KERNEL LOGISTIC REGRESSION MODEL: AN ALTERNATIVE PREDICTION APPROACH FOR TRAVEL BEHAVIOR *

By Chong WEI**, Takamasa IRYO***, and Yasuo ASAKURA****

1. Introduction

Multinomial logit (MNL) model are commonly used to predict travel behavior. MNL model explains the behavior in terms of the linear compensation of the utility. However some researches denoted that people may use non-compensatory rule, and therefore, the bias in prediction may be raised by the linear compensatory model. Machine learning approach is considered as an alternative approach to describe the travel behavior. Machine learning approach describes the data by statistical theory and which is not constrained by the economic explanation.

Some sophisticated machine learning approaches, such as linear discriminant, Support Vector Machine (SVM) and Artificial Neural Networks (ANN), are widely used in the field of transportation research. These machine learning approaches are based on discriminant function. Discriminant function can be described as: $\{g_i(\mathbf{x})\}_{i=1}^{I}$, where i is the index of the classification or the alternative and I is the number of the classification or the alternative. If $g_i(\mathbf{x}) > g_j(\mathbf{x}),\ \forall i \neq j$ then the input $\mathbf{x}$ will be assigned to classification i. The discriminant function based technology may raise some problems in the context of travel behavior research. First, the sizes (i.e.: the number of alternative) and the alternative may be different for each choice set of the SP/RP survey data, however, the general structure of discriminant function based technology cannot address this problem easily. Additionally, discriminant functions based technology is that given a tumor sample, it only predicts a class label but does not provide probability information[1)], on the other hand, choice probability is usually considered as a important information for travel behavior research. In this paper, we propose a new machine learning based behavior model for addressing the problem mentioned above.

2. Methodology

(1) Outline of the proposed model

In the discrete choice model, the choice decision rule is: $C = \mathrm{argmax}_{\forall i} U_i$ where C is the index of the chosen alternative, U_i is the utility of alternative i. In the proposed model we employ a new criterion instead of utility. Let $P(Y=1|\mathbf{x}_i)$ denotes the probability of that the decision maker is satisfied given alternative i, the attribute vector of the alternative is $\mathbf{x}_i$, Y is a binary variable (0 or 1). In the proposed model, we use $P(Y=1|\mathbf{x}_i)$ instead of utility, U_i. The following decision rule was obtained: $C = \mathrm{argmax}_{\forall i} P(Y=1|\mathbf{x}_i)$.

*Keywords: kernel logistic regression, machine learning and modal split

**non-member of JSCE, Master of Engineering, Graduate School of Engineering, Kobe University
(1-1, Rokkodai-cho, Nada, Kobe, 657-8501, Japan, TEL: +81-78-803-6360, FAX: +81-78-803-6360)
***member of JSCE, Doctor of Engineering, Graduate School of Engineering, Kobe University
(1-1, Rokkodai-cho, Nada, Kobe, 657-8501, Japan, TEL: +81-78-803-6360, FAX: +81-78-803-6360)
****member of JSCE, Doctor of Engineering, Graduate School of Engineering, Kobe University
(1-1, Rokkodai-cho, Nada, Kobe, 657-8501, Japan, TEL: +81-78-803-6360, FAX: +81-78-803-6360)

Here, we consider that the distribution of Y is Bernoulli distribution with parameter $\pi(\mathbf{x}_i)$, where $\pi(\mathbf{x}_i)=E(Y|\mathbf{x}_i)$, and $E(Y|\mathbf{x}_i)$ is the expected value of $Y|\mathbf{x}_i$. We have $E(Y|\mathbf{x}_i)=P(Y=1|\mathbf{x}_i)$ and :

$$C=\text{argmax}_{\forall i}E(Y|\mathbf{x}_i)=\text{argmax}_{\forall i}\pi(\mathbf{x}_i)\ldots\ldots(1)$$

where $0\leq\pi(\mathbf{x}_i)\leq 1$. Generally, logistic regression model can be used to predict $\pi(\mathbf{x}_i)$:

$$\text{logit}(\pi(\mathbf{x}_i))=f(\mathbf{x}_i)+\varepsilon_i\ldots\ldots(2)$$

where $f(\mathbf{x}_i)$ and ε_i is the regression equation and the random error component respectively, $\text{logit}(\pi(\mathbf{x}_i))$ denotes logit transformation: $\text{logit}(\pi(\mathbf{x}_i))=\log\left\{\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]^{-1}\right\}$. If we let $\pi(\mathbf{x}_i)=f(\mathbf{x}_i)+\varepsilon_i$ then the value of $f(\mathbf{x}_i)+\varepsilon_i$ has to be between zero and one. This means that a constraint condition must be added to the regression model. A solution to this problem is to use logit transformation map $\pi(\mathbf{x}_i)$ from the range (0, 1) to ($-\infty,+\infty$), therefore, regression model can be applied without constraint conditions.

Let P_i denotes the probability of alternative i is chosen by decision maker given the choice set. According to equation (1), P_i can be denoted as $P_i=P\left(\pi(\mathbf{x}_i)>\pi(\mathbf{x}_j),\forall j\neq i\right)$, where alternative i and j are belong to the same choice set. Equally (we omit the proof) we have: $P_i=P\left(f(\mathbf{x}_i)+\varepsilon_i>f(\mathbf{x}_j)+\varepsilon_j,\forall j\neq i\right)$. If we assume an Gumbel distribution for $\varepsilon_i,\forall i$, then we have:

$$P_i=\exp f(\mathbf{x}_i)\left\{\sum_{j=1}^{I}\exp f(\mathbf{x}_j)\right\}^{-1}\ldots\ldots(3)$$

There is a different between equation (3) and MNL model. $f(\mathbf{x}_i)$ in equation (3) denote a regression model rather than the utility function in MNL model. Regression model describes the property of the training data and which does not explain the mechanism of behavior. We propose a Kernel Logistic Regression (KLR), a machine learning approach for $f(\mathbf{x}_i)$. KLR model is a form of nonlinear regression and which is also good at small size training data (usually, the size of a SP survey data is small).

First we introduce Kernel trick[2]. Kernel or $K(.,.)$ is a kind of function which satisfies Mercer's Theorem[2]. We can express $K(.,.)$ as: $K(\mathbf{x},\mathbf{x}')=\langle\phi(\mathbf{x}),\phi(\mathbf{x}')\rangle$ where $\phi:\mathbf{x}\subset\mathbb{R}^p\rightarrow l_2$, here l_2 denote a infinite dimension space, that means $\phi(\mathbf{x})$ is a vector in the new space. $\mathbf{x}$ and $\mathbf{x}'$ are two vectors from the space of input, $\langle.,.\rangle$ denote the inner product of two vectors. Obviously, it is difficult to calculate the result of $\phi(\mathbf{x})$ (i.e: curse of dimensionality). Fortunately, in the applications of machine learning approach, we usually only want to know the value of inner product of two vectors rather than the full information of the vectors, therefore, we can adopt Kernel trick that use $K(\mathbf{x},\mathbf{x}')$ to calculate the inner product of the vectors without calculating $\phi(\mathbf{x})$. In this paper we employed a Gaussian Kernel:

$$K(\mathbf{x},\mathbf{x}')=\exp\left(-\tfrac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2\right)\ldots\ldots(4)$$

Although the structure of $\phi(\cdot)$ which corresponding to equation (4) is unknown, the value of inner product of these two vectors can be derived from the Gaussian Kernel.

Kernel trick can be used in the regression model[3]. Let $\mathbf{x}$ is the independent variable vector and y is the dependent variable. Function $f(\mathbf{x})$ approximation to y, the parameters of $f(\mathbf{x})$ are estimated using the following regularization problem[4] : $\min_f\left[\sum_{m=1}^{M}L\left(y^{(m)},f\left(\mathbf{x}^{(m)}\right)\right)+\lambda J(f)\right]$. Where $L(.,.)$ is a loss function, $J(.)$ is a penalty functional, m is the index of training data, M is the number of training data. Generally, $f(x)=\sum_{t=1}^{\infty}c_t\phi_t(\mathbf{x})$, where c_t is the unknown parameters, $\phi_t(x)$, $t=1,2,...,\infty$ is a series function which transform the input $\mathbf{x}$ into a infinite dimension space. We have:

$$\min_{\{C_t\}_1^\infty}\left[\sum_{m=1}^{M}L\left(y^{(m)},f\left(\mathbf{x}^{(m)}\right)\right)+\lambda J(f)\right]\ldots\ldots(5)$$

The solution of (5) is[5]

$$f(\mathbf{x})=\sum_{m=1}^{M}\alpha_m K\left(\mathbf{x},\mathbf{x}^{(m)}\right)\ldots\ldots(6)$$

where α_m is the unknown parameters. Equation (6) is the general formalism of KLR model.

Let M is the number of training data (i.e: total N answers are collected by a survey, and I_n is the number of alternative in the question n, M can be expressed as $M=\sum_{n=1}^{N}I_n$). The regression model for this study can be described as:

$$f(\mathbf{x}_q)=\sum_{n=1}^{N}\sum_{i=1}^{I_n}\alpha_{ni}K(\mathbf{x}_{ni},\mathbf{x}_q)\ldots\ldots(7)$$

where $\mathbf{x}_q$ denote the attribute vector of alternative q in the current choice set, $\mathbf{x}_{ni}$ denote the attribute vector of alternative i in the question n.

(2) Estimation

The parameters (α_{ni}) of the proposed model can be estimated using maximum likelihood estimation. The likelihood function is: $L(\boldsymbol{\alpha})=\prod_{n=1}^{N}\prod_{i=1}^{I_n}(P_{ni})^{s_{ni}}$, where $\boldsymbol{\alpha}$ is the parameters vector, $s_{ni}=1$ if alternative i in the question n was chosen, P_{ni} is the choice probability of alternative i in the question n. The log-likelihood function is: $LL(\boldsymbol{\alpha})=\sum_{n=1}^{N}\sum_{i=1}^{I_n}s_{ni}\ln P_{ni}$. We can use Newton-Raphson method or quasi-Newton method to maximize the log-likelihood function $LL(\boldsymbol{\alpha})$. The solution of this unconstraint maximization problem is the estimation result for the parameters.

3. Numerical example

(1) Data set

Although a validation test should be presented in this paper, we would like to give a numerical example to explicitly explain how to use the proposed model.

We explore the modal split among three traffic modes. The attributes considered are 'Travel Time (min)' and 'Fee (1000 yen)' among the traffic mode 'A', 'B' and 'C'. A total of three samples were obtained from a SP survey. Table 1 shows the details of the samples. Let T_{1A} and F_{1A} denote 'Travel Time' and 'Fee' of alternative 'A' in question / sample 1 respectively.

Table 1. The samples

Question Index	1		2			3	
Parameters $\boldsymbol{\alpha}$	α_1	α_2	α_3	α_4	α_5	α_6	α_7
Choice Result	1	0	0	1	0	0	1
Alternative	A	B	A	B	C	A	C
Travel Time (min)	0.22	0.10	0.30	0.15	0.10	0.25	0.16
Fee (1000 yen)	0.20	0.50	0.20	0.60	1.00	0.31	0.51

Table 2. An example of modal split

Alternative	A	B
Travel Time	0.18	0.12
Fee	0.21	0.52

(2) Parameters estimation

The parameters to be estimated in the model are $[\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5\ \alpha_6\ \alpha_7]$. The log-likelihood function to be maximized is:

$$\ln\left\{\exp f(\mathbf{x}_{1A})[\exp f(\mathbf{x}_{1A})+\exp f(\mathbf{x}_{1B})]^{-1}\right\}+\ln\left\{\exp f(\mathbf{x}_{2B})[\exp f(\mathbf{x}_{2A})+\exp f(\mathbf{x}_{2B})+\exp f(\mathbf{x}_{2C})]^{-1}\right\}+$$
$$\ln\left\{\exp f(\mathbf{x}_{3C})[\exp f(\mathbf{x}_{3A})+\exp f(\mathbf{x}_{3C})]^{-1}\right\}\ldots\ldots(8)$$

where:

$$\begin{aligned}f(\mathbf{x}_{1A})=&\alpha_1\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{1A})^2+(F_{1A}-F_{1A})^2\right]\right\}+\alpha_2\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{1B})^2+(F_{1A}-F_{1B})^2\right]\right\}+\\&\alpha_3\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{2A})^2+(F_{1A}-F_{2A})^2\right]\right\}+\alpha_4\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{2B})^2+(F_{1A}-F_{2B})^2\right]\right\}+\\&\alpha_5\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{2C})^2+(F_{1A}-F_{2C})^2\right]\right\}+\alpha_6\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{3A})^2+(F_{1A}-F_{3A})^2\right]\right\}+\\&\alpha_7\exp\left\{-\tfrac{1}{2}\left[(T_{1A}-T_{3B})^2+(F_{1A}-F_{3B})^2\right]\right\}\ldots\ldots(9)\end{aligned}$$

Parameters $[\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5\ \alpha_6\ \alpha_7]$ are coefficients in the statistical model, consequently, which cannot be used to calculate the economic indicators such as the value of travel time savings (VTTS). We can consider that the parameters are the weight of the attribute vector of the alternative in each question.

(3) Prediction

In this section, we present an example of travel behavior prediction using the samples shown in Table 1 and the estimated parameter. We want to predict the modal split between mode ‘A’ and mode ‘B’. Table 2 shows the attributes of mode ‘A’ and mode ‘B’. Let $\mathbf{x}_{1A}=\left[T_{1A}\ F_{1A}\right]$ denote the attribute vector of alternative ‘A’ in question / sample 1, $\mathbf{x}_A$ denote the attribute vector of current alternative ‘A’. The choice probability of mode ‘A’ and mode ‘B’ can be calculated as:

$$P_A=\exp f\left(\mathbf{x}_A\right)\left[\exp f\left(\mathbf{x}_A\right)+\exp f\left(\mathbf{x}_B\right)\right]^{-1} \ldots\ldots(10)$$

$$P_B=\exp f\left(\mathbf{x}_B\right)\left[\exp f\left(\mathbf{x}_A\right)+\exp f\left(\mathbf{x}_B\right)\right]^{-1} \ldots\ldots(11)$$

where:

$$\begin{aligned}
f\left(\mathbf{x}_A\right)=&\ \alpha_1\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{1A}\|^2\right)+\alpha_2\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{1B}\|^2\right)+\alpha_3\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{2A}\|^2\right)+\alpha_4\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{2B}\|^2\right)+\\
&\ \alpha_5\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{2C}\|^2\right)+\alpha_6\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{3A}\|^2\right)+\alpha_7\exp\left(-\tfrac{1}{2}\|\mathbf{x}_A-\mathbf{x}_{3B}\|^2\right)\\
=&\ \alpha_1\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.22\right)^2+\left(0.21-0.20\right)^2\right]\right\}+\alpha_2\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.10\right)^2+\left(0.21-0.50\right)^2\right]\right\}+\\
&\ \alpha_3\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.30\right)^2+\left(0.21-0.20\right)^2\right]\right\}+\alpha_4\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.15\right)^2+\left(0.21-0.60\right)^2\right]\right\}+\\
&\ \alpha_5\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.10\right)^2+\left(0.21-1.00\right)^2\right]\right\}+\alpha_6\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.25\right)^2+\left(0.21-0.31\right)^2\right]\right\}+\\
&\ \alpha_7\exp\left\{-\tfrac{1}{2}\left[\left(0.18-0.16\right)^2+\left(0.21-0.51\right)^2\right]\right\}\ldots\ldots(12)
\end{aligned}$$

4. Conclusion

In this paper we propose a machine learning model to describe the travel behavior data. The structure of the proposed model (see Equation 3) indicates that the model can fit the data set which including different alternatives and size for the choice set (e.g. Table 1). The data are fitted by a Kernel Logistic Regression in the proposed model which is not restricted by the economic explanation, and therefore, Kernel Logistic Regression may be able to better describe the data. On the other hand, the parameters of the proposed model cannot be used to calculate the economic indicators. We can see that the proposed model focus on the accuracy of prediction rather than explain the mechanism of behavior. In this paper we present a numerical example, however a real data set should be used to convince about the performance of the proposed model.

Acknowledgement

The first author thanks Takeshi Nagae for his helpful suggestion for the computer program of the presented study.

References

1) Zhu, J. and Hastie, T.: Classification of gene microarrays by penalized, Biostatistics, Vol 5, pp. 427–443, 2004.

2) Muller, K., *et al.*: An introduction to kernel-based learning algorithms, IEEE Transaction on Neural Networks, Vol. 12, PP. 181-202, 2001.

3) Hastie, T., *et al.*: The Elements of Statistical Learning, Springer, New York, 2001.

4) Le Cessie, S. and Van Houwelingen, J.: Ridge estimators in logistic regression, Applied Statistics, Vol. 41, pp.191-201, 1992.

5) Wahba, G.: Spline models for observational data, Series in Applied Mathematics, Vol. 59, SIMA, Philadelphia, 1990.