

Learning Treed Generalized Linear Models

Hugh Chipman, University of Waterloo

Joint work with Ed George (U. Pennsylvania)

and Rob McCulloch (U. Chicago)

Papers and software available online

<http://www.stats.uwaterloo.ca/~hachipman/>

Relevant online papers:

- “Bayesian Treed Generalized Linear Models”, by Chipman, George, and McCulloch
- “A Bayesian Treed Model of Online Purchasing Behavior Using In-Store Navigational Clickstream.”, Moe, Chipman, George, and McCulloch (for the marketing example)
- Chipman, George, and McCulloch, (2002) “Bayesian Treed Models”, Machine Learning, 48, 299-320.

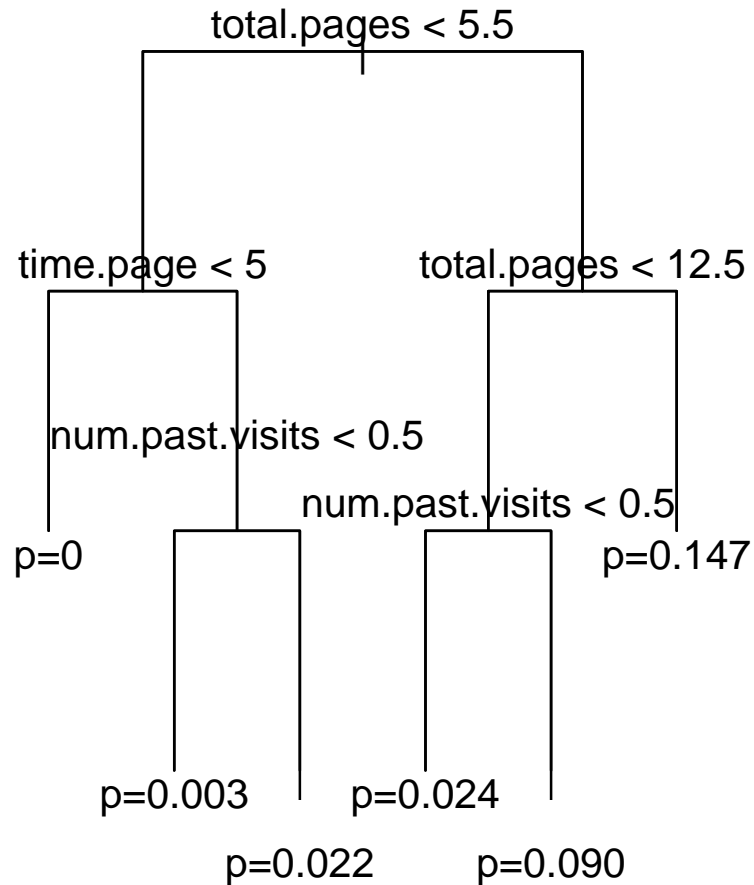
Marketing example: On-line retailing

(Joint work with Moe, U Texas @ Austin)

- Potential customers visit an online store (website)
- Each person's navigation path generates various customer characteristics (e.g. number of product pages, total number of pages, average time per page, etc.)
- (Binary) response variable: Does the customer buy?
⇒ Classification....
- ... But it's important to understand the factors that lead to buying.
- 23 variables, 34,585 observations (from a 6 month period).
- Only 604 out of 34,585 observations are "buyers" (under 2%).

Basic Tree:

(e.g. CART, Classification and Regression Trees, Breiman et. al. 1984)

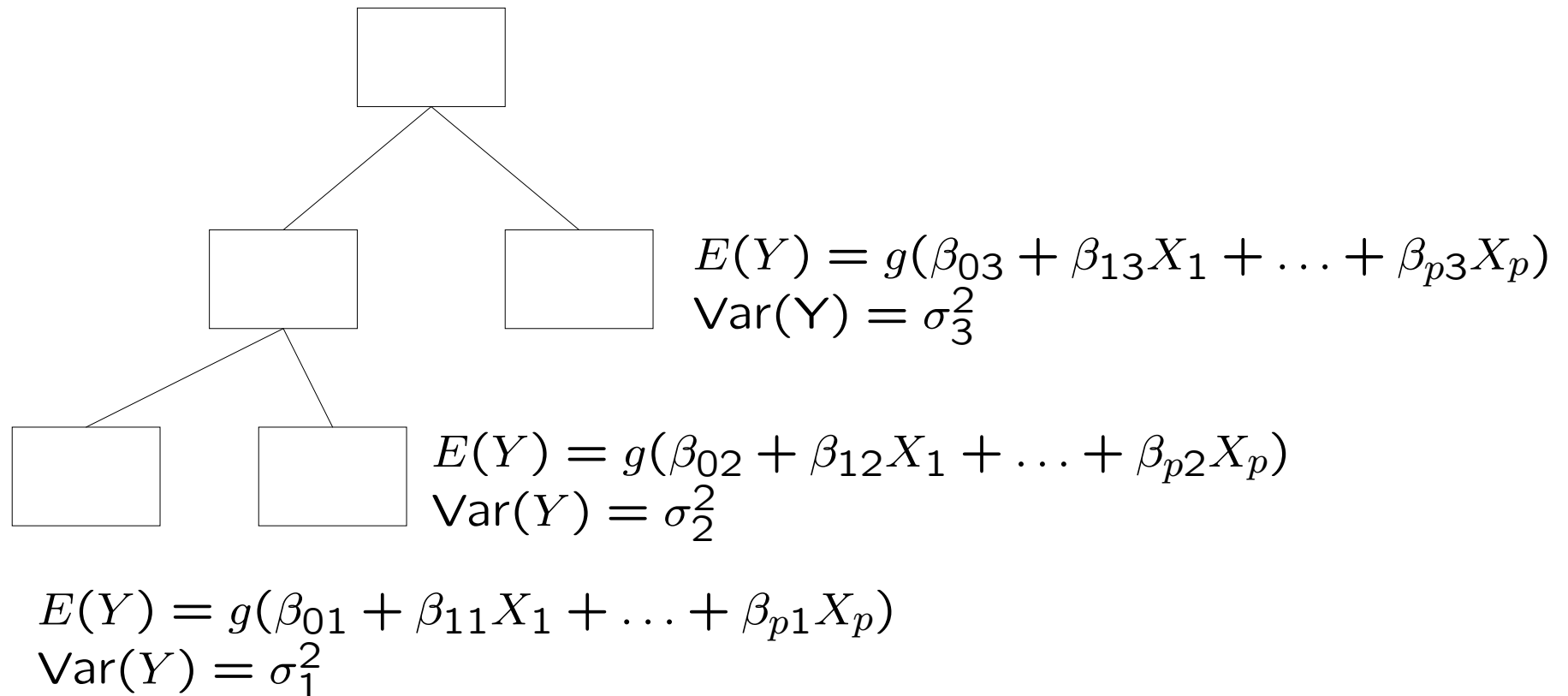


- Small tree for illustration.
- Greedy forward stepwise search, then backwards pruning
- Usually choose tree by cross-validation.
- Constant prediction in each terminal node.

Models in terminal nodes

(Chipman, George and McCulloch 2002, Chaudhuri et. al. 1994, Grimshaw & Alexander 1998, Jordan and Jacobs 1994)

- Linear/generalized linear models in terminal nodes:



Most of the talk is about a (Bayesian) method for fitting treed GLMs

But first, “Why?” :

- **Flexibility:** Adaptive nature of trees + piecewise linear model.
- **Simplicity:**
 - It gives smaller trees.
 - Conventional linear model is a special case (single node tree).
- Easier to rank individuals (no ties in predictions).
- **Other enhancements** of Bayesian approach:
 - Probability distribution on the space of trees
 - Stochastic search for good trees.

Another example: What gets your articles cited?

- McGinnis, Allison, Long (1982) examined careers of 557 biochemists.
- **Question:** what influences later research productivity?
- **Response:** number of citations in years 8, 9 & 10 after Ph.D.
- **Predictors (seven):**
 - number of articles in 3 years before Ph.D. awarded
 - Married @ Ph.D.?
 - Age @ Ph.D.
 - Postdoc?
 - Agricultural college?
 - Quality measures of grad/undergrad schools
- Poisson model seems natural since citations are counts.

The rest of the talk:

To do a Bayesian approach, we need

1. **Priors:** specification can be difficult.
2. **Posteriors:** Calculation involves:
 - Approximations for the posterior in the GLM case.
 - Algorithm to search for high posterior trees.

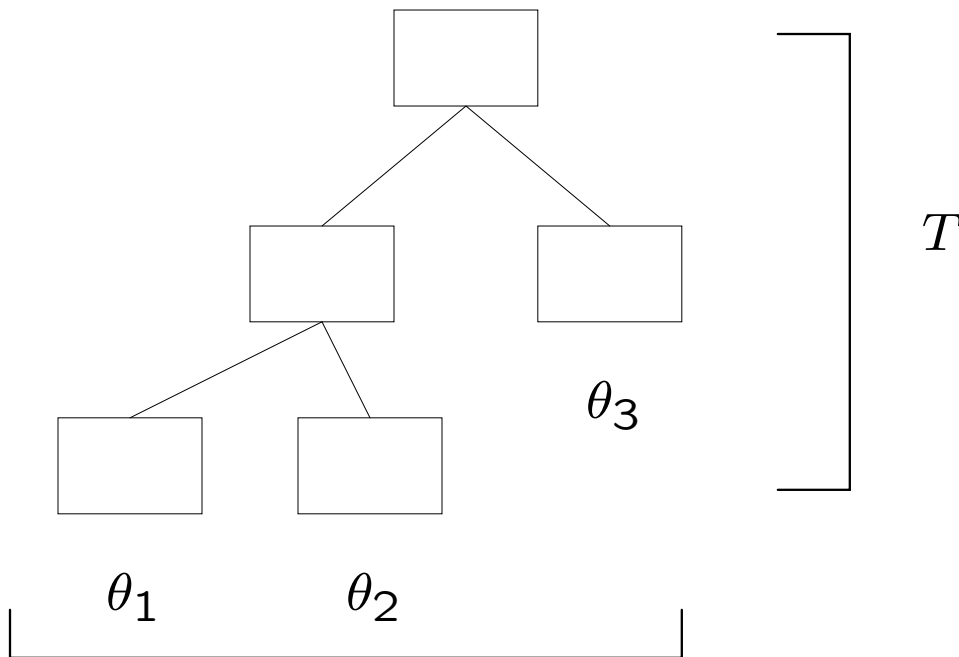
Bayesian approach to CART originally in Chipman, George, & McCulloch (1998) and Denison, Mallick, & Smith (1998).

After covering the Bayesian approach, we'll return to the examples.

After that I'll mention some recent work on Boosting.

Priors

Prior for Θ is $\pi(\Theta, T) = \pi(\Theta|T)\pi(T)$

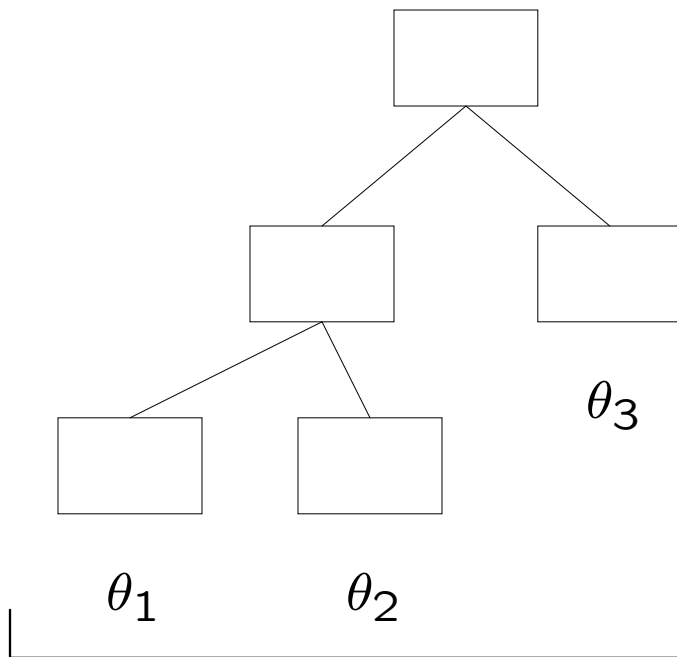


Θ specified
conditional on T

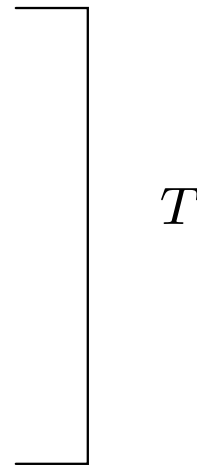
Prior on T specified
in terms of a process
for growing trees.

Priors

Prior for Θ is $\pi(\Theta, T) = \pi(\Theta|T)\pi(T)$



Θ specified
conditional on T



Prior on T specified
in terms of a process
for growing trees.

$\theta_i = (\mu_i, \sigma_i)$ for regression tree
(identifies shifts in μ and σ)

$\theta_i = P(Y = \text{class } j)$ for classification tree

$\theta_i = (\beta_i, \sigma_i)$ for Generalized Linear Model
(Regression coefficients and dispersion)

Priors for $\theta_i = (\beta_i, \sigma_i)$, conditional on the tree T

$$\sigma_i \sim \text{Inverse Gamma}(\nu, \lambda) \quad \beta_i | \sigma_i \sim N(0, \sigma^2 c^2 I)$$

Independent across terminal nodes $i = 1, \dots, b$

Different mean/variance for β possible, but this reasonable if X 's are scaled.

Choice of c is quite important:

- If $\pi(\beta_i | \sigma_i)$ is too informative (c small), you'll shrink β 's too much.
- If $\pi(\beta_i | \sigma_i)$ is too vague (c large), you'll favour the simple tree too much.
- Experiments in the regression case suggest $1 \leq c \leq 5$ is reasonable. Less clear in the GLM case.

Posterior distributions

$$P(T, \Theta | Y) = \frac{L(T, \Theta)\pi(\Theta|T)\pi(T)}{P(Y)} \propto L(T, \Theta)\pi(\Theta|T)\pi(T)$$

L = likelihood = $P(Y|\Theta, T)$

Would like to integrate out the terminal node parameters (Θ).

$$P(T|Y) \propto \pi(T) \int L(T, \Theta)\pi(\Theta|T)d\Theta$$

This tells us what trees are most probable given the data.

Note that since we assume independent observations Y_1, \dots, Y_n , these calculations can be done [separately](#) in each terminal node.

Analytic solution available for linear regression with Gaussian errors, an approximation (Laplace approx.) is necessary for GLMs.

Finding T with large posterior probability

- The space of trees is enormous.
⇒ We need to find T 's that have large posterior probability *without* evaluating $P(T|Y)$ for all possible T .
- Greedy search?
- Instead use the **Metropolis-Hastings algorithm** to sample from the posterior of trees.
⇒ Stochastic search guided by posterior.
- $P(T|Y)$ (up to a normalizing constant) can be used both in the Metropolis-Hastings algorithm, and to rank all trees sampled so far.

A posterior distribution on trees...

This sounds simple, but there are problems.

- **Many local maxima** - even MH tends to gravitate toward one mode.

Solution: Restart the MH algorithm repeatedly to find different modes.

- **Posterior** on individual trees is **diluted** by the prior.

Example: Can split at $X_1 = 1, 2$ or $X_2 = 1, 2, \dots, 100$. Prior mass for splits on X_2 is $1/50$ of mass for splits on X_1 .

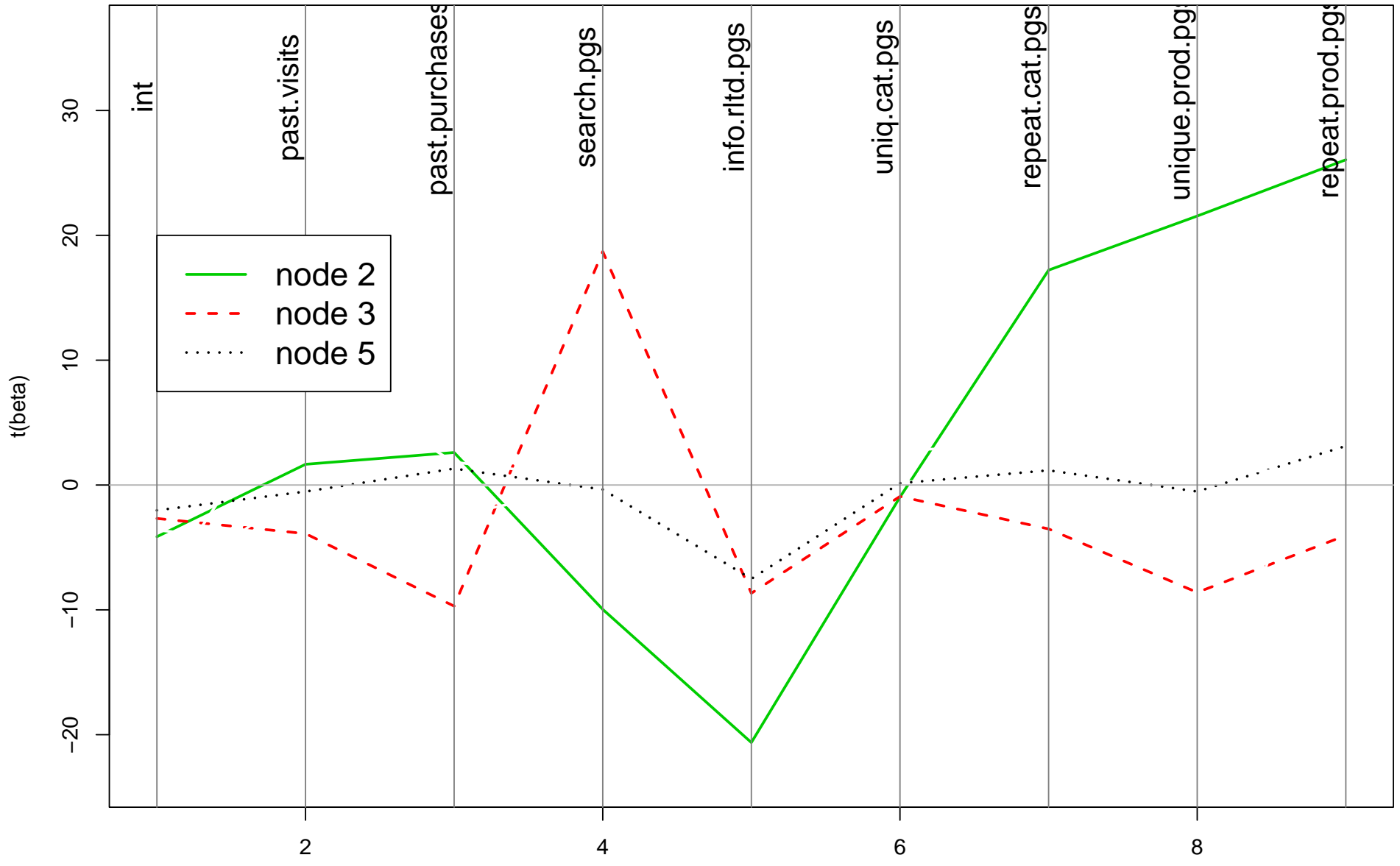
Solution: Don't use the posterior to rank individual trees. Either look at likelihood or sum the posterior over groups of trees.

- **A forest of trees:** Many different trees can fit the same dataset well.

Solution: techniques to sort through the forest and identify groups of similar and different trees.

Web marketing example: Treed logit

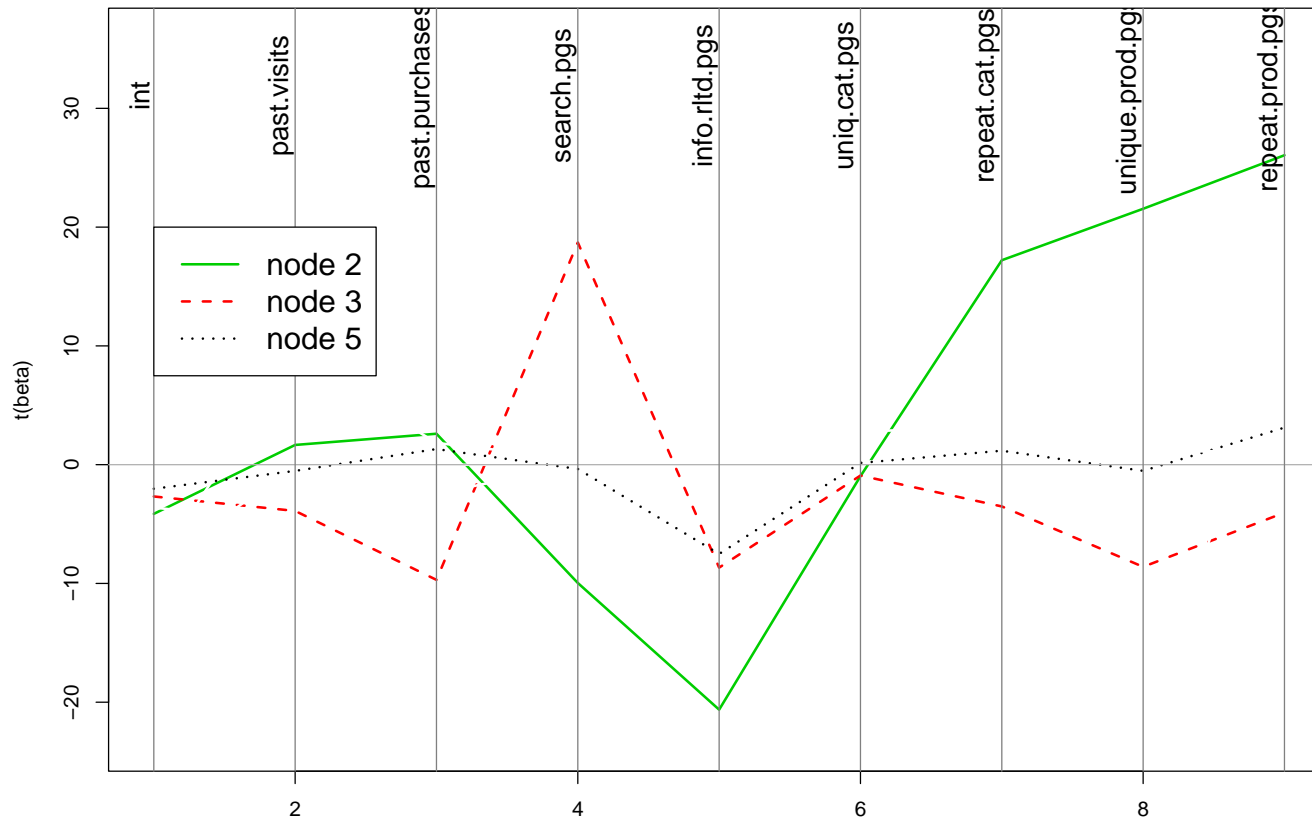
	n	Y=1	Node #
---# shopping-----	20338	7	1
pgs = 1			
---% product ----- prod/cat --	10437	170	2 *
pages <= 84 ratio <=1.29			
--- prod/cat --	376	17	3 *
- ---# shopping -- ratio > 1.29			
pgs = 2...9			
---% product -----	553	5	4
pages > 84			
-----% category -----	2826	360	5 *
- ---# shopping -- pages <=85			
pgs > 9			
-----% category -----	55	1	6
pages > 85			



Web marketing example: Treed logit

	n	Y=1	Node #
---# shopping----- pgs = 1	20338	7	1
---% product ----- pages <= 84 prod/cat -- ratio <=1.29	10437	170	2 *
--- prod/cat -- - ---# shopping -- ratio > 1.29	376	17	3 *
pgs = 2...9			
---% product ----- pages > 84	553	5	4
-----% category ----- ---# shopping -- pages <=85	2826	360	5 *
pgs > 9			
-----% category ----- pages > 85	55	1	6

Focus on nodes 2 (medium # pages) and 5 (lots of pages)



Node 2 (2-9 pgs)

large + coef for

- # repeat cat pgs
- # unique prod pgs
- # repeat prod pgs

Node 5 (> 9 pgs)
has small coefficients

Tree model is legitimate
(outperforms single logit on test data.)

⇒ If you view a moderate number of pages, looking at more product and category pages increases likelihood of buying.

The other example: What gets your articles cited?

- Tried several models:
 - Loglinear model (ie Poisson, log link).
 - Treed Poisson (log-linear model in each node).
 - Conventional tree for Poisson data (constant in each node).
- Assess performance using out-of-sample deviance (smaller is better). This was evaluated with 10-fold cross-validation.

Model	Mean deviance
Loglinear	5.60
Treed Poisson	4.51
Conventional tree	4.32

- We're not doing so well.

The other example: What gets your articles cited?

Why are we doing poorly?

- One (big) problem is *overdispersion*. If distribution was Poisson, then mean deviance should be about 1.0, meaning $E(Y) = Var(Y)$.
- Our model assumes no overdispersion.
- When there is overdispersion, our model may “chase after noise”.
- To reduce overfitting we manually fit a 2-node tree model:

Model	Mean deviance
Loglinear	5.60
Treed Poisson	4.51
Conventional tree	4.32
2-node Treed Poisson	3.90

The other example: What gets your articles cited?

This is suggestive, although the current analysis is a bit of a hack.

- **The two-node treed model:** The split is on articles: (0,1,2) vs. (3 or more).
- The linear model for the 0,1,2 group has nothing except articles significant. The linear model for the other group has most variables significant.

Coefficients:

	Estimate	Std.Err	z value	pval	
(Intercept)	2.181	0.4251	5.130	2.89e-07	***
pdoc	-0.078	0.0969	-0.808	0.419	
age	-0.076	0.0117	-6.512	7.42e-11	***
mar	0.739	0.1285	5.752	8.81e-09	***
doc	0.003	0.0004	6.627	3.42e-11	***
und	0.100	0.0229	4.392	1.12e-05	***
ag	-0.670	0.0992	-6.757	1.41e-11	***
arts	0.134	0.0120	11.180	2e-16	***

Null deviance: 703.28 on 56 degrees of freedom

Residual deviance: 350.27 on 49 degrees of freedom

Other uses for the MH tree algorithm: Boosting, Bagging?

- **Bagging** is just model averaging over multiple models. This is trivial for Bayesian procedures like ours.
- **Boosting** gradually builds up a sequences of approximations to $E(Y|x)$

$$F_0(x) = h(x; \hat{\theta}_0),$$

$$F_1(x) = h(x; \hat{\theta}_0) + h(x; \hat{\theta}_1),$$

$$F_2(x) = h(x; \hat{\theta}_0) + h(x; \hat{\theta}_1) + h(x; \hat{\theta}_2),$$

...

using a “weak learner” $h(x)$. This is accomplished by repeatedly refitting residuals.

- Two possible analogies:
 1. Interleave Boosting and MCMC - one MH step, one boosting step.
 2. Each weak learner is a treed model (heavily shrunk). Use MCMC to train them with a Gibbs-sampler like iteration between successive fits.

Conclusions:

- GLMs in terminal nodes enrich tree structure...
- ... But we need to allow for dispersion parameters.
- Bayesian gives stochastic search and some inference.
- Boosting an interesting possibility.
- Of course, there is much more to do - variable selection in nodes, interpreting a forest of trees, soft splits (very similar to hierarchical mixture of experts (Jordan and Jacobs, 1992)), etc.