An Overview of Choice Models

Dilan Görür

Gatsby Computational Neuroscience Unit
University College London

May 08, 2009
Machine Learning II

Outline

Outline

1. Overview
   - Terminology and Notation
   - Economic vs psychological views of decision making
   - Brief History
2. Models of Choice from Psychology
   - Bradley Terry Luce (BTL)
   - Elimination by Aspects
3. Models of Choice from Economics
   - Logit
   - Nested Logit
   - Probit
   - Mixed Logit

Discrete Choice Models

Terminology

Goal of choice models

- understand and model the behavioral process that leads to the subject's choice

Data: Comparison of alternatives from a choice set

- Psychophysics experiments
- Marketing: which cell phone to buy

Alternatives:

- Items or courses of action
- Features (i.e. aspects) may or may not be known

Discrete Choice Models

Terminology

Choice set

- discrete – finitely many alternatives
- exhaustive – all possible alternatives are included
- mutually exclusive – choosing one implies not choosing any other

Data

- repeated choice from subsets of the choice set
- number of times each alternative is chosen over some others

Task: Learn the true choice probabilities

Discrete Choice Models

Different approaches

Algebraic or absolute theories in economics vs probabilistic or stochastic theories in psychology

- **Random utility models** – randomness in the determination of subjective value
- **Constant utility models** – randomness in the decision rule

Discrete Choice Models

Economic view of decision making

Desirability precedes availability

- preferences are predetermined in any choice situation, and do not depend on what alternatives are available.

A vaguely biological flavor

- Preferences are determined from a genetically-coded taste template.

The expressed preferences are functions of the consumer's taste template, experience and personal characteristics.
Unobserved characteristics vary continuously with the observed characteristics of a consumer.

Discrete Choice Models

Psychological views of decision making

Alternatives are viewed as a set of aspects that are known.
The randomness in choice comes from the decision rule.

Discrete Choice Models

Different approaches

- Psychology:
  - Process Models
    - Bradley-Terry-Luce (BTL)
    - Elimination by Aspects (EBA)
  - Diffusion Models
  - Race Models
- Economics: Random Utility Maximization Models
  - logit, nested logit
  - probit
  - mixed logit

Discrete Choice Models

Basic assumptions

Simple scalability

- The alternatives can be scaled such that the choice probability is a monotone function of the scale values of the respective alternatives.

Independence from irrelevant alternatives (IIA)

- The probability of choosing an alternative over another does not depend on the choice set.
- Simple scalability implies IIA

There are cases where both assumptions are violated, e.g. when alternatives share aspects.

Brief History

- Thurstone (1927) introduced a Law of Comperative Judgement: alternative i with true stimulus level V_i is perceived with a normal error as $V_i + \varepsilon_i$, and the choice probability for a paired comparison satisfied $P_{[1,2]}(1) = \Phi(V_1 - V_2)$, a form now called binomial probit model.
- Random Utility Maximization (RUM) model is Thurstone's work introduced by Marschak (1960) into economics, exploring the theoretical implications for choice probabilities of maximization of utilities that contained some random elements.
- Independence from irrelevant alternatives (IIA) axiom introduced by Luce (1959) allowed multinomial choice probabilities to be inferred from binomial choice experiments. Luce showed for positive probabilities that IIA implies strict utilities w_i s.t. $P_C(i) = w_i / \sum_{k \in C} w_k$. Marschak proved for a finite universe of objects that IIA implies RUM.

Brief History

- Multinomial Logit (MNL) was introduced by McFadden (1968) in which the strict utilities were specified as functions of observed attributes of the alternatives, $P_C(i) = \exp(V_i)/\sum_{k\in C}\exp(V_k)$.
- McFadden showed that the Luce model was consistent with a RUM model with iid additive disturbances iff these disturbances had Extreme Value Type I distribution.

Outline

Bradley Terry Luce (BTL)

Starts with the IIA assumption and arrives at the choice probabilities:

$$P(x; A) = \frac{U(x)}{\sum_{y \in A} U(y)}$$

A: choice set
$U(\alpha)$: utility of alternative x

Elimination by aspects

Takes into account similarities between alternatives. Represents characteristics of the alternatives in terms of their aspects:

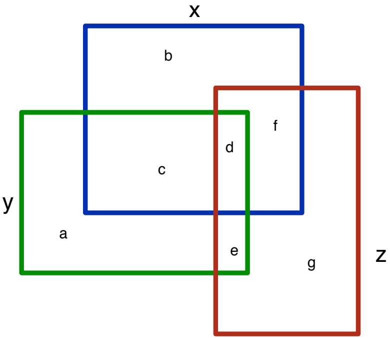
$$x' = \alpha, \beta, \ldots$$

Choice probabilities

$$P(x; A) = \frac{\sum_{\alpha \in x' - A^o} U(\alpha) P(x; A_\alpha)}{\sum_{\beta \in A' - A^o} U(\beta)}$$

A': set of aspects that belongs to at least one alternative in set A
A^o: set of aspects that belong to all alternatives in A
$U(\alpha)$: utility of aspect α

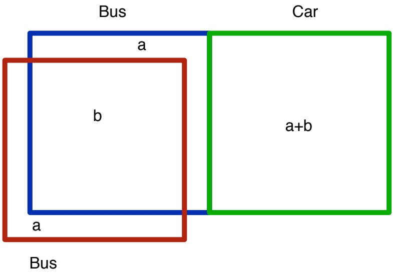
Elimination by aspects



Can cope with choice scenarios where IIA and simple scalability assumptions do not hold.

Elimination by aspects

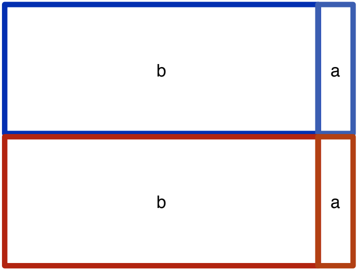
Independence from irrelevant alternatives



IIA suggests the probability of choosing the car to drop from $1/2$ to $1/3$ when a second bus is introduced as the third alternative.

Elimination by aspects

Simple scalability



Simple scalability assumes that the choice probabilities change smoothly with the scale (utility) of an alternative. However, in some situations introducing a small but unique aspect may lead to an immense change in the probabilities.

Outline

Random Utility Maximization

True utility

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

Choice probability

$$P_{ni} = \mathrm{Prob}(U_{ni} > U_{nj} \quad \forall j \neq i)$$

$V_{nj} = V(x_{nj}, s_n) \quad \forall j$
x_{nj} : observed attributes of alternative j
s_n : attributes of decision maker n

Random Utility Maximization

True utility

$$U_{nj} = V_{nj} + \varepsilon_{nj}$$

Choice probability

$$\begin{aligned} P_{ni} &= \mathrm{Prob}(U_{ni} > U_{nj} \ \ \forall j \neq i) \\ &= \mathrm{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \ \forall j \neq i) \\ &= \mathrm{Prob}(\varepsilon_{nj} - \varepsilon_{nj} < V_{ni} - V_{nj} \ \ \forall j \neq i) \\ &= \int_{\varepsilon} \mathbb{I}(\varepsilon_{nj} - \varepsilon_{nj} < V_{ni} - V_{nj} \ \ \forall j \neq i) f(\varepsilon_n)\, \mathrm{d}\varepsilon_n, \end{aligned}$$

where $\varepsilon_n = [\varepsilon_1, \ldots, \varepsilon_J]$

Logit

Extreme Value

Error density and distribution

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}, \qquad F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}$$

Choice probability

$$\begin{aligned} P_{ni} &= \int \left(\prod_{j \neq i} e^{-e^{-\varepsilon_{ni} + V_{ni} - V_{nj}}} \right) e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \, \mathrm{d}\varepsilon_{ni} \\ &= \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}} \\ &= \frac{e^{\beta^T x_{ni}}}{\sum_j e^{\beta^T x_{nj}}} \end{aligned}$$

Nested Logit

Generalized Extreme Value

The set of alternatives are partitioned into subsets called nests such that:

- IIA holds within each nest.
- IA does not hold in general for alternatives in different nests.

Error distribution

$$F(\varepsilon_n) = \exp\Big(-\sum_{k=1}^{K} \big(\sum_{j \in B_k} e^{-\varepsilon_{nj}/\lambda_k}\big)^{\lambda_k} \Big)$$

λ_k is a measure of the degree of independence among the alternatives in nest k.

Nested Logit

Generalized Extreme Value

The set of alternatives are partitioned into subsets called nests such that:

- IIA holds within each nest.
- IA does not hold in general for alternatives in different nests.

Choice probability

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}(\sum_{j \in B_k} e^{V_{nj}/\lambda_k})^{\lambda_k - 1}}{\sum_{l=1}^{K}(\sum_{j \in B_l} e^{V_{nj}/\lambda_l})^{\lambda_l}}$$

λ_k is a measure of the degree of independence among the alternatives in nest k.

Nested Logit

- The model is RUM consistent if $0 \leq \lambda_k \leq 1 \quad \forall k$
- For $\lambda_k > 1$, the model is RUM consistent for some range of explanatory variables.
- $\lambda_k < 0$ is RUM inconsistent as it implies that improving attributes of an alternative can decrease its probability of being selected.
- if $\lambda_k = 1$ the model reduces to Logit
- as $\lambda_k \to 0$ the model approaches the Elimination by Aspects model

Generalizations of Nested Logit

- Higher-level nests
- Overlapping nests
  - **Cross Nested Logit**
    - multiple overlapping nests
  - **Ordered GEV**
    - correlation depends on the ordering of alternatives
  - **Paired Combinatorial Logit**
    - each pair of alternatives constitutes a nest with its own correlation
  - **Generalized Nested Logit**
    - multiple overlapping nests with a different membership weight to each net for each alternative.

Simple procedure for defining any GEV

Define $Y_j = \exp(V_j)$, and consider a function $G = G(Y_1, \ldots, Y_J)$ with partial derivatives $G_i = \frac{\partial G}{\partial Y_i}$

$$P_i = \frac{Y_i G_i}{G} \quad \text{if}$$

1. $G \geq 0$ for all positive values of $Y_j \quad \forall j$.
2. G is homogeneous of degree one, i.e. $G(\rho Y_1, \ldots, \rho Y_J) = \rho G(Y_1, \ldots, Y_J)$
3. $G \to \infty$ as $Y_j \to \infty$ for any j
4. The cross partial derivatives of G have alternating signs.

Simple procedure for defining any GEV

Examples

Logit: $G = \sum_{j=1}^{J} Y_j$

Nested Logit: $G = \sum_{l=1}^{K} \left(\sum_{j \in B_l} Y^{1/\lambda_l} \right)^{\lambda_l}, \quad 0 \leq \lambda_k \leq 1 \quad \forall k$

Paired Combinatorial Logit: $G = \sum_{k=1}^{J-1} \sum_{l=k+1}^{J} \left(Y_k^{1/\lambda_{kl}} + Y_l^{1/\lambda_{kl}}) \right)^{\lambda_{kl}}$

Generalized Nested Logit: $G = \sum_{k=1}^{K} \left(\sum_{j \in B_k} (\alpha_{jk} Y_j)^{1/\lambda_k} \right)^{\lambda_k}$

Probit

Normal

Error distribution: Normal

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{J/2}|\Omega|^{1/2}} e^{-\frac{1}{2}\varepsilon_n^T \Omega^{-1} \varepsilon_n}$$

Choice Probabilities

$$P_{ni} = \int \mathbb{I}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \ \forall j \neq i)\phi(\varepsilon_n)\, \mathrm{d}\varepsilon_n,$$

This integral does not have a closed form.

Mixed Logit

can approximate any RUM model

A hierarchical model where the logit parameters β are given prior distributions so that the choice probabilities are given as:

$$P_{ni} = \int L_{ni}(\beta) f(\beta) \, \mathrm{d}\beta,$$

where $L_{ni}(\beta)$ is the logit probability evaluated at parameters β:

$$L_{ni}(\beta) = \frac{e^{V_{ni}\beta}}{\sum_{i=1}^{J} e^{V_{nj}\beta}}$$

$f(\beta)$ can be discrete or continuous. e.g. if β takes M possible values, we have a latent class model

Mixed Logit

can approximate any RUM model

A hierarchical model where the logit parameters β are given prior distributions so that the choice probabilities are given as:

$$P_{ni} = \int L_{ni}(\beta) f(\beta) \, \mathrm{d}\beta,$$

where $L_{ni}(\beta)$ is the logit probability evaluated at parameters β:

$$L_{ni}(\beta) = \frac{e^{V_{ni}\beta}}{\sum_{i=1}^{J} e^{V_{nj}\beta}}$$

$f(\beta)$ can be discrete or continuous. e.g. if β takes M possible values, we have a latent class model

Extensions of Choice Models

Model extensions

- Nonparametric distributions over noise distribution
- Nonparametric distributions over explanatory variable parameters
- Nonparametric prior over the functions on explanatory variables

Application areas

- Conjoint analysis
- Ranking, information retrieval

References

- McFadden, "Economic Choices", Nobel Prize Lecture, 2000
- Train, "Discrete Choice Methods with Simulation", Cambridge University Press, 2003
- Tversky, "Elimination by Aspects: A Theory of Choice", Psychological Review, 1972
- Greene, "Discrete Choice Modeling", 2008
- Cao et al., "Learning to Rank: From Pairwise to Listwise Approach", ICML 2007
- Chu, Ghahramani, "Preference Learning with Gaussian Processes", ICML 2005