Ch 8.1: Decision Trees

Lecture 24 - CMSE 381

Prof. Mengsen Zhang

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Fri, Oct 31, 2025

Announcements

Last time:

- Cubic Splines

This lecture:

- 8.1 Decision Trees

Announcements:

- HW #6 due Sunday (11/2)
- HW #7 due 11/9

21	F	10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/26
22	M	10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
23	W	10/29	Regression Splines	7.4	
24	F	10/31	Decision Trees	8.1	HW #6 Due Sun 11/2
25	M	11/3	Random Forests	8.2.1, 8.2.2	
26	W	11/5	Maximal Margin Classifier	9.1	
27	F	11/7	SVC	9.2	HW #7 Due Sun 11/9
28	M	11/10	SVM	9.3, 9.4	
29	W	11/12	Single Layer NN	10.1	
30	F	11/13	Multi Layer NN	10.2	HW #8 Due Sun 11/16
31	M	11/17	CNN	10.3	
32	W	11/19	Unsupervised learning / clustering	12.1, 12.4	
33	F	11/21	***Review***		HW #9 Due Sun 11/23
	M	11/24	**Midterm #3**		
	W	11/26	Virtual: Project Office Hours		
	F	11/28	Thanksgiving		
	M	12/1	Virtual: Project Office Hours		
	W	12/3	Virtual: Project Office Hours		
	F	12/5			**Project Due**
	M	12/8			
	W	12/10			
	F	12/12	**No final exam**		**Honors Project Due**

What will you learn today?

- How does a decision tree make decisions?
  - Given a tree and a data point with multiple predictors $(x_1, x_2, \cdots, x_p)$, you should be able to derive what the predicted outcome $\hat{y}$ should be.
  - You should be able to do this for both regression and classification decision trees.
- What are different parts of a decision tree called? What do they represent?
  - You should be able to point out what are: leaves (terminal nodes), internal nodes, branches/edges.
  - Given a simple example (e.g., two predictors), you should be able to point out which region in the predictor space map to which leave.
- How do you fit a decision tree?
  - What is the cost function for regression and classification tree respectively?
  - How does recursive binary splitting work?
  - How does it end?
- What meta-parameter of decision trees determine their flexibility?
- Why do you need to prune the tree?
- What are the advantages and disadvantages of using decision trees vs. linear regression?

Section 1

Decision Trees

Big idea

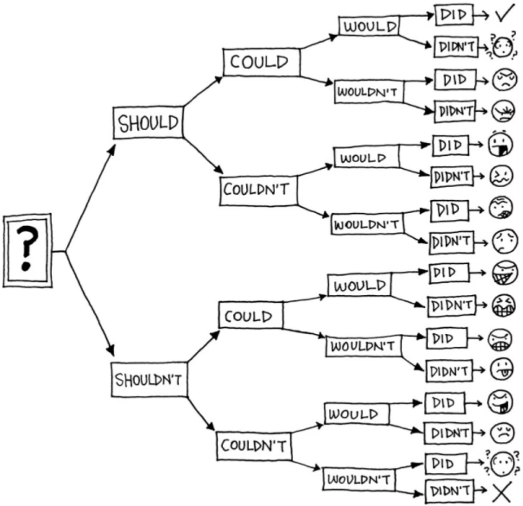


Image: https://marekbennett.com/2014/02/14/decision-tree/

Subset of Hitters data

	Hits	Years	Salary	LogSalary
1	81	14	475.0	6.163315
2	130	3	480.0	6.173786
3	141	11	500.0	6.214608
4	87	2	91.5	4.516339
5	169	11	750.0	6.620073
...	...	...	...	...
317	127	5	700.0	6.551080
318	136	12	875.0	6.774224
319	126	6	385.0	5.953243
320	144	8	960.0	6.866933
321	170	11	1000.0	6.907755

First decision tree example

	Hits	Years	LogSalary
1	81	14	6.163315
2	130	3	6.173786
3	141	11	6.214608
4	87	2	4.516339
5	169	11	6.620073
...	...	...	...
317	127	5	6.551080
318	136	12	6.774224
319	126	6	5.953243
320	144	8	6.866933
321	170	11	6.907755

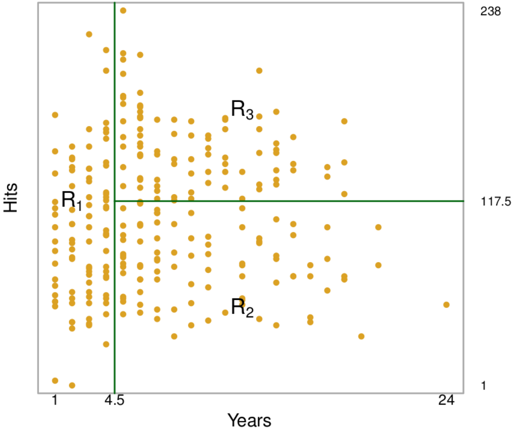
Test your understand: PollEv

Interpretation of example

Coding a regression decision tree

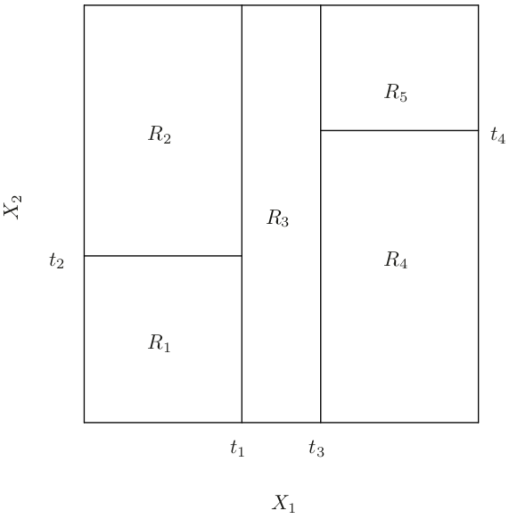
Regions defined by the tree

Viewing Regions Defined by Tree



How do we actually get the tree? Two steps

1. We divide the predictor space – that is, the set of possible values for $X_1, X_2, \cdots, X_p$ — into J distinct and non-overlapping regions, $R_1, R_2, \cdots, R_J$.
2. For every observation that falls into the region R_j, we make the same prediction = the mean of the response values for the training observations in R_j.



Step 1: How do we decide on R_js?

Goal:

Find boxes $R_1, \cdots, R_J$ that minimize

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$\hat{y}_{R_j}$ = mean response for training observations in jth box

Recursive Binary Splitting

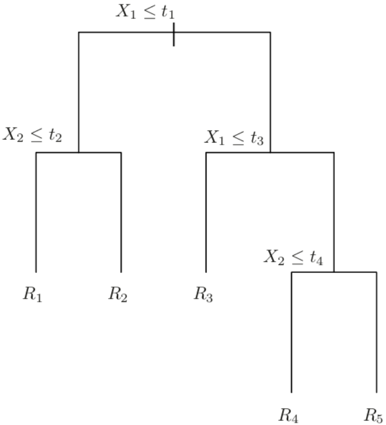
One split:

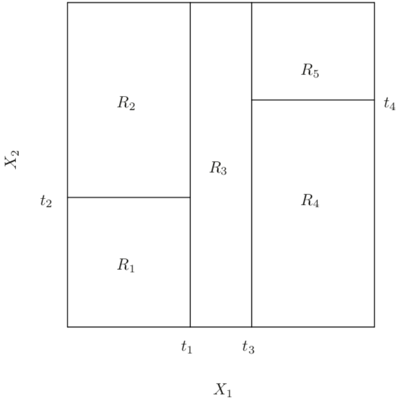
- Pick X_j and cutpoint s
- so that splitting into $\{X \mid X_j < s\}$ and $\{X \mid X_j \geq s\}$ results in largest possible reduction in RSS

$$R_1(j,s) = \{X \mid X_j < s\}$$
$$R_2(j,s) = \{X \mid X_j \geq s\}$$

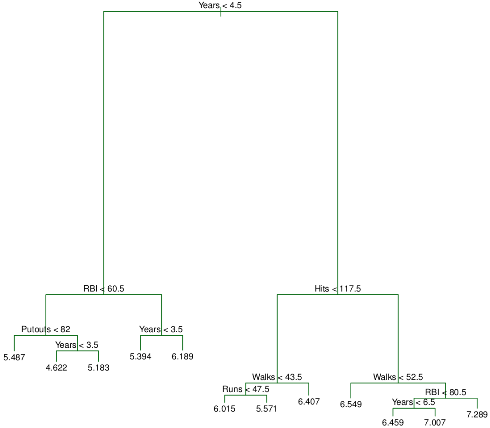
$$\sum_{i|x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i|x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Rinse and repeat



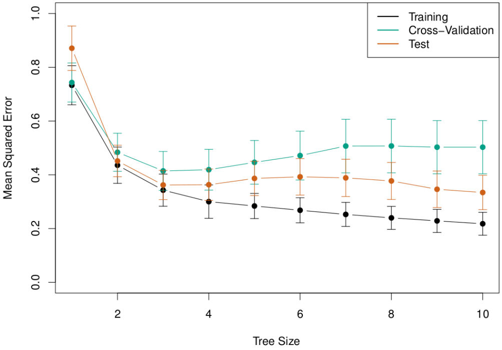


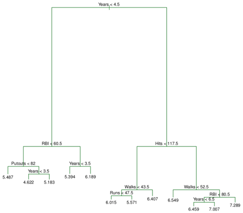
Pruning



Years < 4.5
RBI < 60.5
Hits < 117.5
Putouts < 82
Years < 3.5
Years < 3.5
5.487
4.622
5.183
5.394
6.189
Walks < 43.5
Runs < 47.5
6.015
5.571
6.407
Walks < 52.5
RBI < 80.5
Years < 6.5
6.549
6.459
7.007
7.289

Result of pruning

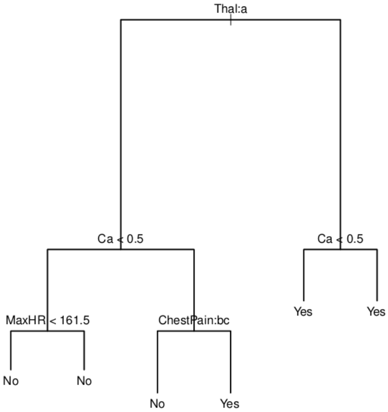




Section 2

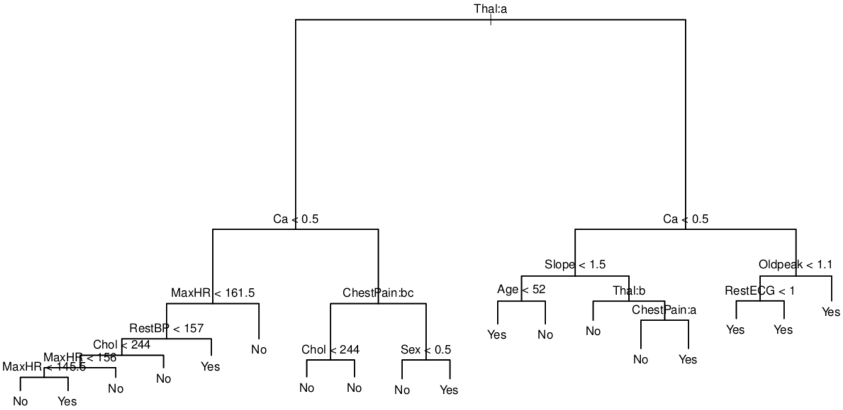
Classification Decision Tree

Basic idea



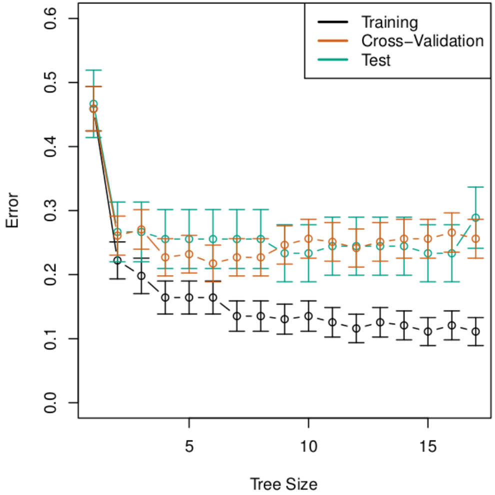
- $\hat{p}_{mk}$ = proportion of training observations in R_m from the kth class
- $E = 1 - \max_k(\hat{p}_{mk})$

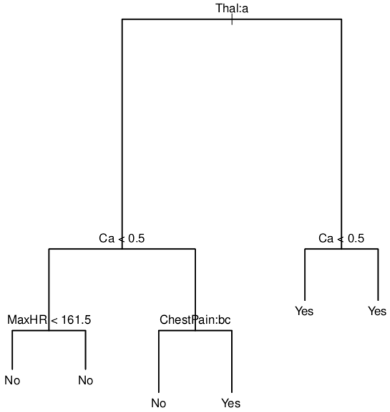
Example



Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Slope < 1.5
Oldpeak < 1.1
Age < 52
Thal:b
RestECG < 1
ChestPain:a
RestBP < 157
Chol < 244
MaxHR < 156
MaxHR < 145.5
Chol < 244
Sex < 0.5
No
Yes
No
No
Yes
No
No
No
No
No
Yes
Yes
No
No
No
Yes
Yes
Yes
Yes

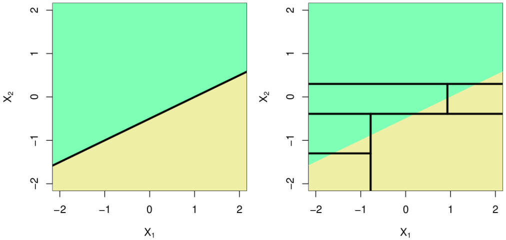
Pruning the example

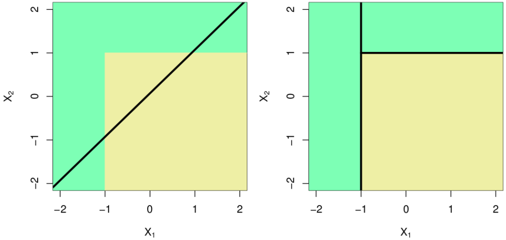




More coding!

Linear models vs trees





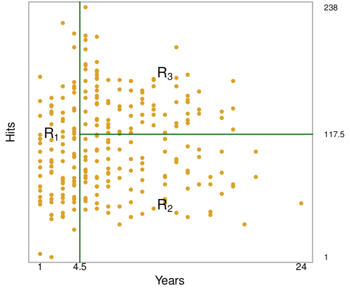
Pros/Cons

Pros:

Cons:

TL;DR

- Split into regions by greedily decreasing RSS
- Prune tree by using cost complexity
- Not robust - Next time, figure out how to aggregate trees



Next time

21	F	10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/26
22	M	10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
23	W	10/29	Regression Splines	7.4	
24	F	10/31	Decision Trees	8.1	HW #6 Due Sun 11/2
25	M	11/3	Random Forests	8.2.1, 8.2.2	
26	W	11/5	Maximal Margin Classifier	9.1	
27	F	11/7	SVC	9.2	HW #7 Due Sun 11/9
28	M	11/10	SVM	9.3, 9.4	
29	W	11/12	Single Layer NN	10.1	
30	F	11/13	Multi Layer NN	10.2	HW #8 Due Sun 11/16
31	M	11/17	CNN	10.3	
32	W	11/19	Unsupervised learning / clustering	12.1, 12.4	
33	F	11/21	***Review***		HW #9 Due Sun 11/23
	M	11/24	**Midterm #3**		
	W	11/26	Virtual: Project Office Hours		
	F	11/28	Thanksgiving		
	M	12/1	Virtual: Project Office Hours		
	W	12/3	Virtual: Project Office Hours		
	F	12/5			**Project Due**
	M	12/8			
	W	12/10			