Bayesian flexible modeling of trip durations

Hugh Chipman, Edward George, Jason Lemp, and Robert McCulloch

Transportation Research Part B

Volume 44, Issue 5，Pages 686-698，June 2010

2010/05/31（月）

論文ゼミ#2

M1　戸叶洋道

BARTモデル

- BART(Bayesian Additive Regression Trees)は，従属変数を，回帰木の和として表現したモデル

- 回帰木は，ツリーの構造

$$T$$

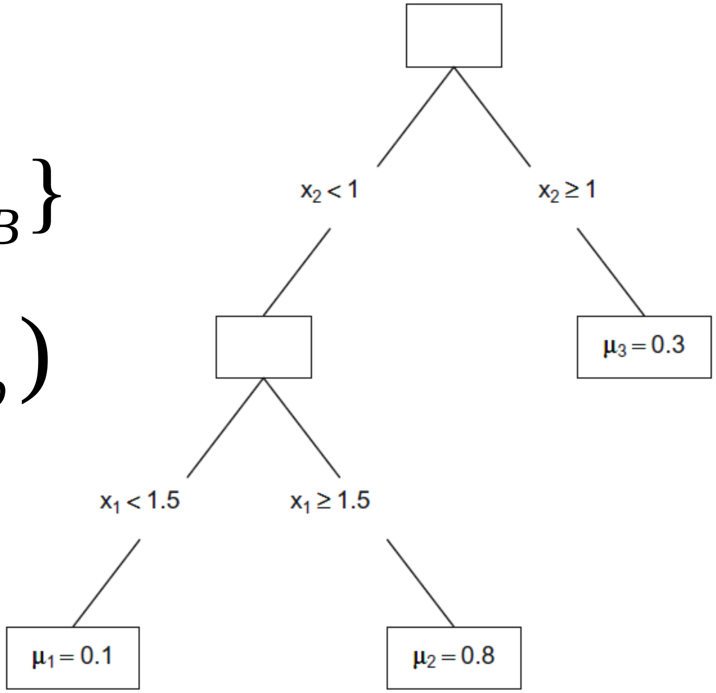
ターミナルノードの値

$$M=\{\mu_1,\ldots,\mu_B\}$$

分岐の条件

$$x=(x_1,\ldots,x_p)$$

で構成　$g(x;T,M)$

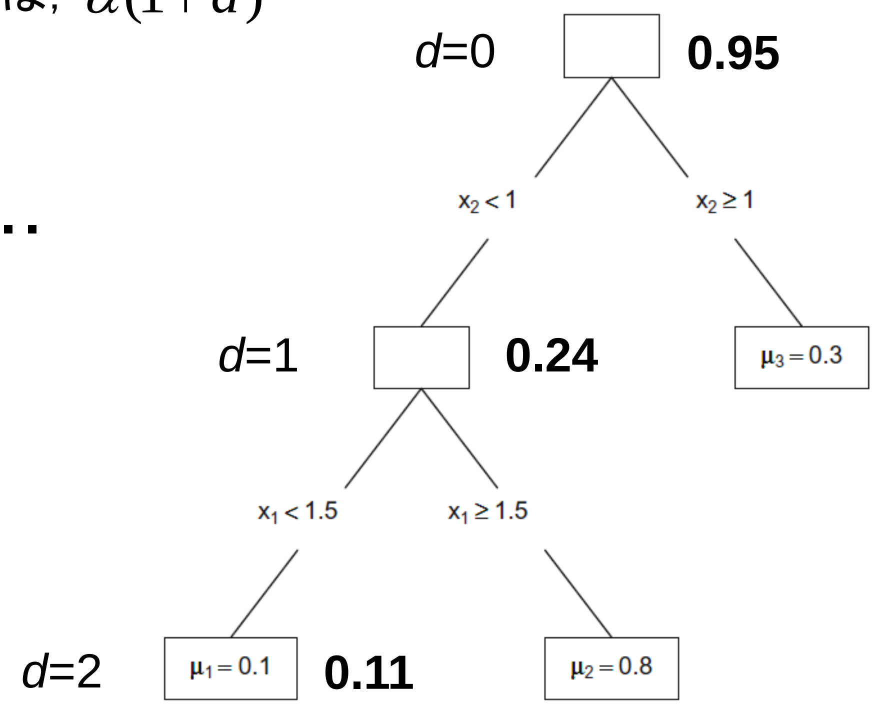


regression tree (回帰木)

- 回帰木は，一番最初にルートノードがあり，そこから子供ノードが枝分かれし，一番下のターミナルノードへと細分化される構造をとる．各ターミナルノードにμの値が格納されている．
- 図は，ターミナルノードが3つの例．　$M = \{0.1, 0.8, 0.3\}$
- あるノードがターミナルノードでない確率は，$\alpha(1+d)^{-\beta}$

  α=0.95,β=2とすると（*d*はノードの深さ）

  それぞれのノードに子供がいる確率は・・・



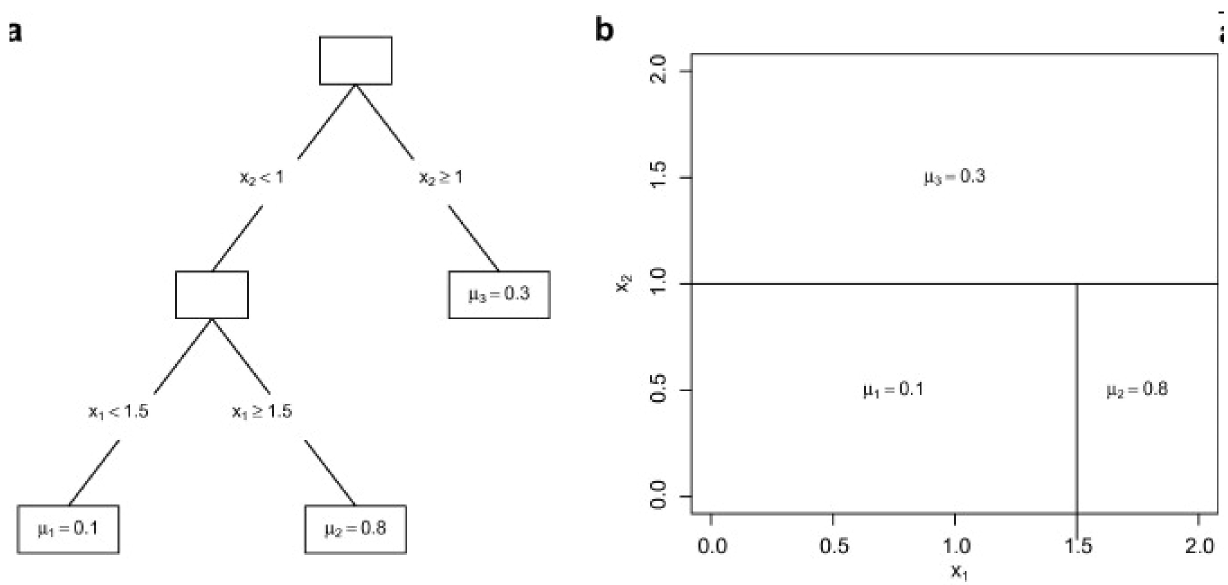
- 枝分かれの変数*x*とその値もランダムに決まる

回帰木の事前確率

$\alpha = 0.95, \beta = 2$

x_1, x_2 は，0.0~2.0まで0.1刻みの値をとる

$0.95 \times$ (root node is nonterminal)
$0.5 \times$ (split is on X_2, one of two variables)
$0.05 \times$ (split is on 1 of 20 possible locations)
$0.2375 \times$ (left child is nonterminal)
$0.5 \times$ (left child splits on X_1, one of two variables)
$0.05 \times$ (split is on 1 of 20 possible locations)
$(1 - 0.1056) \times (1 - 0.1056) \times$ (two children are terminal)
$(1 - .2375)$(right child of root node is terminal) $= 8.6 \times 10^{-5}$



- この木は，$(x_1, x_2) = (1.0, 0.5)$ のとき，$\mu_1 = 0.1$ を返す．

A sum of trees model

$$Y = g(x;T_1,M_1) + g(x;T_2,M_2) + \cdots + g(x;T_m,M_m) + \varepsilon, \qquad \varepsilon \sim N(0,\sigma^2)$$

$$f(x) = g(x;T_1,M_1) + g(x;T_2,M_2) + \cdots + g(x;T_m,M_m)$$

$\mu_{i,b}$ は，$\mu \sim N(0,\sigma_\mu^2)$ に従うものとする.

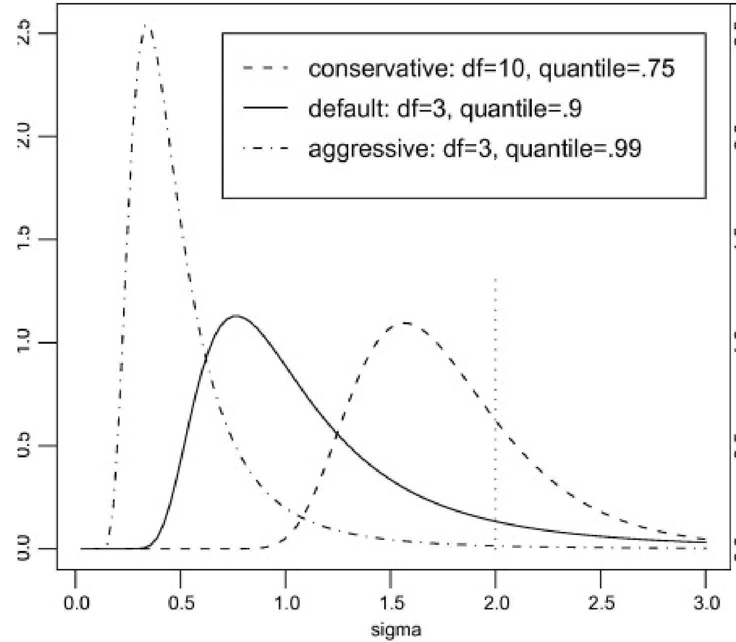
σ は. $\sigma^2 \sim \nu\lambda / \chi_\nu^2$ に従うものとする.

$\hat{\sigma}$：（σのオーバーエスティメート）を定める. 実用的には，標本標準偏差を用いる. 右の図では2.0

ν： 自由度（degrees of freedom）を3~10の範囲で定める.

q： 分布がシグマハット以下である確率(puantile)

デフォルト $(\nu, q) = (3, 0.9)$



A sum of trees model

decision rule

$$Y = \sum_i g(x; \boxed{T_i}, \boxed{M_i}) + \boxed{\varepsilon}$$

$\mu \sim N(0, \sigma_\mu^2)$

$\varepsilon \sim N(0, \sigma^2)$

$\sigma^2 \sim \nu\lambda / \chi_\nu^2$

MCMC algorithm

- yという観測値が与えられた時の，事後分布を導く

$$p((T_1, M_1), \ldots, (T_m, M_m), \sigma \mid y)$$

- 基本的には，ギブスサンプラー
  - まずTとMをひとつずつサンプリングし，

$$(T_1, M_1) \mid T_{(1)}, M_{(1)}, \sigma, y$$
$$(T_2, M_2) \mid T_{(2)}, M_{(2)}, \sigma, y$$
$$\vdots$$
$$(T_m, M_m) \mid T_{(m)}, M_{(m)}, \sigma, y$$

  - 次にσをサンプリングする

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

  - これを繰り返す

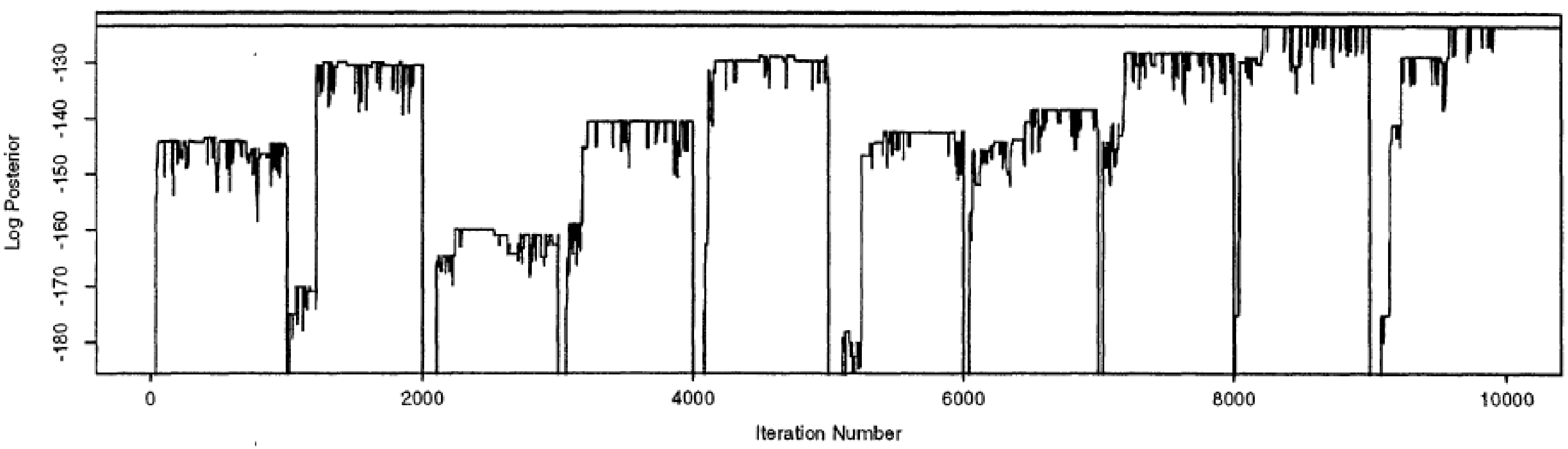
MCMC algorithm

- M, σは事前分布に標準正規分布を持つ
  ⇒通常のベイズ更新を繰り返す

- Tは木の構造⇒どうやって更新？
  - GROW
    - ターミナルノードを増やす(0.25)
  - PRUNE
    - ターミナルノードを減らす(0.25)
  - CHANGE
    - 分岐条件を変更(0.4)
  - SWAP
    - 親子の分岐条件を交換(0.1)
- Tは事前分布を持たないので，ギブスサンプラーでは出来ない
  - ⇒メトロポリスヘイスティングス
  - Chipman(1998)によれば，更新の選択確率は常に1

なぜ sum of？

- Chipmanは，1998年に単回帰木のモデルを提案している
- 木が一つだと柔軟性に乏しく，すぐにある値に収束してしまうが，それが真の値かどうかはわからない．
- ある「good tree」を見つけて何度も再スタートを繰り返さなければならない

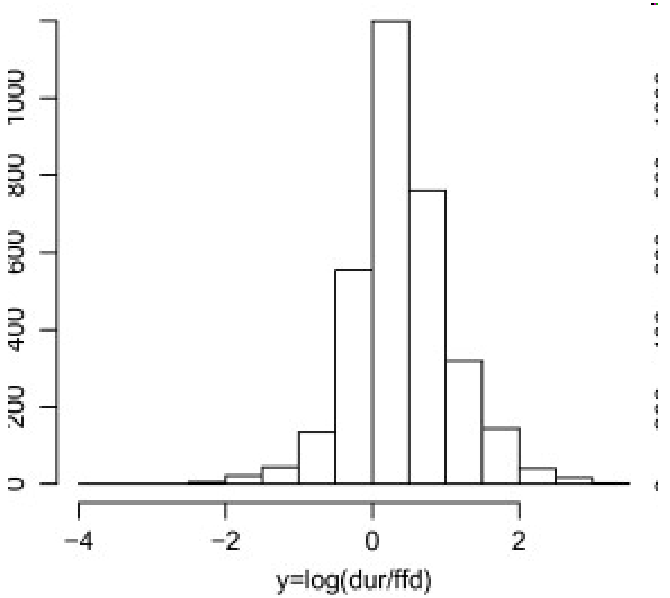


Fitting trip duration data with BART

- データは，テキサス州の車によるトリップ
- 各トリップの変数(17個)
  - 世帯属性
    - 世帯人数
    - 収入
    - 子供の数　etc…
  - トリップメーカーの変数
    - 年齢
    - 職業　etc…
  - トリップの変数
    - 日時
    - トリップタイプ
    - 出発時間
    - 自由流れトリップ距離
    - 自由流れトリップ時間
    - トリップ時間　etc…

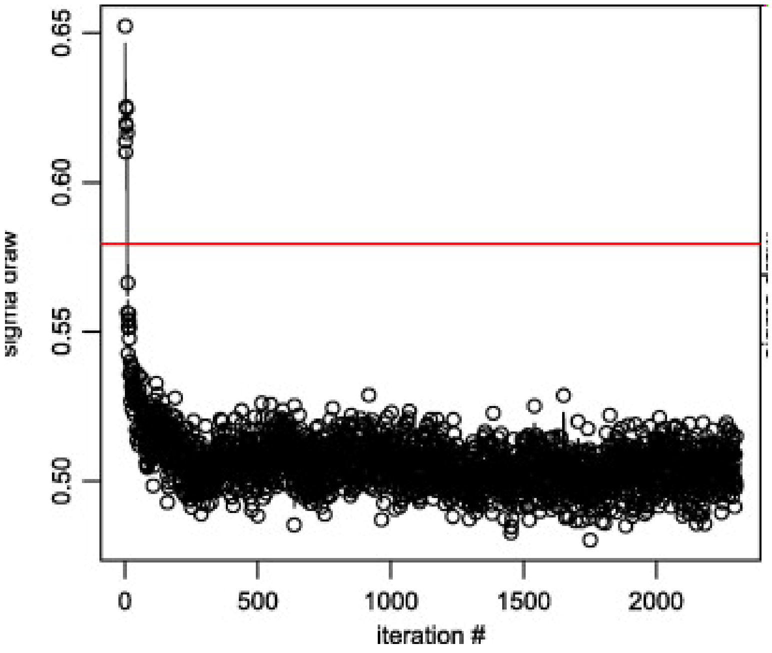
従属変数 $y = \dfrac{dur}{ffd}$

トリップ時間は5分単位で丸められて報告されていることが多いので注意！



BART results, all variables

- BARTモデル結果
  - 図は, σのMCMCの結果のプロット
  - burn-inは300で，総繰り返し数2300
  - 計算時間226秒(2.93GHz,Core 2 Duo)
  - σの平均値は0.5
  - yとBARTのR二乗値は48%
  - xに関する変換は必要ない
    (sum of tree が自動的に柔軟に形を変える)
- 通常の線形回帰モデル結果
  - σの推定値は0.58(図の赤線)
  - R二乗値は28%
  - xに関して変換が必要



結果の解釈

- 各繰り返しで得られる$f^*(x)$は，真の$f(x)$の事後分布からのサンプルとして考えることが出来る.
- $f^*(x)$の平均値は，$f(x)$の平均値と推定することが出来る.
- BARTを予測デバイスと考えるならば，$f^*(x)$の平均値に各xを代入してあげればyを予測することが出来る.
- 実際にどのような関数になっているか（どの変数が被説明変数に対してどのように寄与しているか）はわからない.

結果の解釈

- Chipmanは，変数選択法を提案
- 分岐に利用されている回数が多い変数ほど，被説明変数に対する寄与が大きいと考える．
  - 今回の場合，
    - free flow destance(19%)
    - trip type(4%)
    - departure time(3.4%)
  - が分岐で多く利用された．

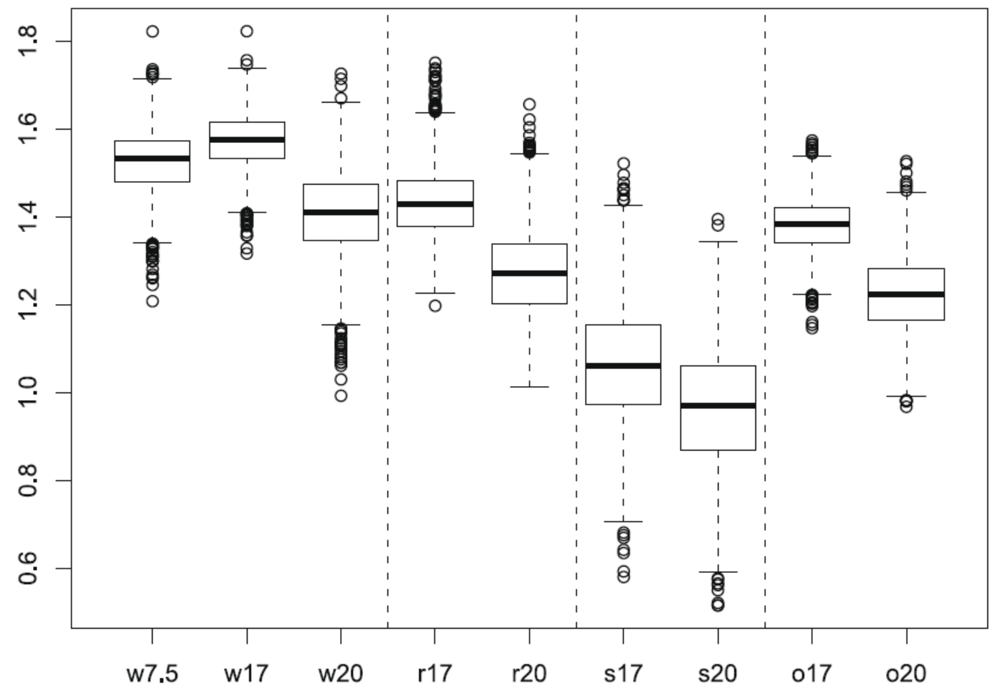
結果の解釈

- 重要だと思われる変数をの組み合わせをいくつか抽出して，比較する.

■ワークトリップはソーシャルトリップよりも長い

■20時よりも17時の方が基本的に長い時間となっている⇒出発時間がtrip durationに大きな影響を与えている

■ソーシャルトリップは他のトリップよりも不確実性が高い



w,r,s,o=work.retail,social,other

Transforming independent variables

- 線形回帰モデルを，変数変換で改善
  - free flow distanceを対数変換
    - R二乗値28⇒40%
    - σの推定値0.58⇒0.53
  - departure timeの変換(popuri et al(2008)による)

$$g_1(T) = \exp(\sin(\frac{2\pi T}{24})), g_2(T) = \exp(\cos(\frac{2\pi T}{24}))$$
$$g_3(T) = \exp(\sin(\frac{4\pi T}{24})), g_4(T) = \exp(\cos(\frac{4\pi T}{24}))$$

$$g(T) = \sum_{i,j} \beta_{i,j} g_i(T)^j$$

    - R二乗値40⇒41%
    - σの推定値0.53⇒0.52
  - ちなみに，今回のBARTモデルは，R二乗値48%，σの平均0.5

まとめ

- sum of trees ＋ MCMC によって，柔軟に，そして自動的に説明変数の関数が変化していくモデル.
- 線形な回帰モデルと違い，変数の変換を自動的にやってくれるため，予測のモデルとしては非常に有用.
- しかし，Chipmanもその解釈に奮闘中であるように，yとxの関係がブラックボックスの中にある感じは否めない.