SEEC Toolbox seminars

Classification and Regression Trees

Ian Durbach
ian.durbach@uct.ac.za

indurbach

iandurbach

SEEC - Statistics in Ecology, Environment and Conservation

What are trees? :)

- Trees are a type of *supervised statistical learning* method
- Very general: methods that relate a response variable y to a set of predictors X, with the aim of predicting the response for future observations
- Alternative to linear and logistic regression, neural networks, etc
- *Regression* trees for continuous response, *classification* for discrete

Example

- We will look at counts of *Aloe dichotoma* (now *Aloidendron dichotomum*) collected by Jack et al. (2016)
- Extensive roadside survey returned 1,138 transects containing aloes
- Our goal is to predict the number of trees in a transect
- Predictors are latitude, longitude, MAP, MAT

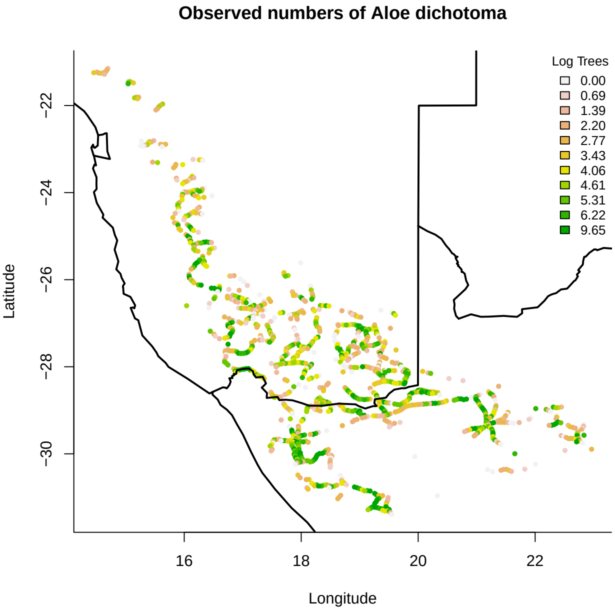
Example

```
> aloe <- read.csv("aloedichotoma.csv", header=TRUE)
> head(aloe)
```

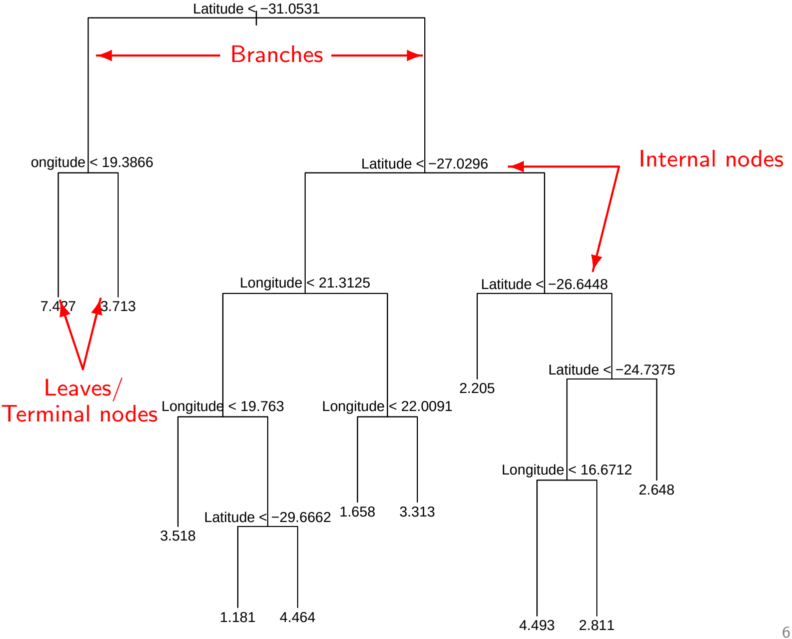
	ntrees	latitude	longitude	MAP	MAT
1	4	-21.14909	14.69328	111	21.7
2	129	-21.47578	15.04399	101	22
3	25	-21.47936	15.1299	130	21.6
4	245	-21.49967	15.04117	95	21.9
5	16	-21.18775	14.67602	108	21.6

We begin by considering only latitude and longitude as potential predictors

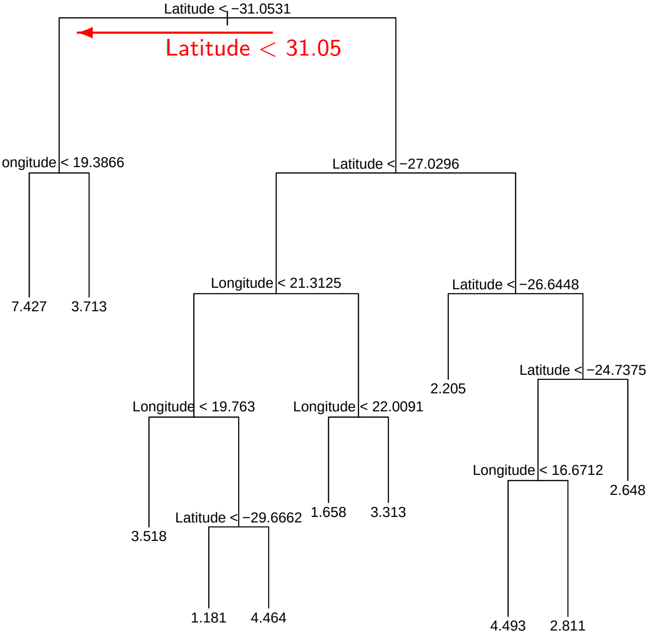
Example


Observed numbers of Aloe dichotoma
Log Trees
0.00
0.69
1.39
2.20
2.77
3.43
4.06
4.61
5.31
6.22
9.65
Latitude
−22
−24
−26
−28
−30
Longitude
16
18
20
22

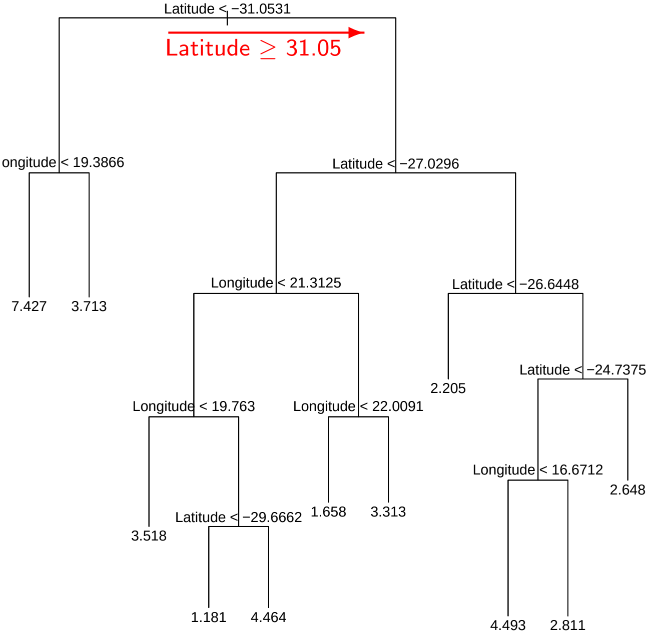
Example Regression Tree



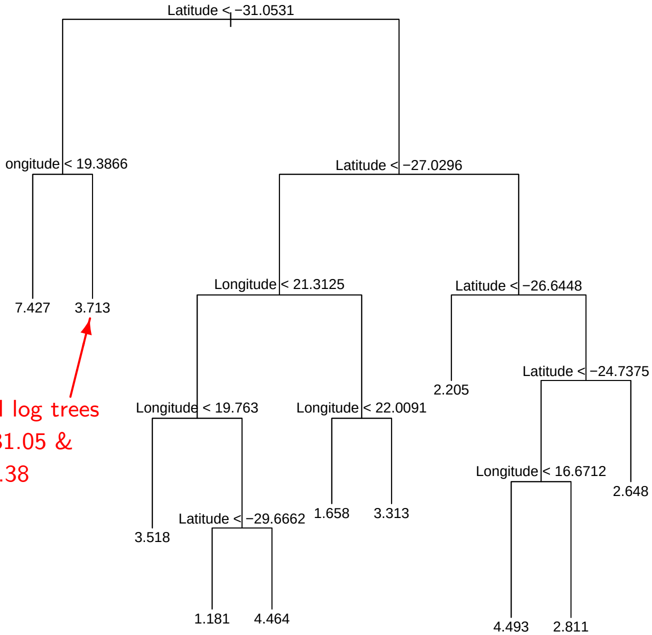
Example Regression Tree



Example Regression Tree

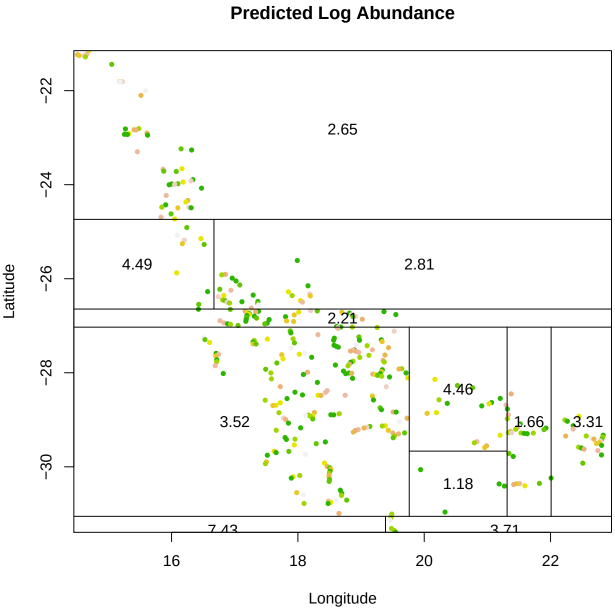


Example Regression Tree

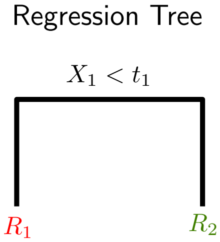


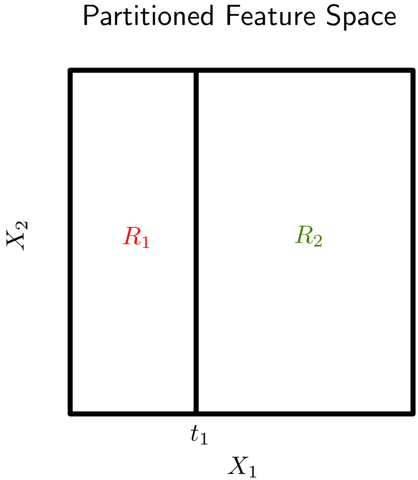
Predicted log trees if lat < 31.05 & lon ≥ 19.38

Partitioned Feature Space



Recursive Binary Splitting

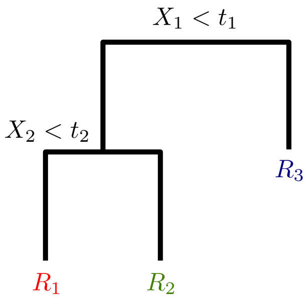




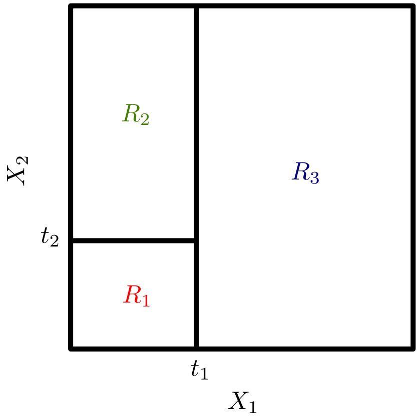
Need to choose **splitting criterion** (RSS)

Recursive Binary Splitting

Regression Tree

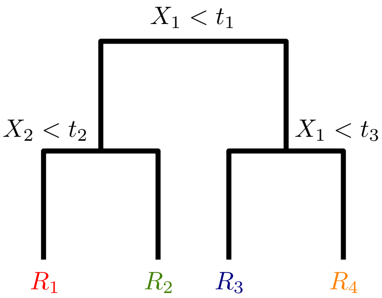


Partitioned Feature Space

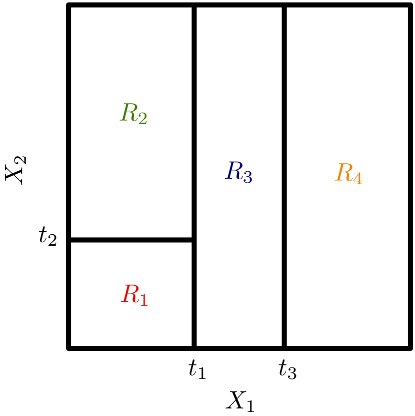


Recursive Binary Splitting

Regression Tree

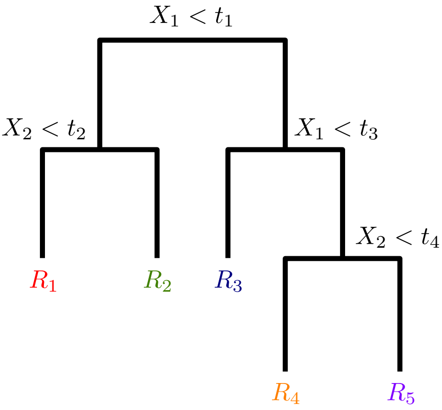


Partitioned Feature Space

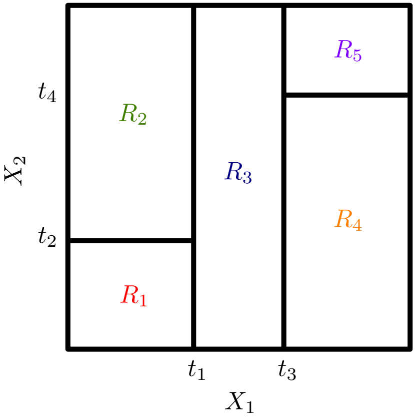


Recursive Binary Splitting

Regression Tree



Partitioned Feature Space



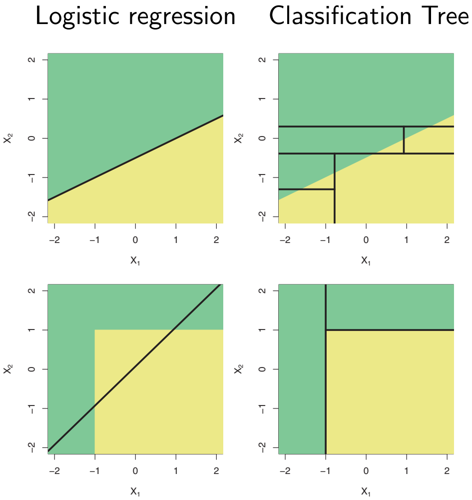
Need to choose **stopping criterion**

Classification Trees

- Used to predict a categorical response
- Similar to regression trees, except the predicted value in a region will now be the *most commonly occurring class*
- The *class proportions* in each terminal node give us an indication of the reliability of the prediction
- Suggested splitting criteria: Gini index, deviance (not % correct)

Trees versus Linear Models

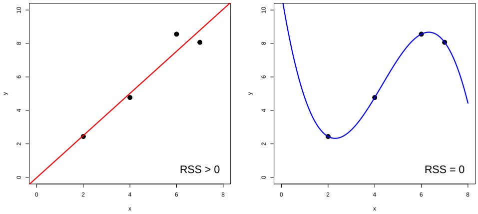
- We could use either logistic regression or decision trees for classification
- Which is better depends on the problem



Model validation

Don't overfit! Do validate!

- A model can be made to fit sample data arbitrarily well



- You are interested in how well your model does on *unseen* data
- Always do validation - **always always always!**

Model validation

Best practice

1. Divide your dataset in 3 parts: *training*, *validation* and *test* sets
2. Fit model on training data
3. Assess model on validation data
4. Choose model with the lowest *validation error*
5. Assess selected model on test data for final model $\Leftarrow$ this is your prediction error

Needs a lot of data

K-fold cross-validation

1. Divide data into K equal-size *folds*
2. Fit model model to all data excluding the kth fold
3. Assess performance using the kth fold
4. Repeat for all folds
5. Combine validation errors across folds

Most often $k = 10$. $K = n$, is *leave-one-out CV*

Cross-Validation

Example: 4-fold cross-validation for the linear model

	x	y
6	0.26	1.39
15	0.63	1.59
8	0.38	1.19
16	0.66	1.57
17	0.73	1.89
1	0.00	1.03
18	0.84	1.80
12	0.52	1.19
7	0.33	1.50
20	0.99	1.99
10	0.43	1.34
5	0.19	1.36
11	0.49	1.59
9	0.38	1.27
19	0.86	2.07
13	0.55	1.62
14	0.63	2.11
4	0.11	0.75
3	0.02	1.08
2	0.01	0.81

Randomise!

Cross-Validation

Example: 4-fold cross-validation for the linear model

	x	y	$\hat{y}$	$\hat{e}^2$
6	0.26	1.39	1.24	0.023
15	0.63	1.59	1.68	0.008
8	0.38	1.19	1.38	0.036
16	0.66	1.57	1.71	0.021
17	0.73	1.89	1.79	0.010
1	0.00	1.03		
18	0.84	1.80		
12	0.52	1.19		
7	0.33	1.50		
20	0.99	1.99		
10	0.43	1.34		
5	0.19	1.36		
11	0.49	1.59		
9	0.38	1.27		
19	0.86	2.07		
13	0.55	1.62		
14	0.63	2.11		
4	0.11	0.75		
3	0.02	1.08		
2	0.01	0.81		

Test set

Training set

$\hat{y} = 0.932 + 1.184x$

Cross-Validation

Example: 4-fold cross-validation for the linear model

	x	y	$\hat{y}$	$\hat{e}^2$
6	0.26	1.39	1.24	0.023
15	0.63	1.59	1.68	0.008
8	0.38	1.19	1.38	0.036
16	0.66	1.57	1.71	0.021
17	0.73	1.89	1.79	0.010
1	0.00	1.03	0.87	0.026
18	0.84	1.80	2.02	0.046
12	0.52	1.19	1.58	0.149
7	0.33	1.50	1.32	0.031
20	0.99	1.99	2.22	0.053
10	0.43	1.34		
5	0.19	1.36		
11	0.49	1.59		
9	0.38	1.27		
19	0.86	2.07		
13	0.55	1.62		
14	0.63	2.11		
4	0.11	0.75		
3	0.02	1.08		
2	0.01	0.81		

Test set

Training set

$\hat{y} = 0.867 + 1.363x$

Cross-Validation

Example: 4-fold cross-validation for the linear model

	x	y	$\hat{y}$	$\hat{e}^2$
6	0.26	1.39	1.24	0.023
15	0.63	1.59	1.68	0.008
8	0.38	1.19	1.38	0.036
16	0.66	1.57	1.71	0.021
17	0.73	1.89	1.79	0.010
1	0.00	1.03	0.87	0.026
18	0.84	1.80	2.02	0.046
12	0.52	1.19	1.58	0.149
7	0.33	1.50	1.32	0.031
20	0.99	1.99	2.22	0.053
10	0.43	1.34	1.42	0.006
5	0.19	1.36	1.14	0.049
11	0.49	1.59	1.48	0.011
9	0.38	1.27	1.36	0.010
19	0.86	2.07	1.92	0.023
13	0.55	1.62		
14	0.63	2.11		
4	0.11	0.75		
3	0.02	1.08		
2	0.01	0.81		

Training set
$\hat{y} = 0.921 + 1.154x$

Test set

Cross-Validation

Example: 4-fold cross-validation for the linear model

	x	y	$\hat{y}$	$\hat{e}^2$	
6	0.26	1.39	1.24	0.023	Training set $\hat{y} = 1.012 + 0.985x$
15	0.63	1.59	1.68	0.008	
8	0.38	1.19	1.38	0.036	
16	0.66	1.57	1.71	0.021	
17	0.73	1.89	1.79	0.010	
1	0.00	1.03	0.87	0.026	
18	0.84	1.80	2.02	0.046	
12	0.52	1.19	1.58	0.149	
7	0.33	1.50	1.32	0.031	
20	0.99	1.99	2.22	0.053	
10	0.43	1.34	1.42	0.006	
5	0.19	1.36	1.14	0.049	
11	0.49	1.59	1.48	0.011	
9	0.38	1.27	1.36	0.010	
19	0.86	2.07	1.92	0.023	
13	0.55	1.62	1.55	0.005	Test set
14	0.63	2.11	1.63	0.232	
4	0.11	0.75	1.12	0.135	
3	0.02	1.08	1.03	0.002	
2	0.01	0.81	1.02	0.044	

Cross-Validation

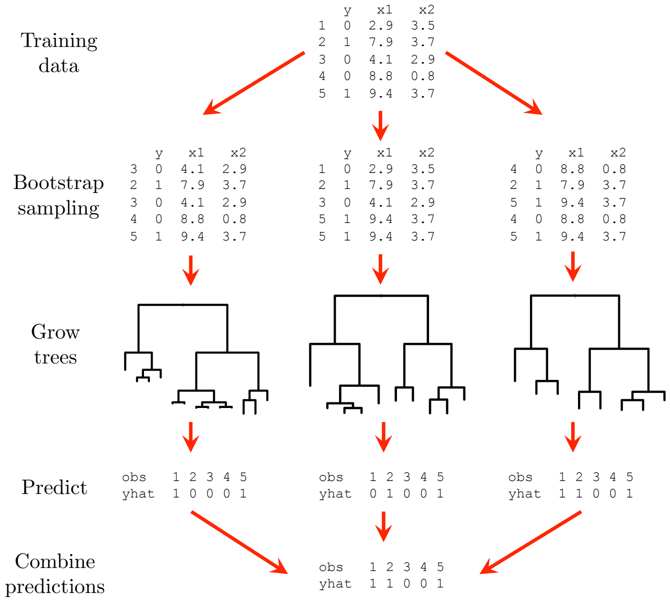
Example: 4-fold cross-validation for the linear model

	x	y	$\hat{y}$	$\hat{e}^2$
6	0.26	1.39	1.24	0.023
15	0.63	1.59	1.68	0.008
8	0.38	1.19	1.38	0.036
16	0.66	1.57	1.71	0.021
17	0.73	1.89	1.79	0.010
1	0.00	1.03	0.87	0.026
18	0.84	1.80	2.02	0.046
12	0.52	1.19	1.58	0.149
7	0.33	1.50	1.32	0.031
20	0.99	1.99	2.22	0.053
10	0.43	1.34	1.42	0.006
5	0.19	1.36	1.14	0.049
11	0.49	1.59	1.48	0.011
9	0.38	1.27	1.36	0.010
19	0.86	2.07	1.92	0.023
13	0.55	1.62	1.55	0.005
14	0.63	2.11	1.63	0.232
4	0.11	0.75	1.12	0.135
3	0.02	1.08	1.03	0.002
2	0.01	0.81	1.02	0.044

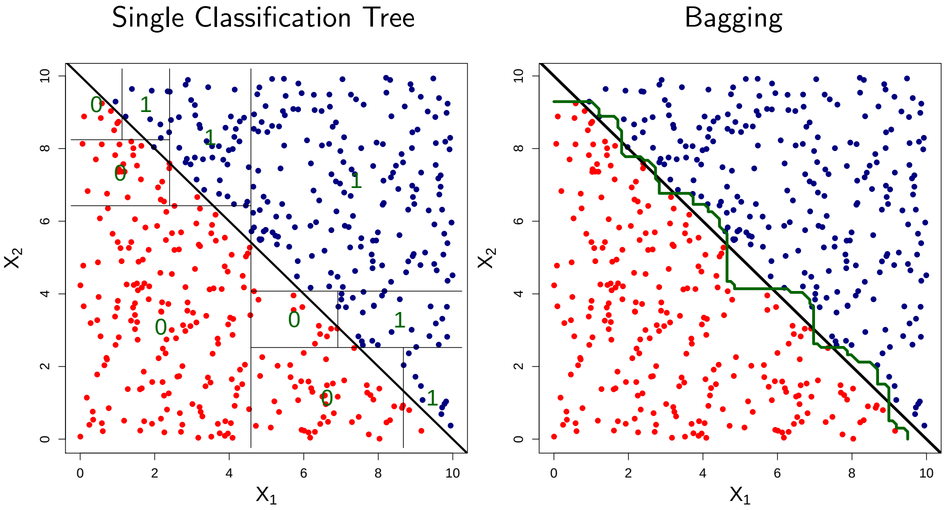
$$\begin{aligned}\text{CV error} &= \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 \\ &= 0.046\end{aligned}$$

Extensions

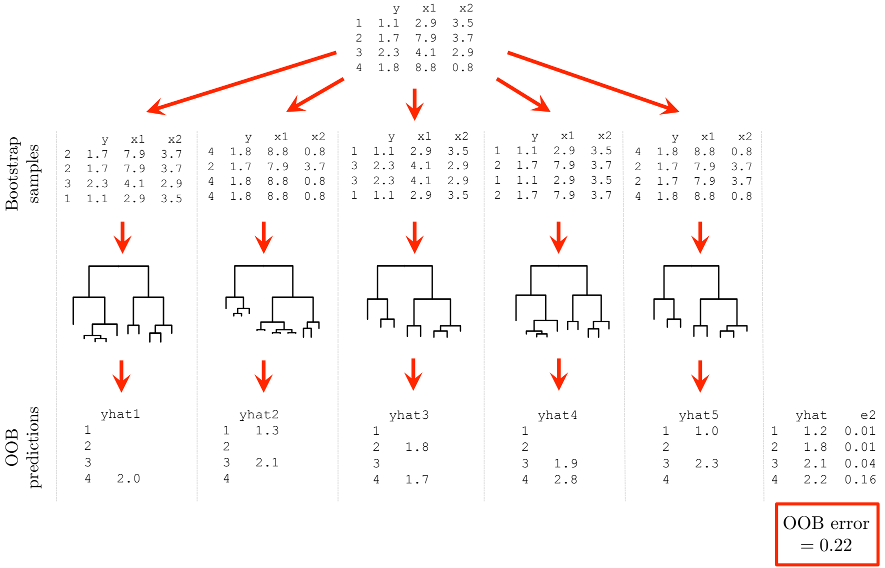
Bagging


Training data
y x1 x2
1 0 2.9 3.5
2 1 7.9 3.7
3 0 4.1 2.9
4 0 8.8 0.8
5 1 9.4 3.7
Bootstrap sampling
y x1 x2
3 0 4.1 2.9
2 1 7.9 3.7
3 0 4.1 2.9
4 0 8.8 0.8
5 1 9.4 3.7
y x1 x2
1 0 2.9 3.5
2 1 7.9 3.7
3 0 4.1 2.9
5 1 9.4 3.7
5 1 9.4 3.7
y x1 x2
4 0 8.8 0.8
2 1 7.9 3.7
5 1 9.4 3.7
4 0 8.8 0.8
5 1 9.4 3.7
Grow trees
Predict
obs 1 2 3 4 5
yhat 1 0 0 0 1
obs 1 2 3 4 5
yhat 0 1 0 0 1
obs 1 2 3 4 5
yhat 1 1 0 0 1
Combine predictions
obs 1 2 3 4 5
yhat 1 1 0 0 1

Why does bagging help?



Cross-validation for bagging: Out-of-Bag Error



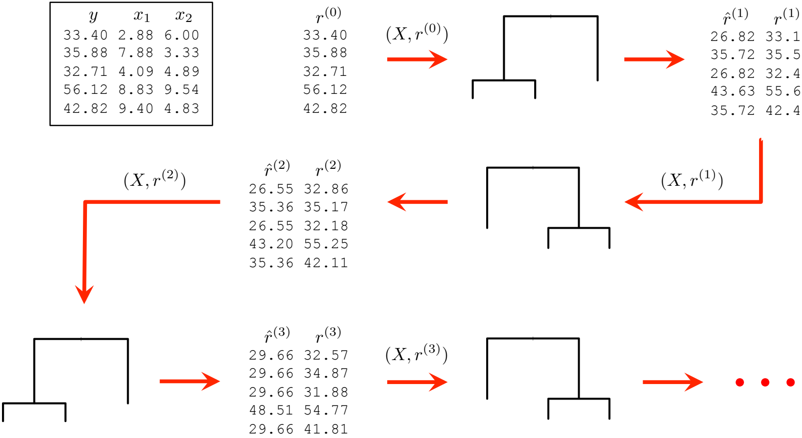
Random forests

- A small tweak that *decorrelates* the trees produced by bagging
- Each time a split is considered, *a random sample of* $m < p$ *predictors* are chosen as split candidates
- Bagging is a special case with $m = p$

Boosting

- Bagging and RFs: each tree is grown independently of all other trees
- Boosting: grows trees *sequentially* using information from previously trees
- First, grow a regression tree with a small number of splits, d
- The residuals of this tree are then treated as the response variable and used to grow another tree
- And so on...

Boosting



Boosting algorithm with
$d = 2$ and $\lambda = 0.01$

$$\hat{y} = \lambda \sum_{b=1}^{B} \hat{r}^{(b)}$$

Effects of predictor variables

Variable Importance

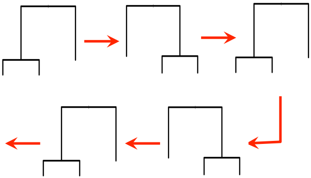
- No inference with trees – no significance testing
- Variable "importance": amount by which the splitting criterion improved
- Only a *relative* measure, and no *how* information

Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3
1	35.70	2.17	1.77	5.78
2	52.28	2.42	5.63	6.46
3	38.18	0.78	2.74	4.36
4	35.99	0.09	3.04	3.45
5	21.19	2.21	0.50	3.40
6	54.38	-2.64	3.63	6.81
7	23.59	2.26	0.23	4.52
8	32.27	0.45	1.34	5.62
9	47.84	0.43	4.24	5.61
10	38.87	-0.84	2.60	4.84

Data



Predictive Model

Construct a partial dependence plot for X_3

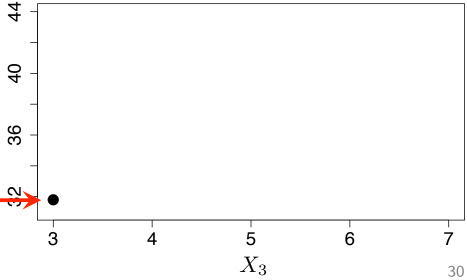
Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3	yhat
1	35.70	2.17	1.77	3	25.87
2	52.28	2.42	5.63	3	42.86
3	38.18	0.78	2.74	3	32.48
4	35.99	0.09	3.04	3	34.93
5	21.19	2.21	0.50	3	20.10
6	54.38	-2.64	3.63	3	41.99
7	23.59	2.26	0.23	3	18.80
8	32.27	0.45	1.34	3	26.70
9	47.84	0.43	4.24	3	39.79
10	38.87	-0.84	2.60	3	34.44

$$\hat{y}(X_3 = 3, X_{-3} = x_{-3,j})$$

$$\hat{y}(X_3 = 3) = 31.80$$



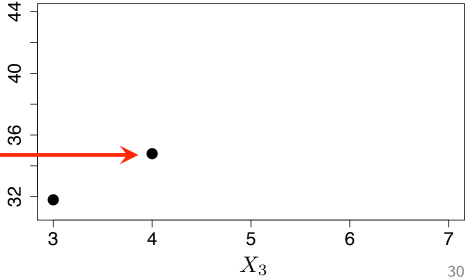
Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3	yhat
1	35.70	2.17	1.77	4	28.86
2	52.28	2.42	5.63	4	45.85
3	38.18	0.78	2.74	4	35.47
4	35.99	0.09	3.04	4	37.93
5	21.19	2.21	0.50	4	23.09
6	54.38	-2.64	3.63	4	44.98
7	23.59	2.26	0.23	4	21.79
8	32.27	0.45	1.34	4	29.69
9	47.84	0.43	4.24	4	42.78
10	38.87	-0.84	2.60	4	37.43

$\hat{y}(X_3 = 4, X_{-3} = x_{-3,j})$

$\hat{y}(X_3 = 4) = 34.79$



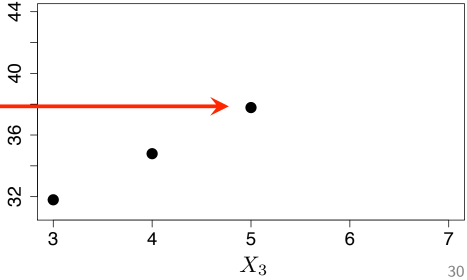
Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3	yhat
1	35.70	2.17	1.77	5	31.85
2	52.28	2.42	5.63	5	48.84
3	38.18	0.78	2.74	5	38.47
4	35.99	0.09	3.04	5	40.92
5	21.19	2.21	0.50	5	26.08
6	54.38	-2.64	3.63	5	47.98
7	23.59	2.26	0.23	5	24.78
8	32.27	0.45	1.34	5	32.69
9	47.84	0.43	4.24	5	45.77
10	38.87	-0.84	2.60	5	40.42

$\hat{y}(X_3 = 5, X_{-3} = x_{-3,j})$

$\hat{y}(X_3 = 5) = 37.78$



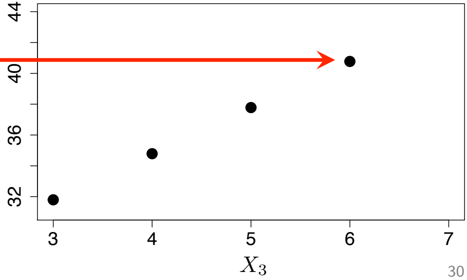
Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3	yhat
1	35.70	2.17	1.77	6	34.84
2	52.28	2.42	5.63	6	51.84
3	38.18	0.78	2.74	6	41.46
4	35.99	0.09	3.04	6	43.91
5	21.19	2.21	0.50	6	29.07
6	54.38	-2.64	3.63	6	50.97
7	23.59	2.26	0.23	6	27.77
8	32.27	0.45	1.34	6	35.68
9	47.84	0.43	4.24	6	48.77
10	38.87	-0.84	2.60	6	43.42

$\hat{y}(X_3 = 6, X_{-3} = x_{-3,j})$

$\hat{y}(X_3 = 6) =$ 40.77



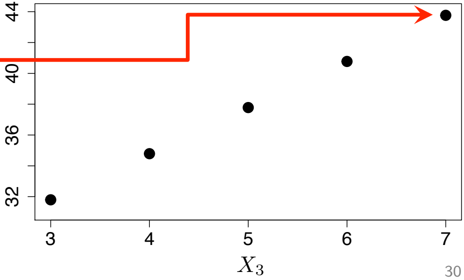
Constructing Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

	y	x1	x2	x3	yhat
1	35.70	2.17	1.77	7	37.84
2	52.28	2.42	5.63	7	54.83
3	38.18	0.78	2.74	7	44.45
4	35.99	0.09	3.04	7	46.90
5	21.19	2.21	0.50	7	32.07
6	54.38	-2.64	3.63	7	53.96
7	23.59	2.26	0.23	7	30.77
8	32.27	0.45	1.34	7	38.67
9	47.84	0.43	4.24	7	51.76
10	38.87	-0.84	2.60	7	46.41

$\hat{y}(X_3 = 7, X_{-3} = x_{-3,j})$

$\hat{y}(X_3 = 7) = 43.76$



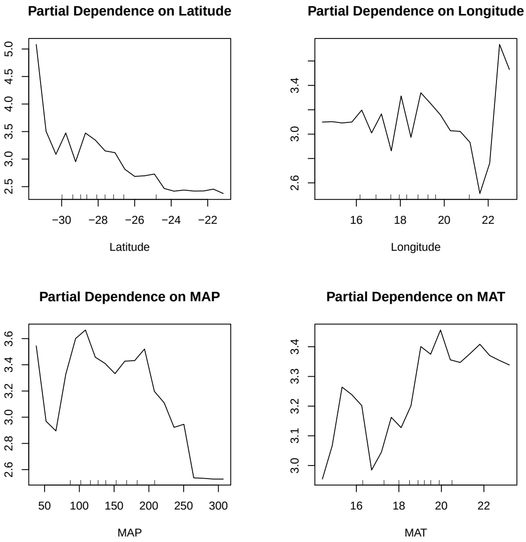
Partial Dependence Plots

Visually shows the effect of X_i on predictions *after accounting for other predictors*

- Fix all sample data except for the data for X_i
- Replace all data for X_i with a small value, say x
- Get mean prediction $\hat{y}$
- Increase x by a small amount and repeat
- Plot all $(x, \hat{y})$ pairs

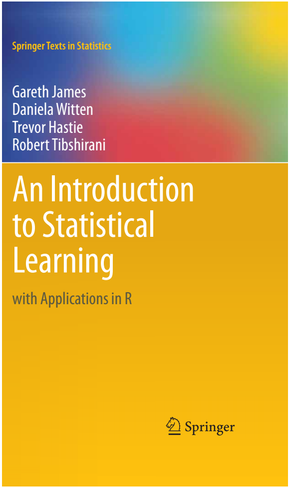
Note this is an *estimate* of the "true" partial dependency (since we use sample data)

Partial Dependence Plots for Aloe Abundance



Further resources

http://www-bcf.usc.edu/∼gareth/ISL/



- Death et al. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81:3178-3192.
- Cutler et al. (2007). Random forests for classification in ecology. Ecology 88(11): 2783–2792.
- Elith et al. (2008). A working guide to boosted regression trees. Journal of Animal Ecology 77: 802-813.
- Jack, S. L., Hoffman, M. T., Rohde, R. F., & Durbach, I. (2016). Climate change sentinel or false prophet? The case of *Aloe dichotoma*. Diversity and Distributions, 22(7), 745–757.
- iandurbach/trees-tutorial