Notes on BART

Adam Kapelner

April 15, 2011

1 Bayesian Additive Regression Trees (BART) Model

1.1 The Basics

We use the same notation as above where T_j represents a binary tree and $M_j = \{\mu_{1j}, \ldots, \mu_{bj}\}$ represents the expected values of y given the data or the most likely class given the data. We now introduce the sum of m trees model as follows where g represents the tree function:

$$y_i = \left(\sum_{j=1}^{m} g(\boldsymbol{x}_i, T_j, M_j)\right) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0,\, \sigma^2\right)$$

Since the tree function evaluates $\boldsymbol{x}$ by a bunch of if-else statements and arrives at a leaf, it is clear that:

$$y_i = \left(\sum_{j=1}^{m} \mu_{ij}\right) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0,\, \sigma^2\right)$$

Here each tree only explains a piece of the expected value or class of y. Now, given a fixed number of trees m, this model is completely parameterized by $\{\{T_1, M_1\}, \ldots, \{T_m, M_m\}, \sigma^2\}$. We choose a fixed m in this implementation.

Now we impose a prior on all the parameters. Keep in mind, we need to keep individual tree effects small.

1.2 The prior

Similar to the prior for the Bayesian Decision trees, the prior has the following symmetry (all probabilities are given the data $\boldsymbol{X}$ and we leave it out for notational simplicity)

$$\begin{aligned}\mathbb{P}\left(\{T_1, M_1\}, \ldots, \{T_m, M_m\}, \sigma^2\right) &= \left[\prod_{j=1}^{m} \mathbb{P}\left(T_j, M_j\right)\right] \mathbb{P}\left(\sigma^2\right) = \left(\prod_{j=1}^{m} \mathbb{P}\left(M_j | T_j\right) \mathbb{P}\left(T_j\right)\right) \mathbb{P}\left(\sigma^2\right) \\ &= \left(\prod_{j=1}^{m} \left(\prod_{i=1}^{b} \mathbb{P}\left(\mu_{ij} | T_j\right)\right) \mathbb{P}\left(T_j\right)\right) \mathbb{P}\left(\sigma^2\right)\end{aligned}$$

Let's look closely at each prior.

1.2.1 $\mathbb{P}\left(\mu_{ij} \mid T_j\right)$

First of all, we rescale each response as follows:

$$y_i' = \frac{y_i - y_{min}}{y_{max}} - 0.5$$

which means that the response is constrained to live within $[-0.5, 0.5]$. We want to design a prior that assigns substantial probability to this region. The default pick is normal, centered at 0, with some variance, call it σ_μ^2:

$$\mu_{1j}, \ldots, \mu_{b_j j} \overset{iid}{\sim} \mathcal{N}\left(0,\, \sigma_\mu^2\right) \quad \textit{i.e.} \text{ for all leaves on all trees} \tag{1}$$

How do we pick this variance σ_μ^2? First, select a value a percentage of the distribution you want to cover, let's say 95%, then *a la* Stat 101, calculate the inverse z value that corresponding to that probability being within $[-z, z]$, then solve for σ_μ:

$$z = \frac{X - \mu}{\sigma} \quad \Rightarrow \quad 1.96 = \frac{0.5 - 0}{\sigma_\mu} \quad \Rightarrow \quad \sigma_\mu^2 = \left(\frac{0.5}{1.96}\right)^2 = 0.255^2$$

This would be the case if there were one tree. Because there are many trees, we want the variance to break up this way:

$$\mathbb{V}\text{ar}\left[Y_i\right] = \mathbb{V}\text{ar}\left[g_1\right] + \ldots + \mathbb{V}\text{ar}\left[g_m\right] = m\sigma_\mu^2 \quad \Rightarrow \quad 1.96 = \frac{0.5 - 0}{\sigma_\mu \sqrt{m}} \quad \Rightarrow \quad \sigma_\mu^2 = \left(\frac{0.5}{1.96\sqrt{m}}\right)^2$$

1.2.2 $\mathbb{P}\left(\sigma^2\right)$

Remember σ^2 is the variance of the normally-distributed $\overset{iid}{\sim}$ errors centered at 0. The standard conjugate normal prior on the variance is the inverse gamma:

$$\mathbb{P}\left(\sigma^2\right) = \text{InvGamma}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$$

We use a naive point estimate of σ^2 by calculating just the sample standard deviation of the transformed responses, $s_{y'}^2$. Now we pick a coverage percentage similar to before, let's say $q = 90\%$. So we set:

$$0.90 = \mathbb{P}\left(\sigma^2 \leq s_{y'}^2\right)$$

We then pick a $\nu \in [3, 10]$ to get an appropriate shape, and then we just use a quick grid search to calculated λ. In my implementation, I chose a $\nu = 3$.

1.2.3 $\mathbb{P}\left(T_j\right)$

This is the exact same as the prior trees in CGM98.

1.3 Sampling from the posterior

We create a Gibbs sampler which samples for each tree and then σ^2. For the jth tree, we're sampling, denote all other trees as (j). The posterior sampling for each tree looks like:

$$T_j, M_j \mid T_{(j)}, M_{(j)}, \boldsymbol{y}, \sigma^2$$

Why do T_j, M_j depend on $T_{(j)}, M_{(j)}$? Because those other trees fit part of the data. Let's subtract out their fit and define the residuals below:

$$\begin{aligned}\boldsymbol{R}_j &:= \boldsymbol{y} - \sum_{k \neq j} g_k \\ &= \left(\sum_k g_k + \boldsymbol{\epsilon}\right) - \sum_{k \neq j} g_k \\ &= g_j + \boldsymbol{\epsilon}\end{aligned}$$

We can immediately find the distribution of the residuals:

$$\boldsymbol{R}_j = g_j + \boldsymbol{\epsilon} \quad \Rightarrow \quad \boldsymbol{R}_j \sim \mathcal{N}\left(\begin{bmatrix}\mu_{k_1 j} \\ \vdots \\ \mu_{k_n j}\end{bmatrix}, \left(\sigma_\mu^2 + \sigma^2\right)\boldsymbol{I}_n\right) \quad \text{(via independence)} \tag{2}$$

Note that $k_1, \ldots, k_n$ represent the indices of the leaves that correspond to where $\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$ get classified to. $\boldsymbol{R}_j$ encapsulates the dependencies of $T_{(j)}, M_{(j)}, \boldsymbol{y}$ so now the sampling for each tree becomes:

$$T_j, M_j \mid \boldsymbol{R}_j, \sigma^2$$

We can further split this up below:

$$\mathbb{P}\left(T_j, M_j \mid \boldsymbol{R}_j, \sigma^2\right) = \mathbb{P}\left(T_j \mid M_j, \boldsymbol{R}_j, \sigma^2\right)\mathbb{P}\left(M_j \mid \boldsymbol{R}_j, \sigma^2\right) \tag{3}$$

1.3.1 Sampling M_j

A sampling of M_j is actually b samplings — one for each of the b leaves. They are all equivalent. Based on equation 3, we have:

$$\begin{aligned}&\mu_{1j} \mid \boldsymbol{R}_j, \sigma^2 \\ &\mu_{2j} \mid \boldsymbol{R}_j, \sigma^2 \\ &\vdots \\ &\mu_{bj} \mid \boldsymbol{R}_j, \sigma^2\end{aligned}$$

What is the sampling distribution? Using the Bayesian machinery we have:

$$\begin{aligned}
\mathbb{P}\left(\mu_{ij} \mid \boldsymbol{R}_j, \sigma^2\right) &\propto \mathcal{L}\left(\boldsymbol{R}_j, \sigma^2; \mu_{ij}\right)\mathbb{P}\left(\mu_{ij}\right)\\
\mathbb{P}\left(\mu_{ij} \mid \boldsymbol{R}_j, \sigma^2\right) &\propto \mathcal{L}\left(\boldsymbol{R}_j; \mu_{ij}, \sigma^2\right)\underbrace{\mathcal{L}\left(\sigma^2; \mu_{ij}\right)}_{\text{constant}}\mathbb{P}\left(\mu_{ij}\right)\\
\mathbb{P}\left(\mu_{ij} \mid \boldsymbol{R}_j, \sigma^2\right) &\propto \underbrace{\mathcal{L}\left(\boldsymbol{R}_j; \mu_{ij}, \sigma^2\right)}_{\text{likelihood}}\underbrace{\mathbb{P}\left(\mu_{ij}\right)}_{\text{prior}}
\end{aligned}$$

This prior is just a sampling from the prior for leaf values (see equation 1). For the likelihood, we know the distribution of $\boldsymbol{R}_j$ from equation 2 so now we can compute the posterior via the nice conjugate properties:

$$\begin{aligned}
&\propto \mathcal{L}\left(\boldsymbol{R}_j; \mu_{ij}, \sigma^2\right)\mathbb{P}\left(\mu_{ij}\right)\\
&= \left(\frac{1}{\sqrt{2\pi(\sigma_\mu^2+\sigma^2)}}\exp\left(-\frac{1}{2\left(\sigma_\mu^2+\sigma^2\right)}\left(\boldsymbol{R}_j-\mu_{ij}\right)^2\right)\right)\left(\frac{1}{\sqrt{2\pi(\sigma_\mu^2)}}\exp\left(-\frac{1}{2\sigma_\mu^2}\mu_{ij}^2\right)\right)\\
&\propto \exp\left(-\frac{1}{2\left(\sigma_\mu^2+\sigma^2\right)}\left(\boldsymbol{R}_j-\mu_{ij}\right)^2\right)\exp\left(-\frac{1}{2\sigma_\mu^2}\mu_{ij}^2\right)\\
&= \exp\left(-\frac{1}{2\left(\sigma_\mu^2+\sigma^2\right)}\left(\underbrace{\boldsymbol{R}_j^2}_{\text{const}}-2\boldsymbol{R}_j\mu_{ij}+\mu_{ij}^2\right)\right)\exp\left(-\frac{1}{2\sigma_\mu^2}\mu_{ij}^2\right)\\
&\propto \exp\left(\underbrace{\frac{\boldsymbol{R}_j}{\sigma_\mu^2+\sigma^2}}_{B}\mu_{ij}\underbrace{-\left(\frac{1}{2\left(\sigma_\mu^2+\sigma^2\right)}+\frac{1}{2\sigma_\mu^2}\right)}_{A}\mu_{ij}^2\right)=\exp\left(B\mu_{ij}+A\mu_{ij}^2\right)\\
&= \exp\left(A\left(\frac{B}{A}\mu_{ij}+\mu_{ij}^2\right)\right)\\
&\propto \exp\left(A\left(-\frac{B}{2A}-\mu_{ij}\right)^2\right) \quad \text{(completed the square)}\\
&\propto \mathcal{N}\left(-\frac{B}{2A}, -\frac{1}{2A}\right)=\mathcal{N}\left(\frac{\boldsymbol{R}_j}{2+\frac{\sigma^2}{\sigma_\mu^2}}, \frac{\sigma_\mu^2+\sigma^2}{2+\frac{\sigma^2}{\sigma_\mu^2}}\right)
\end{aligned}$$

But $\boldsymbol{R}_j$ is a vector? Would this mean that you build the tree, see which R_{ij}'s get classified into certain leaves, and then pick a random R_{ij} for each of the b leaves, then sample from the posterior based on the above to assign the b μ_{ij}'s?

1.3.2 Sampling T_j

Let's examine $T_j \mid M_j, \boldsymbol{R}_j, \sigma^2$ from equation 3. We know via the standard Bayes machinery that:

$$\mathbb{P}\left(T_j \mid M_j, \boldsymbol{R}_j, \sigma^2\right) \quad \propto \quad \underbrace{\mathbb{P}\left(\boldsymbol{R}_j \mid T_j, M_j, \sigma^2\right)}_{\text{likelihood}} \underbrace{\mathbb{P}\left(T_j\right)}_{\text{prior}}$$

And we can explore the posterior using the Metropolis-Hastings iteration in CGM98 using $\boldsymbol{R}_j$ as the response. Hence the sampling is *one* iteration of the M-H algorithm.

In order to do this, we need to find the likelihood. Recall the mean-shift model in CGM98 p939 section 4.1:

$$\begin{aligned} \mu_1, \ldots, \mu_b \mid \sigma^2, T &\overset{iid}{\sim} \mathcal{N}\left(\bar{\mu}, \frac{\sigma^2}{a}\right) \\ \sigma^2 \mid T &\overset{iid}{\sim} \text{InvGamma}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right) \end{aligned}$$

Under this prior, both σ^2 and the μ_i's were margined out and a likelihood could be computed (up to a norming constant) as a function of the y_i's, their locations among the leaves, and the four hyperparameters, $\bar{\mu}, a, \nu, \lambda$.

What has changed in the BART implementation? For starters, the leaves have to account for both variance in the leaf values and all the residual error left over. Additionally, with the transformation of the $\boldsymbol{y}$ response because we assume the distribution is centered at zero which is a big win:

$$\mu_{1j}, \ldots, \mu_{b_j j} \mid \sigma^2, \sigma^2_\mu, T_j \quad \overset{iid}{\sim} \quad \mathcal{N}\left(0, \, \sigma^2_\mu + \sigma^2\right)$$

Now, we don't have to worry about the prior on the terms in the variance because they're given exogenously. Hence, forget about all the hyperparameters. Our goal now is to obtain just $\boldsymbol{R}_j | T_j$ by margining out:

$$\mathbb{P}\left(\boldsymbol{R}_j \mid T_j, \sigma^2, \sigma^2_\mu\right) \propto \underbrace{\int \ldots \int}_{\text{for the } M_j} \mathbb{P}\left(\boldsymbol{R}_j \mid T_j, M_j, \sigma^2, \sigma^2_\mu\right) \mathbb{P}\left(M_j \mid T_j, \sigma^2, \sigma^2_\mu\right) dM_j$$

1.3.3 Sampling σ^2

The sampling for σ^2 is just a draw from the inverse gamma distribution.