6th International Conference on Applied Informatics
Eger, Hungary, January 27–31, 2004.

Construction of decision trees using the MCMC algorithm

Ilona Krasznahorkay

Department of Applied Mathematics and Probability Theory,
Institute of Informatics,
University of Debrecen
e-mail: krasznil@inf.unideb.hu

Abstract

The construction of decision trees is a commonly used and easily applied way of supervised learning. The aim is the prediction of one or more target variable on the basis of many predictor variables. This technique divides the field of predictors along a predictor variable one after another. The goal is that the target variable should be more and more homogeneous along the resulting partition. I modified the CART algorithm developed by Breiman et al. [1], which aims for the minimizing of a concave risk function defined on the partitions generated by the trees. I improved this algorithm with a stochastic search on the set of decision trees applying the Markov Chain Monte Carlo method. It was first proposed in a Bayesian framework by Chipman et al. [2]. By empirical experience finding the optimal tree this technique is much more effective than the former deterministic methods.

Key Words and Phrases: Decision trees, MCMC-method, Metropolis-Hastings algorithm

1. Introduction

The main problem of the supervised learning is the following. If $\mathbf{X} = (X_1, \ldots, X_d)$ is the vector of predictor variables, and Y is the target variable, then we are looking for the function that has values the nearest to Y.

To measure this distance we use the method of least squares:

$$\mathbb{E}(Y - f(\mathbf{X}))^2 \to \min, \qquad f : \mathcal{X} \to \mathbb{R}.$$

The solution of this problem is the conditional expectation

$$\pi(\mathbf{x}) := \mathbb{E}(Y|\mathbf{X} = \mathbf{x}), \qquad x \in \mathcal{X}.$$

If the values of Y are discrete, then we speak about classification problem, and if they are continuous, we speak about regression problem. In this paper we will concentrate to the case of a binary discrete target variable. These binary valued discrete problems are also important in our lives. Just think about the case when a bank must decide whether the costumer is a good debtor or not.

In Section 2 a short overwiev is presented on the decision trees. In Section 3 Metropolis-Hastings algorithm is introduced and applied for decision trees. Section 4 gives some necessary tools to the proof of the main theorem, which remains for Section 5.

2. Decision trees

There are different methods to construct a decision tree, for example CHAID, AID, C4.5, and C5 (see [4]). I used the CART algorithm developed by Breiman et al. [1], which is a deterministic method. Decision trees have many advantages over the other nonlinear methods, for example their easy interpretation.

Definition 2.1. By a decision tree we mean a pair of $\mathbf{T} = (T, \tau)$, where T is a binary tree, and $\tau : T \to \{1, \dots, d\} \times \mathbb{R}$ is a mapping.

If s is a node of the tree, than $\tau(s) = (k, \alpha), 1 \leq k \leq d, \alpha \in \mathbb{R}$, means that we cut the s node in the kth variable at the α value.

Breiman's book's main point is defining a risk function with which we can measure the goodness of a decision tree. Let's denote $\ell(T)$ the number of the tree's leaves, $\mathcal{X}_1, \dots, \mathcal{X}_{\ell(T)}$ the partition of the field of predictor variables defined by the binary tree. Let φ be a concave, symmetric loss function on $[0, 1]$. Define

$$vol(\mathcal{X}_k) := \int_{\mathcal{X}_k} \lambda(\mathbf{x}) d\mathbf{x},$$
$$\pi_k := \frac{1}{vol(\mathcal{X}_k)} \int_{\mathcal{X}_k} \pi(\mathbf{x}) \lambda(\mathbf{x}) d\mathbf{x},$$

$k = 1, \dots, \ell(T)$, where λ is the density function of the predictor variables.

Definition 2.2. A function $R : \mathbf{T} \to [0, 1]$ is called risk function, if

$$R(\mathbf{T}) := \sum_{k=1}^{\ell(T)} \varphi(\pi_k) vol(\mathcal{X}_k). \tag{1}$$

The main advantage of risk function of this form is that we can compute it recursively.

In this case we suppose that λ and π_k are known. If we have a sample $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, than we must estimate $vol(\mathcal{X}_k)$ and π_k. For example:

$$\widehat{vol}(\mathcal{X}_k) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i \in \mathcal{X}_k) \qquad \hat{\pi}_k = \frac{\sum_{i=1}^{n} I(y_i = 1, \mathbf{x}_i \in \mathcal{X}_k)}{\sum_{i=1}^{n} I(\mathbf{x}_i \in \mathcal{X}_k)}.$$

We get another in practice very important risk function if we punish the too big trees in direct ratio to the number of their leaves:

$$R_\alpha(T) := R(T) + \alpha\ell(T), \qquad \alpha > 0. \tag{2}$$

The φ function can be chosen differently, the most often used are misclassification, Gini-index and entropy (see [1], Chapter 4). The concaveness of φ guarantees that the risk can be minimized.

3. The Metropolis–Hastings algorithm

There can be cases, when although we always choose the best cut, at the end we do not get the optimal tree. So the global optimum - as in many other cases - is not the consequence of the local optimum. That is why I integrated the Metropolis-Hastings algorithm to find the optimal tree.

Let $q(T'|T)$, $T, T' \in \mathbf{T}$ be an arbitrary transition kernel. Define α as the following:

$$\alpha(T, T') := \min\left\{1, \frac{\pi(T')q(T|T')}{\pi(T)q(T'|T)}\right\}, \tag{3}$$

where the second term of the minimum denotes the Hastings quotient. Create a Markov chain $\xi_1, \ldots, \xi_n, \ldots$. If $\xi_i = T$ then let

$$\xi_{i+1} = \begin{cases} T' & \text{with probability } \alpha(T, T'), \\ T & \text{otherwise.} \end{cases}$$

Then $\pi(T)$ will be a stationary distribution of the chain (see [9]).

Consider the following very simple case. Let $\mathbf{X}$ be a continuous predictor variable with values in the $[a, b]$ interval. Construct the trees with the following transition probabilities. Denote e the elementary tree, i.e. e is a tree with two leaves. Let's denote $T \to T'$ that T' can be reached from T with adding an elementary tree or deleting the last added elementary tree. Suppose that all these cases have the same probability $1/(\ell(T) + 1)$. We can now define the transition probabilities of the Metropolis-Hastings algorithm as the following:

$$q(T'|T) = \begin{cases} (\ell(T) + 1)^{-1}, & \text{if } T \to T', \\ 0, & \text{otherwise.} \end{cases}$$

4. General state space Markov chains

We should prove that the Markov chain constructed in Section 3 converges to the unique stationary distribution $\pi(T)$, and the convergence is as fast as possible. Here I explain just the tools we will need later, all the other definitions and proofs can be found in the book of Meyn and Tweedie [6].

Let $(E, \mathcal{E})$ be a measurable space, where E is an abstract set and $\mathcal{E}$ is a countably generated sigma algebra. The distribution of a time-homogenous Markov chain $\{\xi_n, n \in \mathbb{N}\}$ on the state-space E is specified by its initial distribution and its transition kernel $P(x, A)$.

Definition 4.1. An ergodic Markov chain with invariant distribution π is geometrically ergodic if there exists a function $M : E \to \mathbb{R}_+$ such that $\int_E M(x)\pi(dx) < \infty$ and a positive constant $r < 1$ such that

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq M(x)r^n$$

for all $x \in E$ and $n \in \mathbb{N}$.

Theorem 4.1. *Suppose that an ergodic Markov chain has the property that for a function $V : E \to [1, \infty)$, some constant $\beta > 0$ and $b < \infty$ and a small set $C \subset E$*

$$PV(x) - V(x) \leq -\beta V(x) + bI_C(x)$$

for all $x \in E$. Then the chain is geometrically ergodic.

5. Risk fuction with penalty

Define now the target invariant distribution π_α as the following:

$$\pi_\alpha(T) = \exp\{-(R(T) + \alpha\ell(T))\}, \qquad T \in \mathbf{T}.$$

So by using the risk function with penalty, although that by growing the tree the risk gets lower and lower, the optimal tree will be finite.

We use this fuction instead of the one suggested by Breiman et al., because so we are to find the tree with the greatest probability (in the other case we should have found the tree with lowest probability).

We can give a distribution on the set of decision trees in direct ratio to π_α. To this first we need the number of binary trees with ℓ leaves (see Cormen et al. [3]):

$$b_l \approx \frac{1}{4\sqrt{\pi}} \frac{4^l}{l^{3/2}}.$$

We need:

$$\sum_T \pi_\alpha(T) < \infty.$$

One can see that π_α can be normalized to be a distribution, if and only if $\alpha > \ln 4$. Because this normalizing factor disappears in the definition of the Hastings quotient, we are going to speak about the π_α distribution. Let q be defined as in Section 3.

By the definition of π_α and q we get that the Hastings quotient is:

$$h(T,T') = \exp\{-(R(T)-R(T'))\} \cdot \begin{cases} e^{-\alpha}\dfrac{\ell(T)+2}{\ell(T)+1}, & \text{if } T' = T \cup_s e, \\ \\ e^{\alpha}\dfrac{\ell(T)}{\ell(T)+1}, & \text{if } T' = T\backslash_s e. \end{cases}$$

We can state the main theorem.

Theorem 5.1. *The $\xi_1, \xi_2, \ldots$ Markov chain defined by the MH algorithm is geometrically ergodic on the set of binary trees with π_α stationary distribution, i.e. the chain converges exponentially to the stationary distribution for all $\alpha > \ln 4$.*

Proof. First we prove the ergodicity. It can be made using the definition of irreducibility, strong aperiodicity and Harris recurrence.

Then we construct a suitable energy function V with that the drift criterion of Theorem 4.1 is satisfied. Search the V function in the following form:

$$V(T) = \sum_{s \in leaves} f(T,s),$$

i.e. let V be the sum of the values given to the leaves of the tree.

If the T tree has enough leaves then the π_α function doesn't change very much after adding an elementary tree to T. So $\pi_\alpha(T')/\pi_\alpha(T)$ is near to 1. So the Hastings quotient is between κ and 1, where κ is a constant, and for great trees it is about $e^{-\alpha}$. In that case when we remove the last added elementary tree, the Hasting quotient is greater than 1, so $\alpha(T,T') = 1$. So we get

$$\begin{aligned} PV(T) - V(T) =& \frac{1}{\ell(T)+1} \sum_{T'=T\cup e} h(T,T')(V(T') - V(T)) + \\ &+ \frac{1}{\ell(T)+1}(V(T\backslash e) - V(T)). \end{aligned}$$

If V is bounded than the second term is small when the number of leaves is great enough. For the first term we have:

$$\begin{aligned} &\frac{\ell(T)-1}{\ell(T)+1} \sum_{s\in\tilde{T}} h(T, T\cup e) f(T,s) \left[\frac{f(T\cup e, s)}{f(T,s)} - 1\right] + \\ &\quad + \frac{1}{\ell(T)+1} \sum_{s\in\tilde{T}} h(T, T\cup_s e) f(T,s) \left[\frac{f(T\cup_s e, s_B) + f(T\cup_s e, s_J)}{f(T,s)} - 1\right], \quad (4) \end{aligned}$$

where s_B and s_J are those leaves of the tree $T \cup_s e$ which were the leaves of the last added elementary tree. From this we can see that we should choose the $f(T\cup e, s)/f(T,s)$ quotient suitable.

Define the V function recursively. Let f be 1 on the leaves of the elementary tree. If we have defined V for a tree T than to $T \cup e$ define the following: Every value of leaves they stay after the addition let be multiplied by $\gamma \in (0, 1)$, and to the new leaves give 1. By pruning we get the following state back. On Figure 1 we can see the recursion of V.

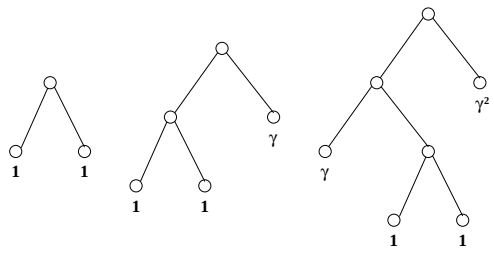

Figure 1: The recursion of the V function

It's easy to see, that $V(T) > 1$ for all tree, and V is bounded on the set of binary trees since $V(T) < 2/(1-\gamma)$. The second term of expression (4) tends to 0 if the number of the leaves tends to infinity. With multiplying by -1 we get

$$\frac{\ell(T)-1}{\ell(T)+1} \sum_{s \in \tilde{T}} h(T, T \cup e) f(T, s)(1-\gamma) \geq \frac{\kappa(1-\gamma)}{3} V(T),$$

if the tree has minimally two leaves. So for trees having sufficiently large number of leaves the drift criterion is satisfied with $\beta = \kappa(1-\gamma)/3$. The trees having less leaves than a given number form a small set. Thus the proof is finished.

6. Continuous predictor variables

Suppose now that the $(X_1, \ldots, X_d)$ predictor variables are continuous, that is the $\lambda(x) = (\lambda_1(x_1), \ldots, \lambda_d(x_d))$ density function exists. Consider the case of the punished risk function. As before we can see that π_α can be normalized to be a density. Now define q by the following way:

$$q(T'|T) = \begin{cases} \dfrac{\lambda_j(x)}{(\ell(T)+1)d}, & \text{if } T' = T \cup_s e, \tau'(s) = (j, x); \\ \dfrac{1}{\ell(T)+1}, & \text{if } T' = T \backslash_s e; \\ 0, & \text{otherwise,} \end{cases}$$

where τ defined according to the Definition 2.1., and we choose from the possibilities with equal probability. We can state the following theorem.

Theorem 6.1. *Suppose that the density functions $\lambda_j, j = 1, \ldots, d$, of the predictor variables are continuous. Then the Markov chain $\xi_1, \xi_2, \ldots$ with the above defined transition kernels on the set of decision trees is geometrically ergodic with stationary distribution π_α, for all $\alpha > \ln 4$.*

The proof is similar to the discrete case.

7. Discussion

The method suggested by Breiman et al. is not always effective and fast enough. So I integrated stochastic search to find the optimal tree. I gave the transition kernels in different cases and proved the geometric ergodicity of the Markov chains constructed this way.

It would be important to construct the chain when we have just a sample, and see how much the estimated tree differs from the theoretical one. Other questions are the mixing rates, i.e. how fast we get to the stationary distribution.

References

[1] Breiman, L., Friedman, J.H., Olsen, A.O., Stone, C.J.: *Classification and Regression Trees.* Wadsworth International Group, 1984.

[2] Chipman, H.A., George, E.I., McCulloch, R.E.: Bayesian CART Model Search. *JASA*, Vol. 93, Num. 443, (1998), 935-960.

[3] Cormen, T.H., Leierson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms.* MIT Press, 2001.

[4] Hastie, T., Tibshirani, R., Friedman, J.H.: *The Elements of Statistical Learning.* Springer, 2001.

[5] Jerrum, M., Sinclair, A.: The Markov chain Monte Carlo method: An approach to approximate counting and integration. *Approximation Algorithm for NP–hard Problems*, Dorit Hochbaum ed., PSW, 1996.

[6] Meyn, S.P., Tweedie, R.L.: *Markov Chains and Stochastic Stability.* Springer, 1993.

[7] Roberts, G.O.: *Markov chain concepts related to sampling algorithms.* In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. 45-57. Chapman & Hall/CRC, 1996.

[8] Tierney, L.: *Introduction to general state-space Markov Chain Theory.* In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. 2-18. Chapman & Hall/CRC, 1996.

[9] Tierney, L.: Markov chains for exploring posterior ditributions (with discussion). *Ann. Statist.*, Vol. 22, (1994), 1701-1762.

Postal address

Ilona Krasznahorkay
Department of Applied Mathematics and Probability Theory
University of Debrecen
Pf. 12
4010, Debrecen
Hungary