Bayesian Additive Regression Trees

Lecture 23 - CS 689, Spring 2023

- Last time: decision trees and their ensembles
  - 1. Individual decision trees
  - 2. DTree ensemble with random forests
  - 3. DTree ensembles with greedy optimization - forward stagewise learning and **gradient boosted trees** (next slide)

- Rest of today: **Bayesian learning** for decision tree ensemble

$L = (y - f(x))^2$

$\frac{\partial L}{\partial f} = -2(y - f(x))$

Algorithm 16.4: Gradient boosting

Initialize $f_0(\mathbf{x}) = \operatorname{argmin}_{\gamma} \sum_{i=1}^{N} L(y_i, \phi(\mathbf{x}_i; \boldsymbol{\gamma}))$;
for $m = 1 : M$ **do**
    Compute the gradient residual using $r_{im} = -\left[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right]_{f(\mathbf{x}_i) = f_{m-1}(\mathbf{x}_i)}$,
    Use the weak learner to compute $\boldsymbol{\gamma}_m$ which minimizes $\sum_{i=1}^{N} (r_{im} - \phi(\mathbf{x}_i; \boldsymbol{\gamma}_m))^2$;
    Update $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu \phi(\mathbf{x}; \boldsymbol{\gamma}_m)$;
Return $f(\mathbf{x}) = f_M(\mathbf{x})$

Learning rate / step size

Review: MCMC for Bayesian pred.

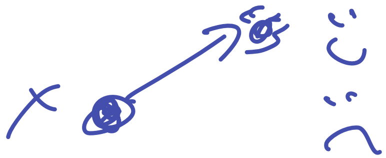
- Goal: want posterior predictive

$$p(y \mid x, X^{Tr}, Y^{Tr}) = \int_\theta \underline{p(\theta \mid X^{Tr}, Y^{Tr})}\, \underbrace{p(y \mid x, \theta)}\, d\theta$$

$$\approx \sum_{s=1}^{S} p(y \mid x, \theta^{(s)}) \quad \text{for } \theta^{(s)} \sim p(\theta \mid X^{Tr}, Y^{Tr})$$

- Method: <u>Approximate with samples</u> from the parameter posterior
    - No closed form, but can use <u>Markov Chain Monte Carlo</u>

MCMC: key algorithms



Algorithm 24.2: Metropolis Hastings algorithm

Initialize x^0 ;
for $s = 0, 1, 2, \ldots$ **do**
  Define $x = x^s$;
  Sample $x' \sim q(x'|x)$;
  Compute acceptance probability

$$\alpha = \frac{\tilde{p}(x')q(x|x')}{\tilde{p}(x)q(x'|x)}$$

  Compute $r = \min(1, \alpha)$;
  Sample $u \sim U(0, 1)$;
  Set new sample to

$$x^{s+1} = \begin{cases} x' & \text{if } u < r \\ x^s & \text{if } u \geq r \end{cases}$$

Algorithm 1 Gibbs sampler

Initialize $x^{(0)} \sim q(x)$
for iteration $i = 1, 2, \ldots$ **do**
  $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)})$
  $x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)})$
  $\vdots$
  $x_D^{(i)} \sim p(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \ldots, X_D = x_{D-1}^{(i)})$
end for

for us

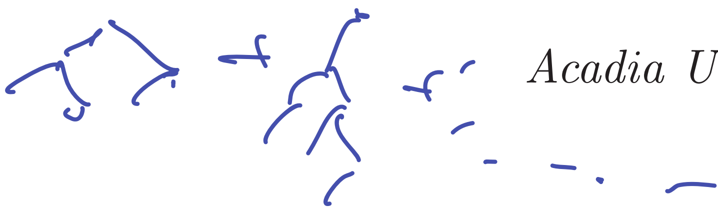
$$\tilde{p}(\theta) = \underbrace{p(\theta)}_{\text{prior}} \underbrace{p(Y^{tr}|X^{tr},\theta)}_{\text{lik}}$$

The Annals of Applied Statistics
2010, Vol. 4, No. 1, 266–298
DOI: 10.1214/09-AOAS285

BART: BAYESIAN ADDITIVE REGRESSION TREES[1,2]

By Hugh A. Chipman, Edward I. George and Robert E. McCulloch

Acadia University, University of Pennsylvania and University of Texas at Austin



- Model:

ensemble $f(x) = \sum_{j=1}^{m} \underbrace{g_j(x)}_{\text{one tree}}$

$P(g)$: prior on trees

$P(y|x,f)$: likelihood, $N(f(x), \sigma^2)$

Posterior over ensembles: $P(f|Y^{tr}, X^{tr}) \propto P(f)\, P(Y^{tr}|X^{tr}, f)$

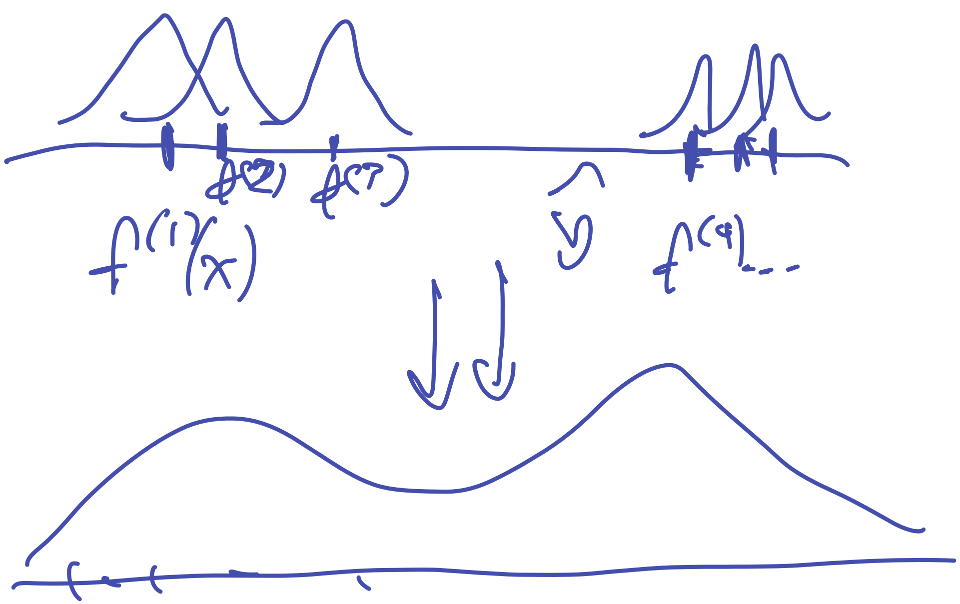
Uncertainty-aware predictions

BART for causal inference

- If we had samples of posterior ensembles, we could do lots of nice probabilistic inferences / predictions!

$$f^{(s)} \sim p(f \mid X, Y)$$

$$p(Y \leq t \mid x) \approx \frac{1}{S}\sum_s p(Y \leq t \mid f^{(s)}(x))$$



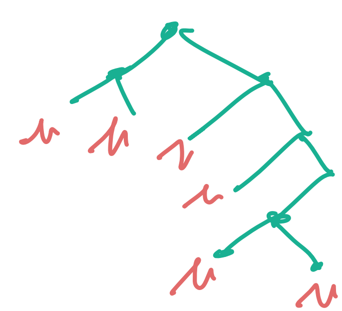
Model and Prior Structure

T: binary tree, $M = \{\mu_1 \dots \mu_b\}$ for each leaf

$$\boxed{\text{fix}} \quad Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$p\big((T_1, M_1) \dots (T_m, M_m), \sigma\big) = p(\sigma) \underbrace{\left[\prod_j \overbrace{p(M_j \mid T_j)}^{\text{leaf preds.}} \overbrace{p(T_j)}^{\text{struct}}\right]}_{\text{Each tree is indep}}$$

$$p(M_j \mid T_j) = \underbrace{\prod_{i=1}^{b} p(\mu_{ij} \mid T_j)}_{\text{Each leaf indep}}$$



Tree structure prior

Need to regularize size/shape of trees

- Node at depth d, is nonterminal with prob

$$\alpha (1+d)^{-\beta}$$

hyperparams

- Uniform distribs: ① what variable to split on
② what threshold to use: any discrete value in training set

Data-driven priors

$$(\mu_{ij} \mid T_j) \sim N(\mu_\mu, \sigma_\mu^2)$$

μ

- Leaf predictions: conjugate normal prior

(1-D Gaussian models)

$$p(y \mid x \in \text{leaf } b) = \int_{\mu_b} p(\mu_b \mid T)\, p(y^{Tr} \mid x^{Tr}, \mu_b)$$

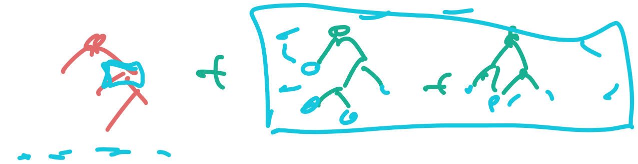
$$= N(\text{- - - -},\ \text{- - - -})$$

Avg of μ_b, and avg y within leaf b

σ

- Variance term: conjugate inverse chi square

μ_μ, σ_μ^2 : based on y^{Tr}



Backfitting MCMC

T_j, M_j $\quad\quad\quad \sigma$

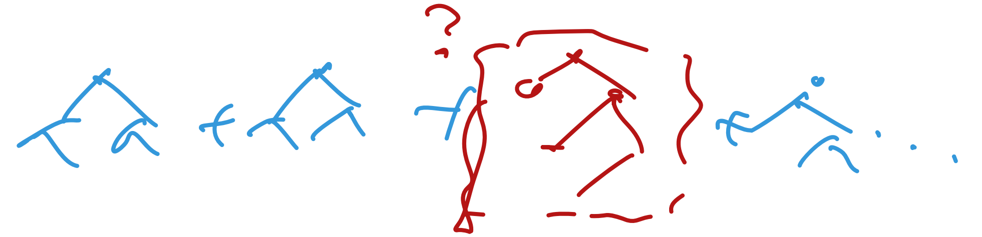
- Gibbs sampler to resample one tree at a time, and variance

$$p\big((T_1, M_1), \ldots (T_m, M_m), \sigma \mid X, Y\big)$$

GS for $p(\sigma \mid T_1 M_1 \ldots T_m M_m, X, Y)$

then each $p(T_j, M_j \mid T_{(j)}, M_{(-j)}, X, Y, \sigma)$

Partial residuals



$$\tilde{R}_j = \left\{ y_i - \sum_{k \neq j} g(x_i; T_k, M_k) \right\}_{i=1..N}$$

$$p(T_j, M_j \mid T_{(j)}, M_{(j)}, X, y, \sigma) = p(T_j, M_j \mid R_j, \sigma)$$

$$\propto \underbrace{p(R_j \mid T_j, M_j, \sigma)}\, p(T_j, M_j)$$

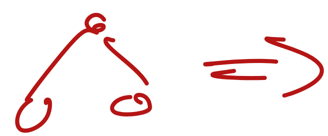
$$R_j = g(x; T_j, M_j) + \varepsilon$$

Tree proposal & resampling

$$p(T_j \mid R_j, \sigma) \propto p(T_j) \underbrace{\int_{M_j} p(R_j \mid M_j, T_j, \sigma)\, p(M_j \mid T_j, \sigma)\, dM_j}_{\text{Gaussian } R_j \sim N(\ldots, \ldots)}$$

- Proposal distribution from old to new tree:

$\rightarrow q(T^{new} \mid T^{old})$

- Grow terminal node 
- Prune pair of nodes 
- Change nonterminal rule 

$$\alpha = \frac{p(T^{new} \mid R_j, \sigma)\; q(T^{old} \mid T^{new})}{p(T^{old} \mid R_j, \sigma)\; q(T^{new} \mid T^{old})}$$

~~Leaf resampling~~

For each outer iteration, s:

For each tree j,

$$(T_j^{(s)}, M_j^{(s)}) \sim \text{MH acceptance for } P(T_j, M_j \mid R_j, \sigma)$$

$$\sigma^{(s)} \sim P(\sigma \mid T, M, \sigma, X, Y) \quad [\text{conjugate}]$$

$\Longrightarrow$

$$\{T_1^{(s)}, M_1^{(s)}, \ldots, T_m^{(s)}, M_m^{(s)}, \sigma^{(s)}\}_{s=1\ldots S}$$