Introduction to Tree-based Methods

Machine Learning for Ecology workshop

SEEC - Statistics in Ecology, Environment and Conservation

What are trees? :)

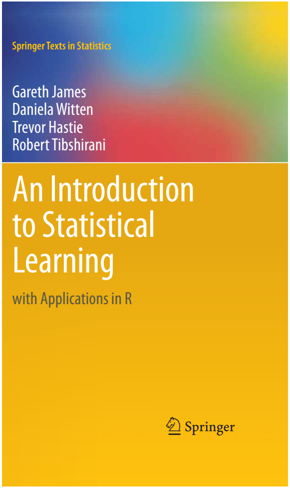
- Trees are a type of *supervised statistical learning* method
- Very general: methods that relate a response variable y to a set of predictors X, with the aim of predicting the response for future observations
- Alternative to linear and logistic regression, neural networks, etc
- *Regression* trees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

What we'll cover

- Classification and regression tree basics – model fitting and interpretation
- Model validation (training/test, cross-validation) and tree pruning
- Extensions to bagged trees, random forests, boosted trees
- Interpreting variable effects – importance and nature of relationships
- **Acknowledgement!** These slides adapted from UCT Stats Hons Analytics course developed by Miguel Lacerda and Stefan Britz

Some 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

Example

- We will look at counts of *Aloe dichotoma* (now *Aloidendron dichotomum*) collected by Jack et al. (2016)
- Extensive roadside survey returned 1,138/3,061 transects containing aloes
- Goal 1: to predict the presence of trees in a transect (classification)
- Goal 2: to predict the number of trees in transects containing at least one (regression)
- Predictors are latitude, longitude, MAP, MAT (and others)

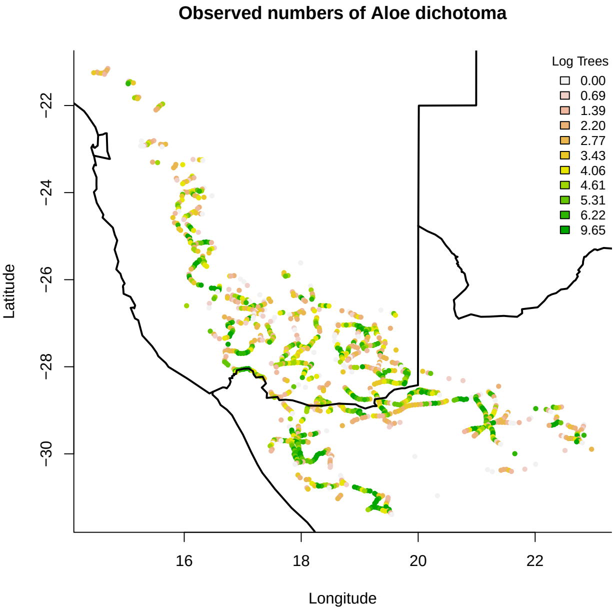
Example

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

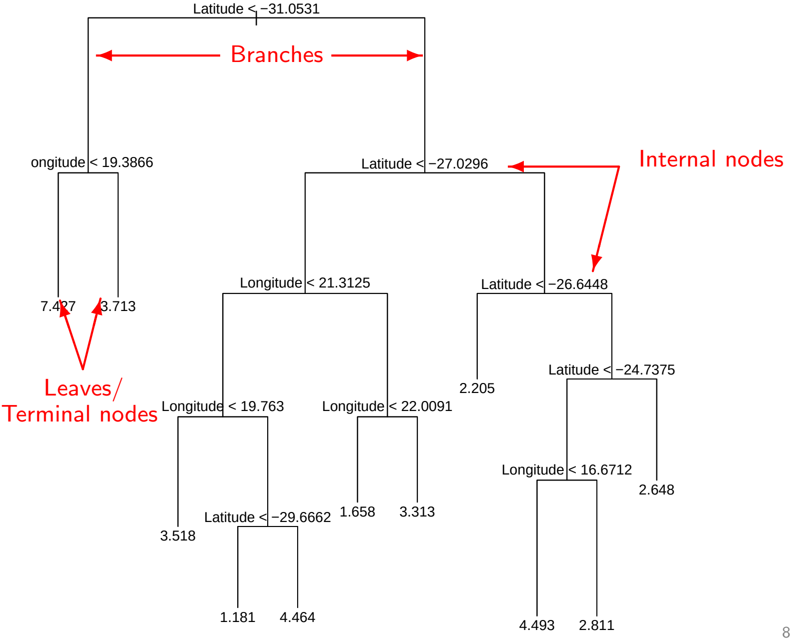
	ntrees	latitude	longitude	MAP	MAT
1	4	-21.14909	14.69328	111	21.7
2	129	-21.47578	15.04399	101	22
3	25	-21.47936	15.1299	130	21.6
4	245	-21.49967	15.04117	95	21.9
5	6	-21.18775	14.67602	108	21.6

We begin by considering only latitude and longitude as potential predictors

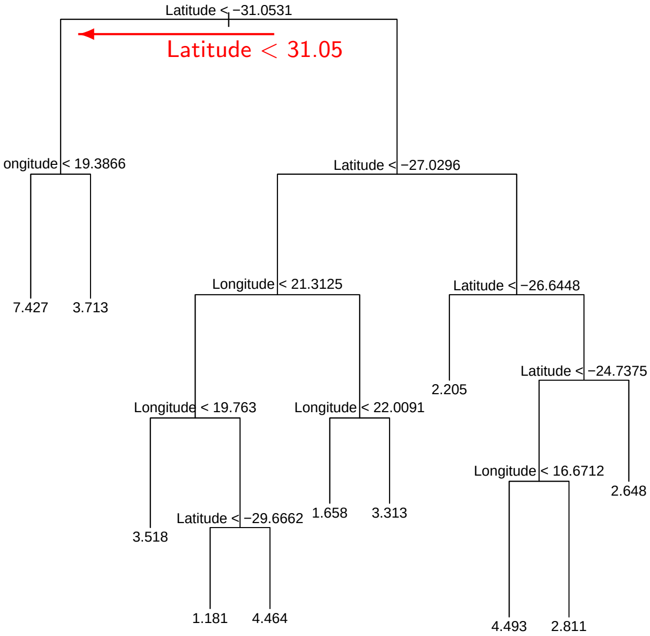
Example


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

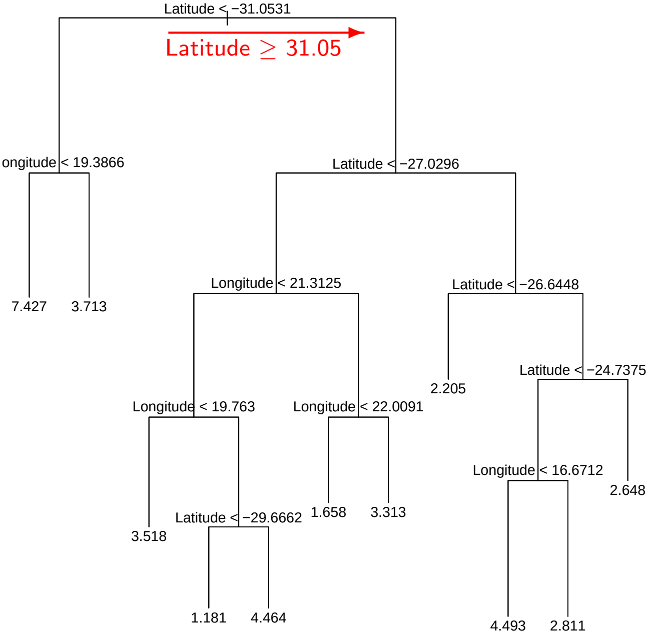
Example Regression Tree



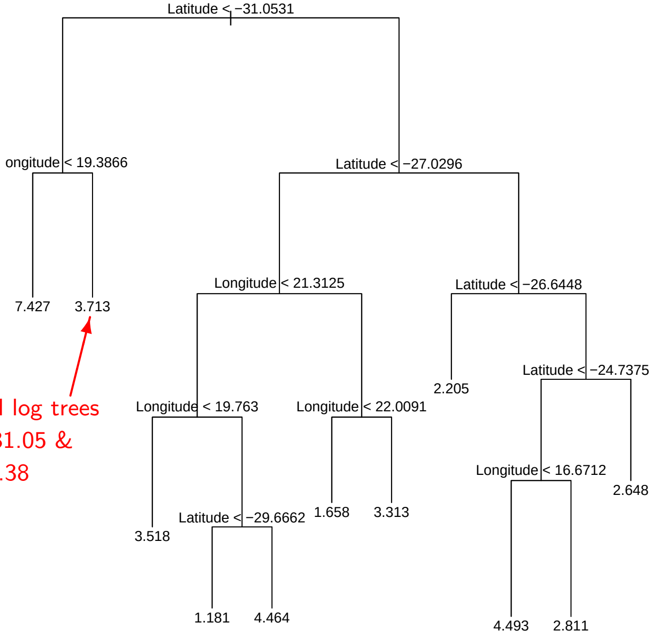
Example Regression Tree



Example Regression Tree

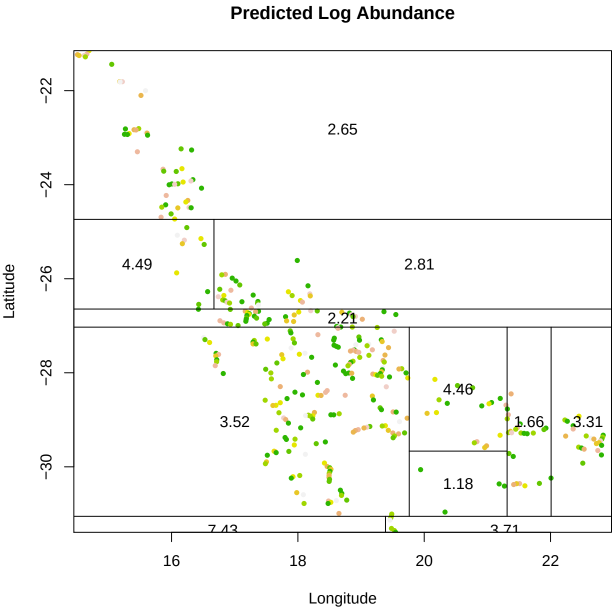


Example Regression Tree

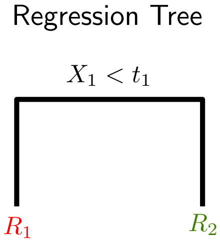


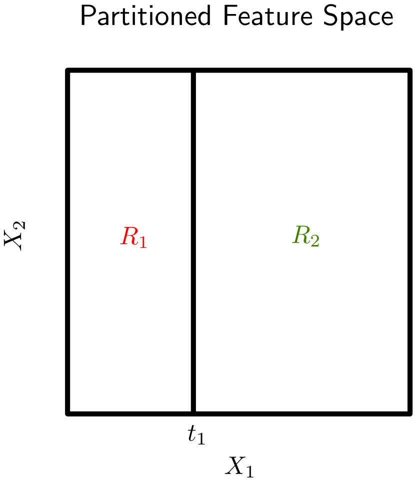
Predicted log trees
if lat < 31.05 &
lon ≥ 19.38

Partitioned Feature Space



Recursive Binary Splitting

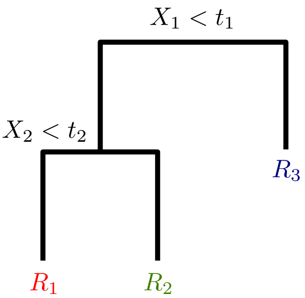




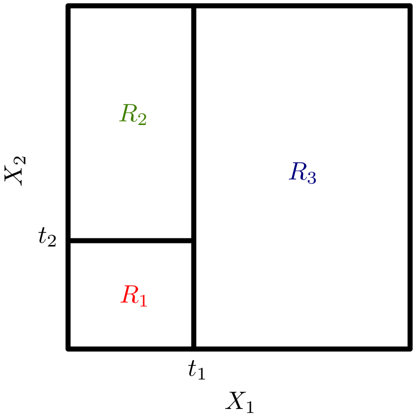
Need to choose **splitting criterion** (RSS)

Recursive Binary Splitting

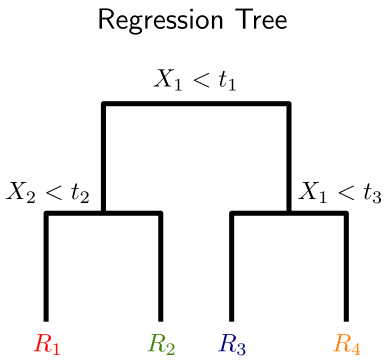
Regression Tree

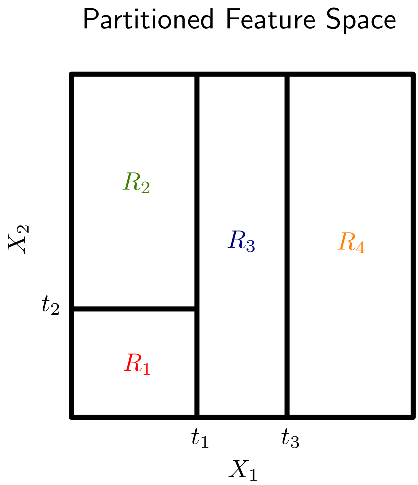


Partitioned Feature Space

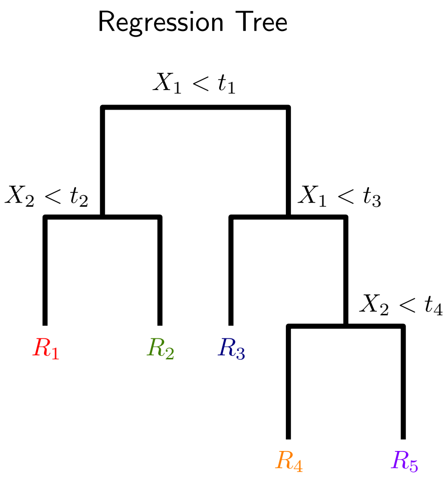


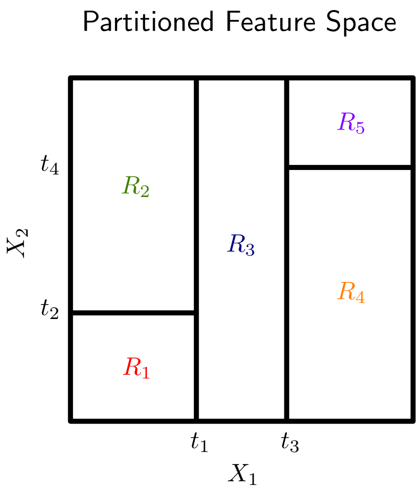
Recursive Binary Splitting





Recursive Binary Splitting



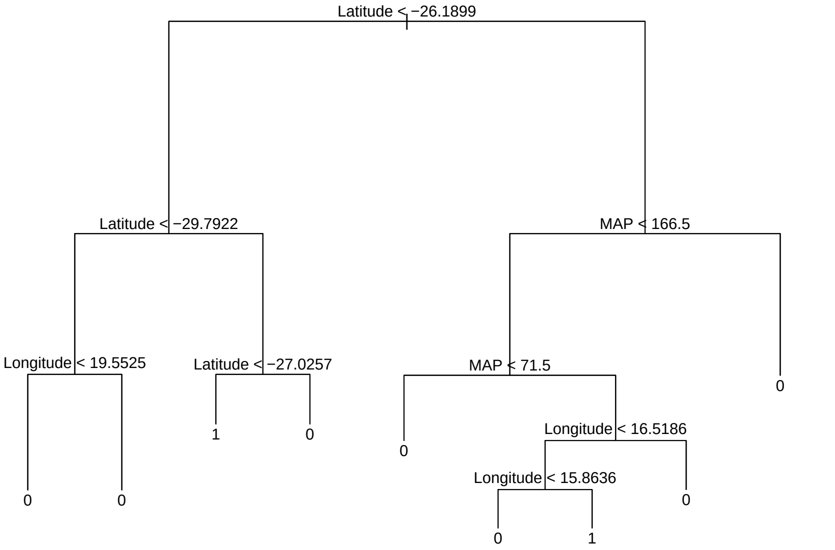


Need to choose **stopping criterion**

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)

Example Classification Tree



Splitting Criteria

- Residual sums of squares
- Classification error
- Gini index
- Deviance

Gini Index

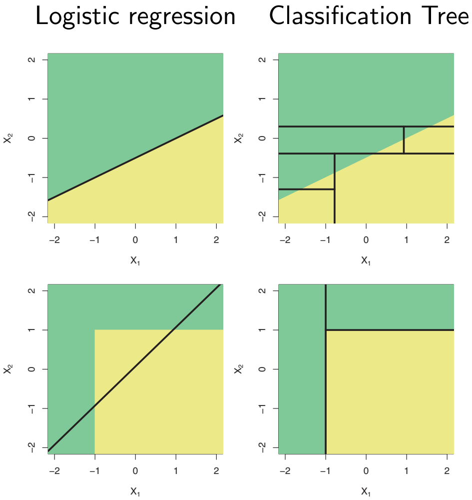
- Measures *node impurity* or variability of response within the terminal nodes
- Total Gini $=$ sum of Gini across all terminal nodes (G_j)
- For a binary response $G_j = 2p_j(1 - p_j)$
- Minimized when each terminal node include observations of only one class

Deviance

- A probability or likelihood-based measure
- Observations in node j come from a binomial distribution with parameter p_j
- Likelihood in node j is $L_j = p_j^{n_{j1}}(1 - p_j)^{n_{j2}}$
- Overall likelihood L is product of L_j
- Deviance is $D = -2 \log L$
- We want the model that makes the data most probable i.e. minimizes deviance

Trees versus Linear Models

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



Summary

- Introduced *trees* – binary recursive splitting methods
- *Regression* trees for continuous response, *classification* for discrete
- Tuning parameters: how to choose split, when to stop
- **Next:** Model validation and tree pruning