ARTICLE IN PRESS

Journal of Choice Modelling xxx (2017) 1–16

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm

Boosting conditional logit model

Haolun Shi [*], Guosheng Yin

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Keywords:
Boosting
Conditional logit model
Discrete choice
Smoothing spline
Utility function

ABSTRACT

A componentwise smoothing spline-based boosting procedure is developed for the conditional logit model to estimate the covariate effects nonparametrically. The proposed method can be applied to discrete choice modeling to predict the choice outcomes. Our boosting procedure possesses the properties of slow over-fitting behaviour, automatical variable selection, consistent approximation to the utility function, and the ability to capture the potential nonlinear covariate effects. We show in the simulation studies that the method can provide accurate estimates of the true functional forms of the covariate effects and can select the predictors that are most related to the choice utility. The proposed boosting conditional logit procedure is also applied to two real datasets and its prediction accuracy is demonstrated to be superior to that of the conventional conditional logit regression.

1. Introduction

Random utility (RU) models are of paramount interest in the analysis of discrete choices. The RU models are based on the economic principle of utility maximization. For an individual n choosing alternative j, the RU model typically assumes that an underlying utility consists of a deterministic component and a random component,

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

where the deterministic component V_{nj} is often defined to be an additive function of the attributes, and the random component ε_{nj} represents the influence from the unobserved attributes on the choice behaviour, and the interpersonal and intrapersonal heterogeneity in utilities (Baltas and Doyle, 2001; Train, 2009).

Depending on whether the independence of irrelevant alternatives (IIA) property is satisfied, the RU models can be classified into IIA models and non-IIA models (Luce, 1959; Ben–Akiva and Lerman, 1985). The IIA assumption requires that the odds of choosing one alternative over the other should be independent of the other alternatives in the choice set. Such a property implies the same degree of substitution effect among the alternatives, which is often considered as restrictive and impractical (Malhotra, 1984; Green and Srinivasan, 1990).

One of the most commonly used IIA model is the conditional logit (CL) model (McFadden, 1974). Under the such a model, the random components are assumed to be independent and follow the type I extreme-value distribution and the values of attributes may differ across individuals and alternatives. Let X_{nj} be the vector of attributes for individual n and alternative j. The systematic utility is assumed to be

* Corresponding author.
E-mail addresses: shl2003@connect.hku.hk (H. Shi), gyin@hku.hk (G. Yin).

http://dx.doi.org/10.1016/j.jocm.2017.07.002
Received 6 November 2016; Received in revised form 7 June 2017; Accepted 12 July 2017
Available online xxxx

$$V_{nj} = X_{nj}^{\top}\beta, \tag{1.1}$$

where the coefficient vector β is the same across alternatives. Let Y_{nj} denote the binary outcome which equals 1 if individual n chooses alternative j and 0 otherwise. If we assume each choice set consists of J alternatives, then the selection probability that individual n chooses alternative j is

$$\Pr\left(Y_{nj}=1\right)=\Pr\left(U_{nj}>U_{nk},k\neq j\right)=\frac{e^{V_{nj}}}{\sum_{k=1}^{J}e^{V_{nk}}}.$$

The IIA models impose a restricted structure to the random components and might be invalid in their applications in the practical context (Tversky, 1972; Currim, 1982). As a result, a variety of non-IIA models are developed, which seek to accommodate various heterogeneous patterns of correlations among the random components. Three main categories of non-IIA models include the generalized extreme value (GEV) model, the probit model and the mixed logit model (Train, 2009). As the most widely used GEV model (Ben–Akiva, 1973; McFadden, 1979), the nested multinomial logit model (McFadden, 1981) partitions the alternatives within a choice set into subsets (called "nests"), where the IIA property only holds for alternatives within a nest, and thus greater substitution effects are implied within a nest than between the nests. The heteroscedastic extreme value model (Steckel and Vanhonacker, 1988; Bhat, 1995) assumes type I extreme-value distribution for the random components but allows their variances to be different across alternatives. Further extensions of the GEV family include the ordered GEV model (Small, 1987), cross-nested logit model (Vovsha, 1997; Bierlaire, 1998; Ben-Akiva and Bierlaire, 1999), paired combinatorial logit model (Chu, 1981, 1989), and generalized logit model (Wen and Koppelman, 2001). The multinomial probit model allows greater flexibility in the pattern of the substitution effects among alternatives: the random components within a choice set are assumed to jointly follow a multivariate normal distribution and their covariance matrices can be arbitrarily specified (Hausman and Wise, 1978; Daganzo, 1979). The mixed logit model (McFadden and Train, 2000) is able to adaptively and flexibly model the choice probabilities over a prespecified mixing distribution. Depending on the prior information on the taste variation among the decision makers, various mixing distributions can be employed, e.g., normal, log-normal, triangular or uniform (Revelt and Train, 1998; Ben–Akiva and Bolduc, 1996; Hensher and Greene, 2001; Train, 2001). A thorough review on the applications of the mixed logit model can be found in Hensher and Greene (2001).

More recently, non-RU based models with focuses on machine learning are proposed as complements to the RU based econometric choice models. Hensher and Ton (2000) explored the use of neural networks to model the commuter mode choice. Karlaftis (2004) adopted the tree-structured classification technique to predict the individual choice decision. Zhang and Xie (2008) applied the support vector machine to model the travel mode choice. Karlaftis and Vlahogianni (2011) provided a general overview on the differences and similarities between the machine learning based models and the conventional econometrics choice models. Reid and Tibshirani (2014) developed a conditional logit model under lasso and elastic net penalties.

One limitation of the conditional logit model in (1.1) is that the systematic part of the utility function is assumed to be linear on X_{nj}. A more flexible way of modeling the utility function is to replace $X_{nj}^{\top}\beta$ in (1.1) by an unknown function $F(X_{nj})$. The probability that an individual n chooses alternative j becomes

$$\Pr\left(Y_{nj}=1\right)=\frac{e^{F\left(X_{nj}\right)}}{\sum_{k=1}^{J}e^{F\left(X_{nk}\right)}}. \tag{1.2}$$

Our goal is to approximate the function $F(\cdot)$ via a machine learning procedure called boosting (Schapire, 1990; Freund, 1995; Freund and Schapire, 1997), which has evolved into one of the most successful and widely used methods for producing accurate predictions in regression and classification problems. The building block of a boosting procedure is a base learner, which can be either an estimator or a classifier for predicting the outcomes. The method fits a base learner repeatedly on the reweighted data, and finally the predictions produced by all the base learners are combined together as the final boosting estimator. The original AdaBoost algorithm is found to be consistently accurate and have slow over-fitting behaviour. To explain its accuracy in model prediction, Schapire et al. (1998) developed a concept called "margin" and studied its relationship to the effectiveness of the boosting algorithm. Breiman (1999) and Friedman et al. (2000) provided a different perspective on boosting, viewing it as an additive stagewise gradient descent algorithm in the functional space. Such a viewpoint has given rise to various extensions of the original AdaBoost algorithm. Friedman (2001) developed a generic framework of the functional gradient descent algorithm and derived variants of the original boosting method by changing the base learner and the loss function. Bühlmann and Yu (2003) proposed a boosting algorithm for minimizing the L_2 loss function with componentwise cubic smoothing spline as its base learner. Li and Luan (2005) applied such a componentwise smoothing spline-based boosting procedure to the proportional hazards model, and Lu and Li (2008) extended the method to a general class of nonlinear transformation models.

A componentwise smoothing spline-based boosting procedure has several desirable properties. First, as a boosting procedure, it has slow over-fitting behaviour. Bühlmann and Yu (2003) showed that the complexity of the fitted boosting estimator only increases by an exponentially diminishing amount as the boosting procedure progresses. Second, it automatically performs variable selection, assigning higher weights to the more important predictors and lower weights to the less important ones, and is particularly useful under high-dimensional settings (Bühlmann, 2003). Third, it leads to a consistent approximation to a function with a fixed finite-dimensional domain (Jiang, 2004; Zhang, 2004; Lugosi and Vayatis, 2004; Zhang and Yu, 2005). Fourth, with smoothing spline as its base learner, it is able to capture the nonlinear predictor effects on the outcomes.

The original AdaBoost algorithm cannot directly model the discrete choice data due to the restriction that the probabilities of all the