Bayesian Variable Selection

B

Oleg Okun
SMARTTECCO, Stockholm, Sweden

INTRODUCTION

Variable selection is an important task in Predictive Analytics as it aims at eliminating redundant or irrelevant variables from a predictive model (either supervised or unsupervised) before this model is deployed in production. When the number of variables exceeds the number of instances, any predictive model will likely overfit the data, implying poor generalization to new, previously unseen instances. Even if for some data sets the number of variables is (much) smaller than the number of instances, some of collected variables may still harm the model performance if left in a data set because these variables may mask (hide) good for prediction variables. Therefore these "harmful" variables need to be discovered and removed.

There are hundreds techniques proposed for variable selection (see, for example, the book of (Liu & Motoda, 2008) entirely devoted to various variable selection methods). The purpose of this chapter is not to present as many of them as possible but concentrate on one type of algorithms, namely Bayesian variable selection (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013). Again, as there are many such algorithms (see, for example, the survey in (O'Hara & Sillanpää, 2009)), we explain the general idea on the example of a particular representative algorithm described in this chapter.

Why Bayesian variable selection? Bayesian variable selection methods come equipped with measures of uncertainty, such as the posterior probability of each model and the variable importance specified by marginal inclusion probabilities. Model uncertainly can be incorporated into prediction through model averaging, which usually improves prediction. Missing data and/or non-Gaussian data distributions are easily handled by Markov Chain Monte Carlo (MCMC) simulations, which are the part of Bayesian variable selection.

Bayesian methods such as "stochastic search variable selection" (George & McCulloch, 1996) have been proposed as alternatives to traditional stepwise variable selection procedures in regression models. Instead of either fixing a regression coefficient at zero or allowing it to be estimated by least squares, as in stepwise procedures, stochastic search variable selection assigns a mixture prior distribution for the given coefficient. Both components of this prior are centered at zero but one with a small variance and the other with a large variance.

In general, there is a vector of regression coefficients β and a vector γ of the same length containing 0/1 indicators, where 1 means a variable is included in a model and 0 implies that a variable is omitted from a model. The classical Bayesian variable selection (George & McCulloch, 1996) thus corresponds to the following model:

1. Mixture ("spike and slab") prior (Mitchell & Beauchamp, 1988) for β_j: $\beta_j \mid \gamma_j \sim \left(1-\gamma_j\right) N\left(0,\tau_j^2\right) + \gamma_j N\left(0,c_j^2\tau_j^2\right)$ where $N\left(\mu,\sigma^2\right)$ is the normal (Gaussian) distribution with the mean μ and variance σ^2. The constant τ_j^2 is small, so that if $\gamma_j = 0$, β_j can be assumed to be 0. The constant c_j^2 is large, so that if $\gamma_j = 1$, β_j can be treated as a non-zero model coefficient.

DOI: 10.4018/978-1-4666-5202-6.ch023

2. The prior for γ_j is a Bernoulli prior: $\gamma_j \sim Bernoulli\left(p_j\right)$, where p_j is the prior probability that the j th variable is included in a model, provided that these probabilities are independent.

As a result, models are influenced by the choice of the priors and their parameters, thus resulting to several variants of the classical model. (O'Hara & Sillanpää, 2009) presented an overview and taxonomy of four types of classical Bayesian variable selection algorithms: indicator model selection, stochastic search variable selection, adaptive shrinkage, and model space approach.

Despite certain differences, these algorithms employ Markov Chain Monte Carlo techniques, such as Gibbs sampling, for fitting Bayesian models.

Algorithms of the first two types use an indicator vector to designate variables included in a model. The model coefficients and indicators can be independent when the priors on the coefficients and indicators are independently placed, or the prior on coefficients can be made dependent on the prior on indicators. The difference between indicator model selection and stochastic search variable selection is in this dependence relationship. Algorithms of the first type assume that values of the prior on coefficients when a variable is omitted do not influence on the posterior, while this is not true for the second type of algorithms.

Algorithms of other two types do not use indicators. Adaptive shrinkage algorithms put the priors on both coefficients and coefficient parameters τ_j^2, leading to shape approximation of the"spike and slab" prior. When there is no evidence in the data for non-zero values, the coefficient prior shrinks towards zero; otherwise, almost no shrinkage occurs. As there are no indicator variables, the problem of knowing which variables are included in the model is decided by comparing the absolute value of a coefficient against threshold.

Unlike the others, model space approach algorithms treat a set of coefficients as a whole and places the priors on the number of variables to be selected and the coefficients of these variables. By choosing a small number of variables, computations are made faster than in other types of algorithms.

Considering variables in groups rather than individually recently paved the way to a new type of Bayesian variable selection algorithms (Carbonetto & Stephens, 2012), (Hernández-Lobato, Hernández-Lobato, & Dupont, 2013), (Quintana & Conti, 2013). Algorithms of this type can be called block feature selection as entire blocks are considered either included in the model or omitted from it. If the indicator variable of a particular group is equal to zero, the model coefficients corresponding to that group are set to zero and the variables of the group are not used for prediction of the targets. On the other hand, if the indicator variable is equal to one, the variables of that particular group are used for prediction and the model coefficients are assumed to be generated from a multivariate distribution.

With multivariate distributions, the posterior can easily become complex and multimodal. Therefore the traditional Markov Chain Monte Carlo methods do not suit well to this case. Instead of them, Variational Bayes (Blei, Jordan, & Paisley, 2012) is used as an alternative. Whereas Monte Carlo techniques provide a numerical approximation to the exact posterior using a set of samples, Variational Bayes provides a locally-optimal, exact analytical solution to an approximation of the posterior.

Although there are many Bayesian variance selection algorithms, we decided to present one algorithm introduced in (Lee, Sha, Dougherty, Vannucci, & Mallick, 2003). The motivation is that the detailed explanation of a single algorithm may be better for understanding the Bayesian variable selection concept than a short presentation of several algorithms.