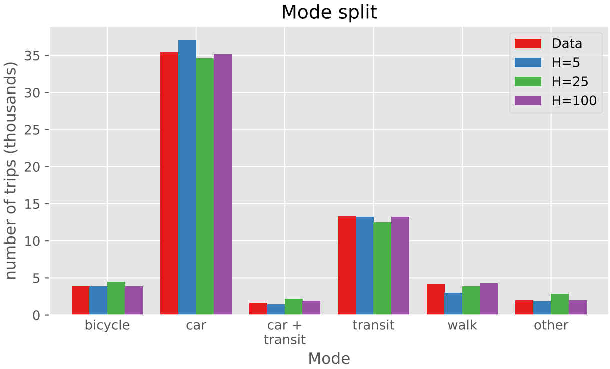


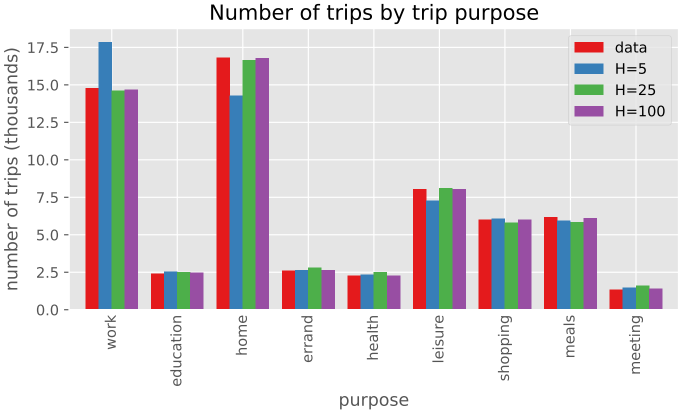
Table of Contents

Results

Simulation results

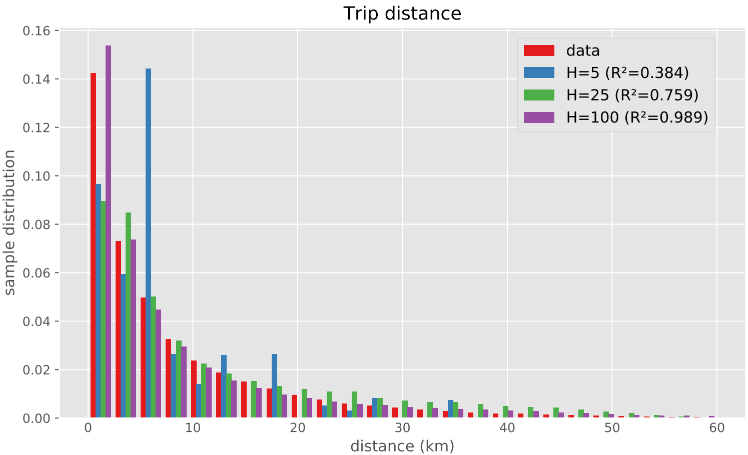


Results



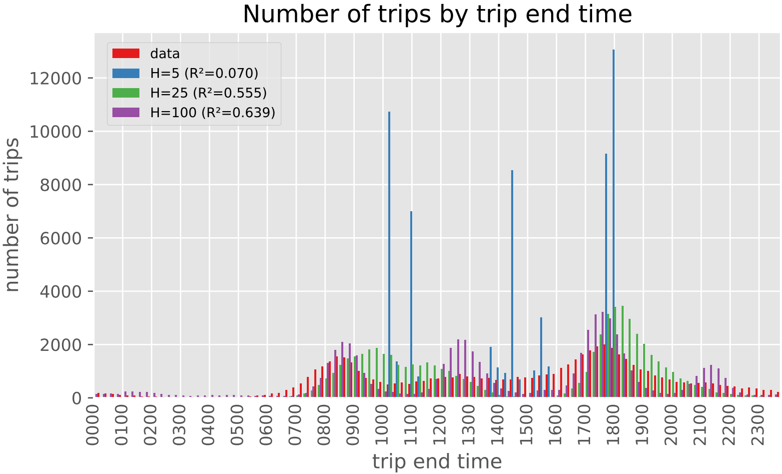
Number of trips by trip purpose
number of trips (thousands)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
data
H=5
H=25
H=100
work
education
home
errand
health
leisure
shopping
meals
meeting
purpose

Results



Trip distance
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
sample distribution
0
10
20
30
40
50
60
distance (km)
data
H=5 (R²=0.384)
H=25 (R²=0.759)
H=100 (R²=0.989)

Results



Ryerson University

LITrans

Results

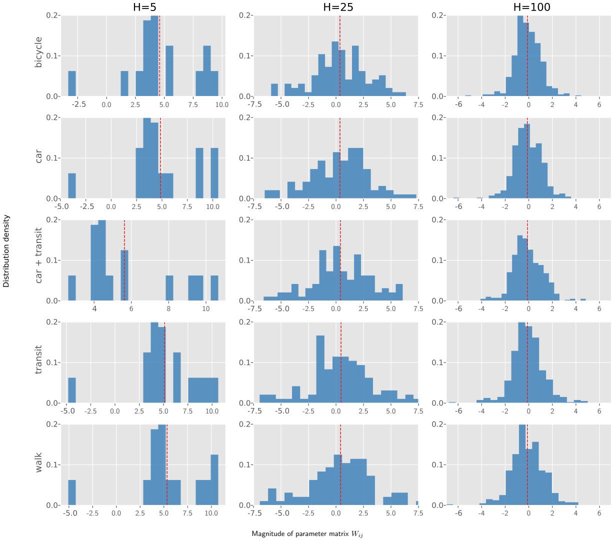

H=5
H=25
H=100
bicycle
car
car + transit
transit
walk
Distribution density
Magnitude of parameter matrix W_{ij}

Table of Contents

Summary

Classical economic models

- Maximize ‘utility’ of preference
- linear-in-parameters

Moving towards generative models

- Minimizing the difference between the expected and preferred outcomes (relative entropy)
- Flexibility gained from data-driven learning models
- Realistic representation of ‘risk’ minimization?

Summary

Key Contributions

- Proposed a generative modelling framework for travel behaviour analysis
- Developed a joint p.d.f. estimation approach for multiple discrete and continuous variables
- Data-driven
  - Captures 'true' behavioural effects in large datasets (error distributions)
  - Leveraging information theory and bayesian inference

On-going research and ideas

- Identifiability in data-driven models
- Regret theory and capturing 'risky' behaviour
- Advanced machine learning methods (RNN, GAN, etc...)

Questions?

Thank you for your attention!

Melvin Wong, Bilal Farooq
Laboratory of Innovations in Transport (LITrans)
Ryerson University, Toronto, Canada
email: melvin.wong@ryerson.ca; bilal.farooq@ryerson.ca