BART from the ground up p266-277

Assume:

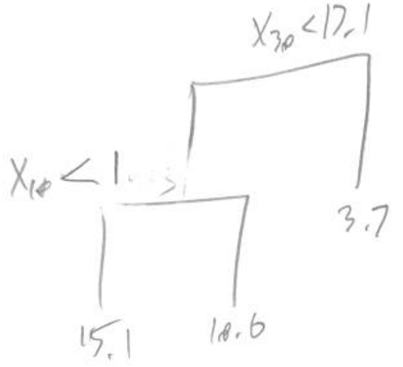
$\vec{y} = f(X) + \vec{\varepsilon}$, $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ comes to Bayesian

$X_{n \times p}$ col's real valued or categorical, $\vec{y} \in \mathbb{R}^n$ (for now)

We approximate f by sum of m trees model

$$\vec{Y} = \sum_{t=1}^{m} g_t(X) + \vec{\varepsilon}$$

Each g_t is a tree w/ binary splits just like the CART setup:



Note how the g_t's are identified by structure and split rules, which we call T_t, and a collection of rules assoc with each component of X which we denote:

$$M_t = \langle \mu_{t1}, \mu_{t2}, \ldots, \mu_{tb_t} \rangle$$

where b_t is the number of nodes in the t^{th} tree

$$\vec{Y} = \sum g_t(X \mid T_t, M_t) + \vec{\varepsilon}$$

Sum of trees: not majority voting.

Goal: to be able to compute $P(\hat{Y} \mid x^*)$, thereby of course make predictions

$$\hat{y} = E[\hat{Y} \mid x^*]$$

Same as regression

We employ a Bayesian setup: $P(\overbrace{T_1, M_1, \ldots, T_m, M_m, \sigma^2}^{\vec{\theta}} \mid y, X) \propto \underbrace{P(y \mid T_1, M_1, \ldots, T_m, M_m, \sigma^2, X)}_{\text{Lik}} \; \underbrace{P(T_1, M_1, \ldots, T_m, M_m, \sigma^2 \mid X)}_{\text{Prior}}$

Since we always condition on X (fixed design), we're going to suppress this notation from now on.

Let's look at the prior first

$$P(T_1, M_1, \ldots, T_m, M_m, \sigma^2) = P(T_1, M_1, \ldots, T_m, M_m)\, P(\sigma^2) \qquad \text{assume } \sigma^2 \perp\!\!\!\perp \text{ tree structure}$$

$$= P(\sigma^2) \prod_{t=1}^{m} P(T_t, M_t) \qquad \text{assume trees ind of each other}$$

$$= P(\sigma^2) \prod_{t=1}^{m} P(M_t \mid T_t)\, P(T_t) \qquad \text{Bayes Rule}$$

$$= \underbrace{P(\sigma^2)}_{\text{III}} \prod_{t=1}^{m} \underbrace{P(T_t)}_{\text{I}} \prod_{\ell=1}^{b_t} \underbrace{P(\mu_{t\ell} \mid T_t)}_{\text{II}} \qquad \text{assume each node ind. of each other}$$

three things to specify

$P(T_t) \to$ physical structure

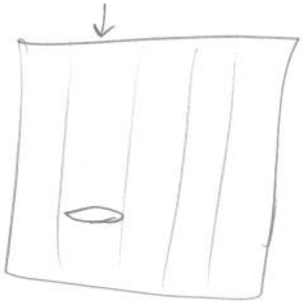
$P(T_t) \to$ split rules

Structure: begin with $\cdot$

$P(\text{split}) = \alpha(1 + \text{depth})^{-\beta}$, $\alpha \in (0,1)$, $\beta \in [0,\infty)$

For each split rule:

Pick $j \in \{1,\dots,p\}$ uniformly



If $x_{\cdot j} \in \mathbb{R}^n$, pick $i \in \{1,\dots,n\}$ rule is

$$x_j < X_{ij}$$

If $x_j \in \{1,\dots,k\}^n$ pick a random $\overset{\text{subset}}{\text{element}} \in 2^{\{1,\dots,k\}}$

$\Rightarrow x_j \in$ subset

$$P(\mu_{t\ell} \mid T_t) = N(\mu_\mu, \sigma^2_\mu)$$

How do we pick these two hyperparams?

We use a half empirical Bayes idea.

Pick μ_μ, σ^2_μ s.t. $[y_{(1)}, y_{(n)}]$ represents 95% of the possible spread. Assume mean is in center.

$m\mu_\mu = \frac{y_{(n)} - y_{(1)}}{2}$

$$E[Y \mid x] = E\left[\sum g_t\right] = m\,E[g_1] = m\,E[\mu_{11}] = m\mu_\mu$$

$$\text{Var}[Y \mid x] = \qquad = m\sigma^2_\mu$$

4

Back to Sam 101:

$$\left[m M_n \pm 1.96\, \sigma_m \sqrt{m} \right] = \left[y_{(1)}, y_{(n)} \right]$$

let $M_m = y_{(c)} \triangleq \frac{y_{(n)} + y_{(1)}}{2}$

$y_{(n)} - y_{(c)} = 1.96\, \sigma_m \sqrt{m}$

$\Rightarrow \sigma_m = \frac{y_{(n)} - y_{(c)}}{1.96 \sqrt{m}}$

$m M_m - 1.96\, \sigma_m \sqrt{m} = y_{(n)} - y_{(c)}$

Before we solve, why not:

$$y := \underbrace{\frac{y - y_{(1)}}{y_{(n)} - y_{(1)}}}_{\text{becomes } [0,1]} - \frac{1}{2} \in \left[-\frac{1}{2}, \frac{1}{2} \right]$$

$$\Rightarrow M_n = 0 \Rightarrow \sigma_m = \frac{\frac{1}{2}}{1.96 \sqrt{m}} \Rightarrow \sigma_m^2 \approx \frac{1}{15.4\, m}$$

K

$$\Rightarrow P(M_{te} \mid T_t) = N(0, \sigma_m^2)$$

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

To pick ν, λ we use a half Emp. Bayes idea.

Let's say $P(\sigma < s_y) = 0.9 = q$ hyperparam

→ or RMSE from 1st regression

(extremely conservative)

We fix $\nu = 3$ for large spread and compute λ from PDF.

How big should m be?

- One of my projects: make a fully Bayesian story
- Use X-validation
- Just use 200 since it doesn't make such a difference...

Now, return to master formula...

If you can't compute

$$P(T_1, M_1, \ldots, T_m, M_m, \sigma^2 \mid y)$$

explicitly, turn to Gibbs sampling...

$T_1, M_1 \mid T_2, M_2, \ldots, T_m, M_m, \sigma^2, y$

$\vdots$

$T_m, M_m \mid T_1, M_1, \ldots, T_{m-1}, M_{m-1}, \sigma^2, y$

$\sigma^2 \mid T_1, M_1, \ldots, T_m, M_m, y$

How does T_1, M_1, depend on other trees?

Remember:

$\vec{y} = g_1 + g_2 + \ldots + g_m + \vec{\varepsilon}$

$$\Rightarrow g_1 = \underbrace{\vec{y} - \underbrace{(g_2 + \ldots + g_m)}_{\text{contribution of other trees}} - \varepsilon}_{R_1}$$

what's left over i.e. what's "remaining", "residual response"

plays the role of the new "y"

How does σ^2 depend on trees?

$$\vec{\varepsilon} = \underbrace{\vec{y} - \sum g_t}_{\text{r.v. } E \text{ for residuals}}$$

Now we have a backfitting Gibbs sampler:

$T_1, M_1 \mid R_1, \sigma^2$

$\vdots$

$T_m, M_m \mid R_m, \sigma^2$

$\sigma^2 \mid E$

We wish we could...

$T_1 \mid R_1, \sigma^2, M_1$

$M_1 \mid R_1, \sigma^2, T_1$

circular logic

We can margin out M_1 to find

$$P(T_1 \mid R_1, \sigma^2) = \int P(T_1, M_1 \mid R_1, \sigma^2)\, dM_1$$

$$\propto \int P(R_1 \mid T_1, M_1, \sigma^2)\, P(T_1, M_1 \mid \sigma^2)\, dM_1$$

$$= P(T_1 \mid \sigma^2) \int \underbrace{P(R_1 \mid T_1, M_1, \sigma^2)}_{\text{Assume Normal}}\, \underbrace{P(M_1 \mid T_1, \sigma^2)}_{\text{Normal}}\, dM_1$$

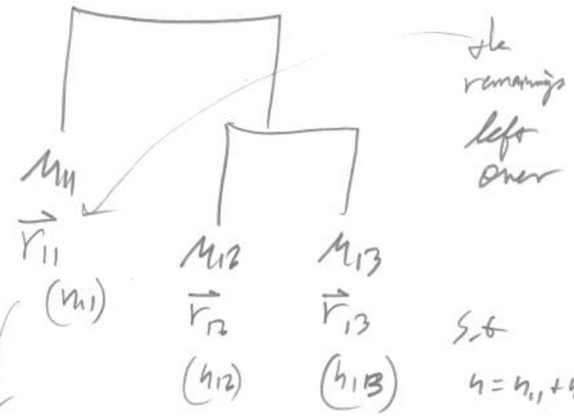
So it can be done...

So let's try to do it!

$$= P(T_1 \mid \sigma^2)\, P(R_1 \mid T_1, \sigma^2)$$

Lik. Assumption

T_1 $\quad (b_1 = 3)$



comes from error

$$r_{11k_1}, \ldots, r_{11k_{n_{11}}} \mid \mu_{11} \overset{iid}{\sim} N(\mu_{11}, \sigma^2)$$

Remember $\quad r_{11k_1} = \mu_{11} + \varepsilon_{11k_1}$

Now, the prior on $\mu_{11} \sim N(0, \sigma^2_\mu)$

So... easy marginny!

$$r_{11k_1} \sim N(0, \sigma^2_\mu) * N(0, \sigma^2) = N(0, \sigma^2 + \sigma^2_\mu)$$

$$Cov[r_{11k_1}, r_{11k_2}] = Cov[\mu_{11} + \varepsilon_{11k_1}, \mu_{11} + \varepsilon_{11k_2}] = \sigma^2_\mu$$

$$\Rightarrow \vec{r}_{11} \sim N_{n_{11}}\Big(0, \underbrace{\sigma^2_\mu J_{n_{11}} + \sigma^2 I_{n_{11}}}_{\Sigma_{11}}\Big)$$

not independent $\qquad$ (equicorrelation)

$$\frac{1}{(2\pi)^n} \prod_{l=1}^{b_1} |\Sigma_{1l}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \vec{r}_{1l}\, \Sigma_{1l}\, \vec{r}_{1l}\right)$$

$$\Rightarrow P(T_1 \mid R_1, \sigma^2) \propto P(T_1 \mid \sigma^2) \cdot$$

$$\prod_{l=1}^{b_1} \frac{1}{(2\pi)^{n_{1l}} \underbrace{|\Sigma_{1l}|^{1/2}}_{(\sigma^2)^{n-1}(\sigma^2_\mu n_{1l} + \sigma^2)}} \exp\left(-\frac{1}{2} \vec{r}_{1l}^{\,T} \underbrace{\Sigma_{1l}^{-1}}_{\frac{1}{\sigma^2_\mu n_{1l} + \sigma^2} \frac{1}{n_{1l}} J_{n_{1l}} + \frac{1}{\sigma^2}\left(I_{n_{1l}} - \frac{1}{n_{1l}} J_{n_{1l}}\right)} \vec{r}_{1l}\right)$$

Since we have a proportional density, we use M-H:

$$\text{if} \quad \ln(U(0,1)) < \ln \frac{q(T_1^* \mid T_0)}{q(T_0 \mid T^*)} \frac{\cancel{P(T^* \mid R_1, \sigma^2)}}{\cancel{P(T_0 \mid R_1, \sigma^2)}} \frac{P(T^*)}{P(T_0)}$$

$P(R_1 \mid T^*, \sigma^2)$

$P(R_1 \mid T_0, \sigma^2)$

$\Rightarrow$ accept

So how to get T^*?

Take T_0, and do one of the 4 steps:

Steps	w.p.
Grow	25%
Prune	25%
change	40%
Swap	10%

Next in our Metropolis-within-Gibbs implementation:

$$\mu_1 \mid R_1, T_1, \sigma^2 \quad \begin{cases} \mu_{11} \mid \vec{r}_{11}, \sigma^2 \\ \mu_{12} \mid \vec{r}_{11}, \sigma^2 \\ \vdots \\ \mu_{1b_1} \mid \vec{r}_{11}, \sigma^2 \end{cases}$$

Simple due to conjugacy!

$$P(\mu_{11} \mid \vec{r}_{11}, \sigma^2) \propto P(\vec{r}_{11} \mid \mu_{11}, \sigma^2)\, P(\mu_{11} \mid \sigma^2)$$

$$= \left(\prod_{i=1}^{n_{11}} N(\mu_{11}, \sigma^2) \right) N(0, \sigma_\mu^2)$$

$$\cdots$$

$$\Rightarrow \rho B1 \qquad = N\left(\frac{\frac{n}{\sigma^2}\bar{r}_{11}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}} \right)$$

Gelman et al

Now we sample $T_2 \mid R_2, \sigma^2$, $\mu_2 \mid R_2, \sigma^2, T_2, \ldots$

until

$$\sigma^2 \mid E$$

Also simple due to conjugacy

$$P(\sigma^2 \mid E) \propto P(E \mid \sigma^2)\, P(\sigma^2)$$

$$= \left(\prod_{i=1}^{n} N(0, \sigma^2) \right) \mathrm{InvG}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2} \right)$$

$$\cdots$$

$$= \mathrm{InvG}\left(\frac{\nu + n}{2}, \frac{\lambda\nu + \sum e_i^2}{2} \right)$$

Now we're good

$$P(T_1, \mu_1, \ldots, T_m, \mu_m, \sigma^2 \mid y)$$

We draw N_G samples and burn the first B

$$\hat{f}_{B+1}, \ldots, \hat{f}_{N_G - B} \quad \text{empirical PDF}$$

To get the best guess: avg of PDF:

$$\hat{y} = \frac{1}{N_G - B} \sum_{g=B+1}^{N_G} \hat{f}_g(x^*) \quad \underline{\underline{OR}} \quad \hat{y} = \hat{f}_{(50\%)}(x^*)$$

median

$$PPI_{y,95\%} = \left[\hat{f}_{25\%}(x^*),\ \hat{f}_{75\%}(x^*) \right]$$

— partial dep function

p277-278 BART probit

no rescaling y,

$$P(\hat{Y}=1 \mid x^*) = \Phi\left(\sum_{t=1}^{m} g_t(x^*)\right)$$

same

$\mu_{ij} \overset{iid}{\sim} N(0,\sigma_\mu^2)$ let $\sigma = 1$ explicitly

augmentation idea

$Z_1, \dots, Z_n \overset{iid}{\sim} N\left(\Sigma g_t, 1\right)$

s.t.

$Z_i \mid y_i=1 \sim \max\{N(\Sigma g_t, 1), 0\}$

$Z_i \mid y_i=0 \sim \min\{N(\Sigma g_t, 1), 0\}$

Use Z_i's as y_i's and do

metropolis-within-gibbs backfitting for

$T_1, M_1, \dots, T_m, M_m$

See 5.1 ~~Bakeoff~~ 42 datasets

normal trans(y). 5-fold

Cross validation

two versions of BART

default / CV

$\nu=3,\ q=90\%,\ k=2,\ m=200$

$N_b=1000,\ B=200$

Competitors

Lasso,

Grad. Boosting,

Random Forest,

Neural Nets

$(\nu, q) \in \{(3, 90\%), (3, 99\%), (10, 75\%)\}$

$k \in \{2, 3, 5\}$

$m \in \{50, 200\}$

18 choices