Ensembling Treed Basis Regressions

Sameer K. Deshpande

May 15, 2019

1 General Setup

Suppose we observe time series data for n individuals sampled irregularly in the intervals $[0, T]$. For individual i, we observe $\mathbf{x}_i \in \mathbb{R}^p$ a vector of covariates and also $y_{i,J}$ at time $t_{i,j} \in [0, T]$. We model

$$y_{i,j} = f(\mathbf{x}_i, t_{i,j}) + \sigma\varepsilon_{i,j}$$

where $\varepsilon_{i,j} \sim N(0, 1)$. We express the unknown function f as the sum of m "functional regression trees."

To set our notation, let T denote a binary decision tree partitioning $\mathbb{R}^p$ that consists of a collection of interior nodes and $L(T)$ terminal or *leaf* nodes. We associate an axis-aligned decision rule of the form $\{x_j < c\}$ or $\{x_j \geq c\}$ to each internal (i.e. non-leaf) node of T. In this way, T defines a partition of $\mathbb{R}^p$ into $L(T)$ rectangular cells, corresponding to the leaves of T, and we let $\ell(\mathbf{x}, T)$ be the function that returns the index of the cell containing the point $\mathbf{x}$. A *functional regression tree* (T, φ, Θ) consists of a decision tree T, a fixed feature map $\varphi : [0, T] \to \mathbb{R}^D$ and a collection $M = \left\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{L(T)}\right\}$ where each $\boldsymbol{\theta}_\ell \in \mathbb{R}^D$. We define the evaluation function

$$g(\mathbf{x}, t; T, \varphi, \Theta) = \varphi(t)^\top \boldsymbol{\theta}_{\ell(\mathbf{x};T)}.$$

For a given tree T and leaf index ℓ, let $I(\ell; T)$ be the collection of indices i such that $\ell(\mathbf{x}_i; T) = \ell$.

For our purposes, we will consider the cosine basis function where $\varphi = (\phi_1, \ldots, \phi_D)$ where

$\phi_d(t) = \left(\frac{2}{T}\right)^{\frac{1}{2}} \cos\left(\frac{d\pi t}{T}\right)$. We approximate

$$f(\mathbf{x}, t) = \sum_{m=1}^{M} g(\mathbf{x}, t; T_m, \phi, \Theta_m).$$

Upon placing a prior on the regression trees and updating it with the observed data, we can induce a posterior distribution over f.

The prior over regression trees consists of two parts: a decision tree prior and the conditional prior of $M \mid T$. For the first part, we use exactly the same prior as Chipman et al. (2010): the probability that node at depth d is internal is $\alpha(1+d)^{-\beta}$ and conditional on a node being internal, the splitting rule is picked uniformly from the set of all available splitting rules. Conditional on the decision tree T, the associated leaf parameters are modeled as i.i.d. $N(0, \sigma_\theta^2 D \Lambda_\theta)$ where $\Lambda_\theta = \text{diag}(e^{-\gamma c_d})$ We follow Lenk (1999) and take $c_d = d$ (the "algebraic smoother") or $\log d$ (the "geometric smoother"). The parameter γ controls how quickly the Fourier coefficients in $\boldsymbol{\theta}_\ell$ decay to zero, implicitly controlling the smoothness of $g(\mathbf{x}, t; T, M)$ as a function of t.

2 A Backfitting Strategy

We now briefly summarize how to extend Chipman et al. (2010)s backfitting strategy to this functional setting. We have m regression trees $(T_1, \Theta_1), \ldots, (T_M, \Theta_M)$, which we update one at a time, holding all else fixed. Let (T_{-m}, Θ_{-m}) be the set of all $M - 1$ regression trees besides (T_m, Θ_m) and define

$$r_{i,j,m} = y_{i,j} - \sum_{m' \neq m} g(\mathbf{x}_i, t_{i,j}; T_{m'}, \Theta_{m'}).$$

Observe that

$$\pi(T_m, \Theta_m \mid \mathbf{y}, T_{-m}, \Theta_{-m}, \sigma) \propto \pi(T_m) \prod_{\ell=1}^{L} \prod_{i \in I(\ell;T)} \prod_{j=1}^{n_i} \exp\left\{-\frac{(r_{i,j,m} - \varphi(t_{i,j})^\top \boldsymbol{\theta}_\ell)^2}{2\sigma^2}\right\}$$
$$\times \prod_{\ell=1}^{L} \sigma_\theta^{-D} |\Lambda_\theta|^{-\frac{1}{2}} \exp\left\{-\frac{\boldsymbol{\theta}_\ell^\top \Lambda_\theta^{-1} \boldsymbol{\theta}_\ell}{2\sigma^2}\right\}$$

From here, we immediately see that the $\boldsymbol{\theta}_\ell$'s are independent *a posteriori* with $\boldsymbol{\theta}_\ell \sim N(M_\ell, V_\ell)$ where

$$V_\ell = \left[\sigma_\theta^{-2}\Lambda_\theta^{-1} + \sigma^{-2} \sum_{i \in I(\ell;T)} \sum_{j=1}^{n_i} \varphi(t_{i,j})\varphi(t_{i,j})^\top\right]^{-1}$$

$$M_\ell = V_\ell \left[\sigma^{-2} \sum_{i \in I(\ell;T)} \sum_{j=1}^{n_i} r_{i,j,m}\phi(t_{i,j})\right].$$

Marginalizing over Θ_m, we have the following conditional posterior probability over the decision tree T

$$\pi(T \mid T_{-m}, \Theta_{-m}, \mathbf{y}, \sigma^2) \propto \pi(T) \prod_{\ell=1}^{L} \sigma_\theta^{-D} |\Lambda|^{-\frac{1}{2}} |V_\ell|^{\frac{1}{2}} \exp\left\{\frac{1}{2} M_\ell^\top V_\ell^{-1} M_\ell\right\}$$

To carry out the update $(T, \Theta) \to (T^*, \Theta^*)$, we first propose a new tree T_{prop} by either growing T at one leaf node or by pruning two leafs back to their common parent node. We accept this proposal with probability

$$\alpha(T_{prop}, T) = \min\left\{1, \frac{q(T, T_{prop})\pi(T_{prop} \mid \mathbf{y}, T_{-m}, \Theta_{-m}, \sigma)}{q(T_{prop}, T)\pi(T \mid \mathbf{y}, T_{-m}, \Theta_{-m}, \sigma)}\right\}.$$

If we accept the proposal we set $T^* = T_{prop}$; otherwise we set $T^* = T$. We then update Θ^* conditionally on T^* by making conjugate normal draws. It should be noted that to carry out this update requires computing the inverse and determinant of $L(T) + 1$ covariance matrices V_ℓ. So the computational cost of each regression tree update is $O(L(T)D^3)$. [skd]: there's probability a factor of n lurking in there somewhere because we need to sum over $\varphi(t_{i,j})\varphi(t_{i,j})^\top$

References

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298.

Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a Fourier representation. *Journal of the Royal Statistical Society (Series B)*, 61(4):863 – 879.