# FIT5149: Applied Data Analysis
# Tree-Based Methods

Faculty of Information Technology, Monash University, Australia

Week 8

# Outline

MONASH University

1. The Basic Decision Trees
   - Regression Tree
   - Classification Tree

2. Advanced Tree-based Methods
   - Bagging
   - Random forests
   - Boosting

## Where Does Tree-based methods sit in the Unit?

| | Discrete Labels | Continuous Labels |
|---|---|---|
| Supervised data | Classification<br>• Logistic regression<br>• Softmax regression<br>• LDA & QDA<br>• **Decision tree for classification**<br>• SVM<br>• GAM for classification | Regression<br>• Simple Linear regression<br>• Multiple linear regression<br>• Polynomial regression<br>• Splines<br>• **Decision tree for regression**<br>• GAM for regression |
| Unsupervised data | Clustering and dimensionality reduction<br>• PCA<br>• K-mean clustering<br>• Hierarchical clustering | |

Figure: Major topics covered in FIT5149

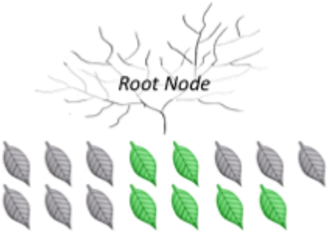
## Learning Outcomes

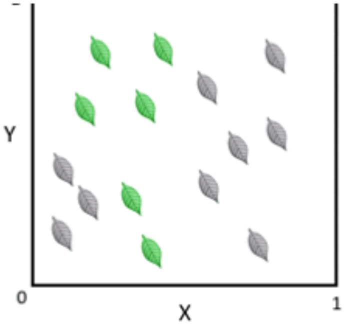
MONASH University

- Weekly learning outcomes
  - Differentiate between tree-based methods and the other methods
  - Understand the advantages and disadvantages of trees
  - Generate more powerful prediction model with bagging, random forest and boosting.
- Unit learning outcomes
  - Analyse data sets with a range of statistical, graphical and machine-learning tools;
  - Evaluate the limitations, appropriateness and benefits of data analytics methods for given tasks;
  - Assess the results of an analysis;

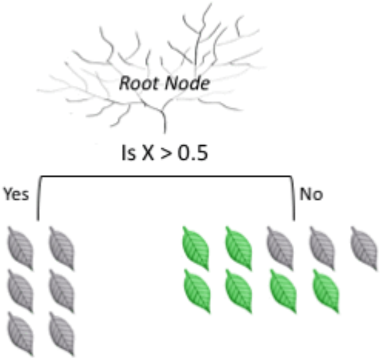
# The Basic Decision Trees

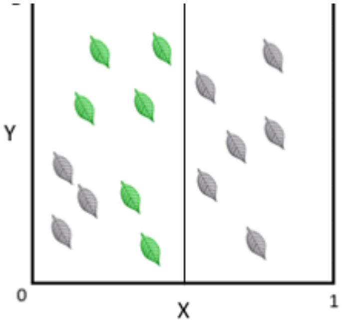
# The idea of decision tree



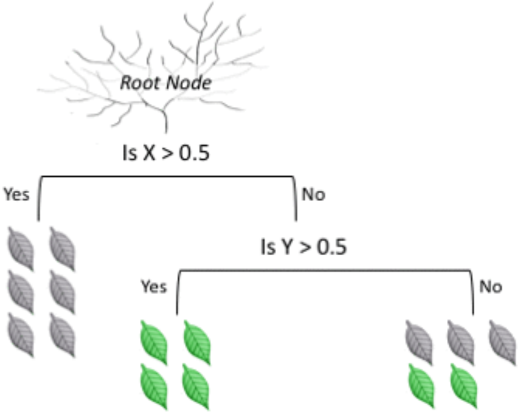



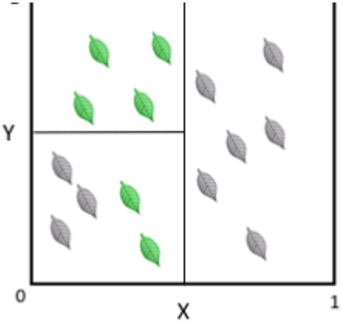
# The idea of decision tree



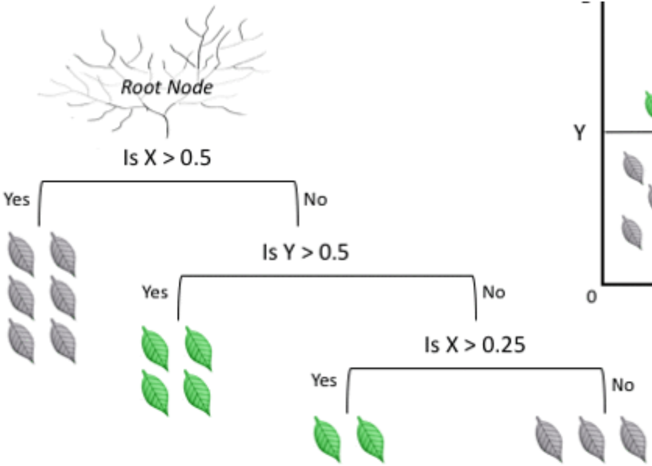



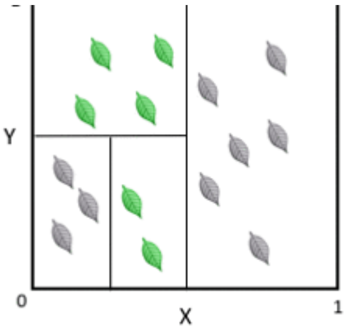
# The idea of decision tree



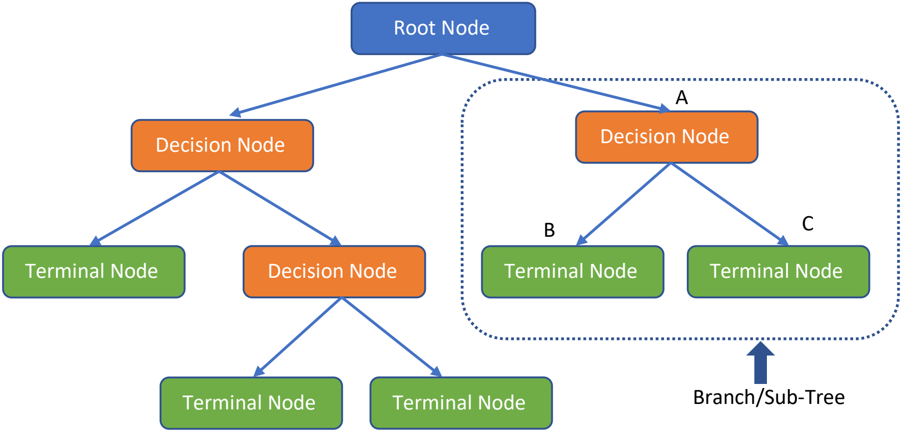
## The idea of decision tree



# The basic idea of decision trees

MONASH University



## Two types of trees: regression or classification tree

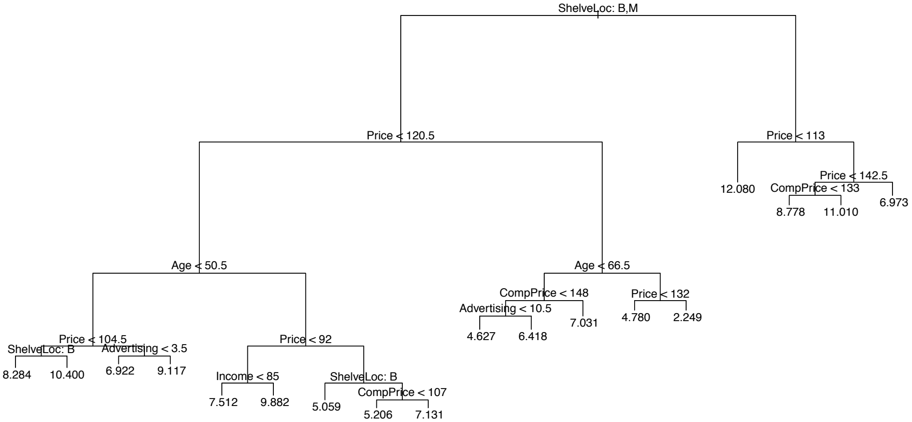



Figure: Task: Predict sales of child car seats at different stores

# Two types of trees: regression or classification tree

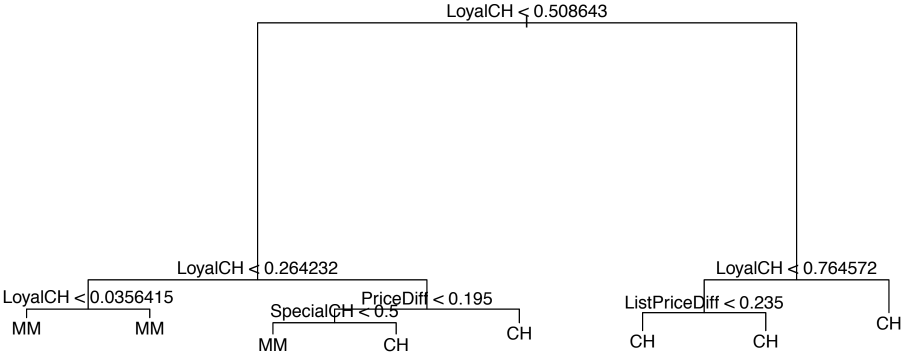



Figure: Task: Predict whether the customer purchased Citrus Hill (CH) or Minute Maid (MM) Orange Juice

## Outline

MONASH University

1. The Basic Decision Trees
   - Regression Tree
   - Classification Tree
2. Advanced Tree-based Methods

## Regression Tree: Predict quantitative variable

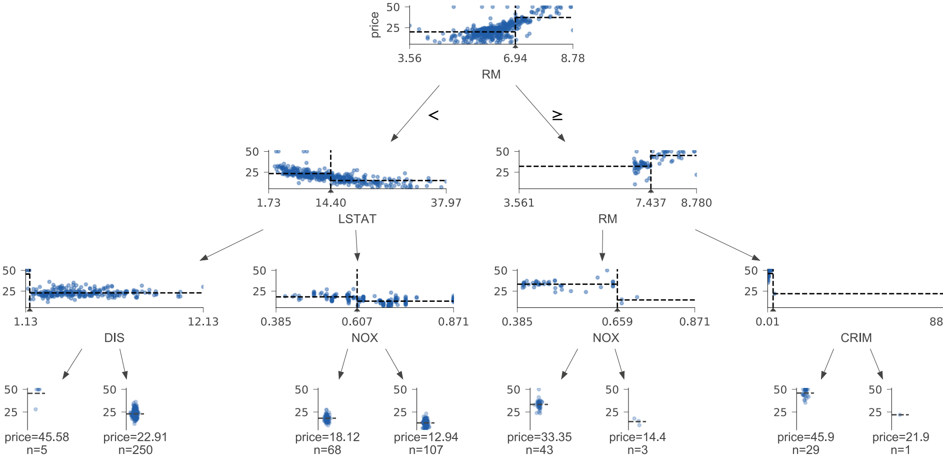
MONASH University

Task: predict media value of owner-occupied homes in $1000s in Boston

```
> str(Boston)
'data.frame':       506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Refer to the data dictionary for the meaning of variables.

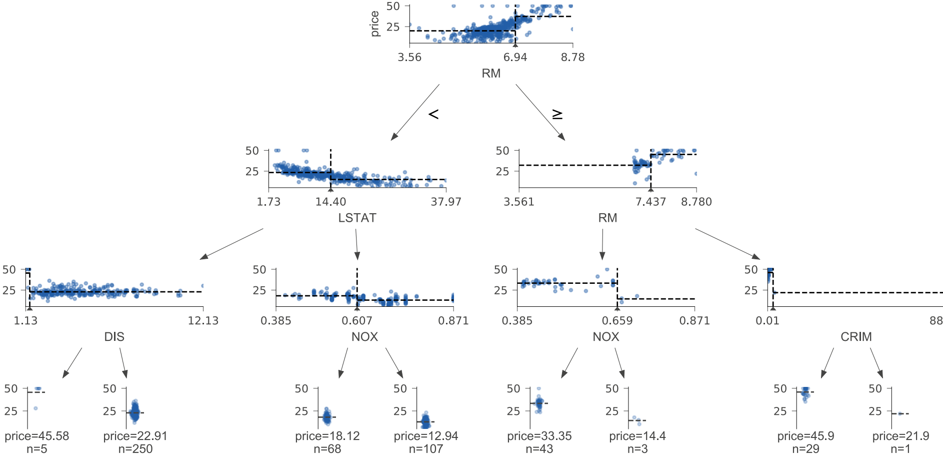
# Regression Tree for predicting house price[1]




[1] The plot is from https://explained.ai/decision-tree-viz/, which is slightly different from the lab results.

# Regression Tree: Interpretation

Discussion: What are the rules used to split the sample space? How many partitions are generated by the decision tree?

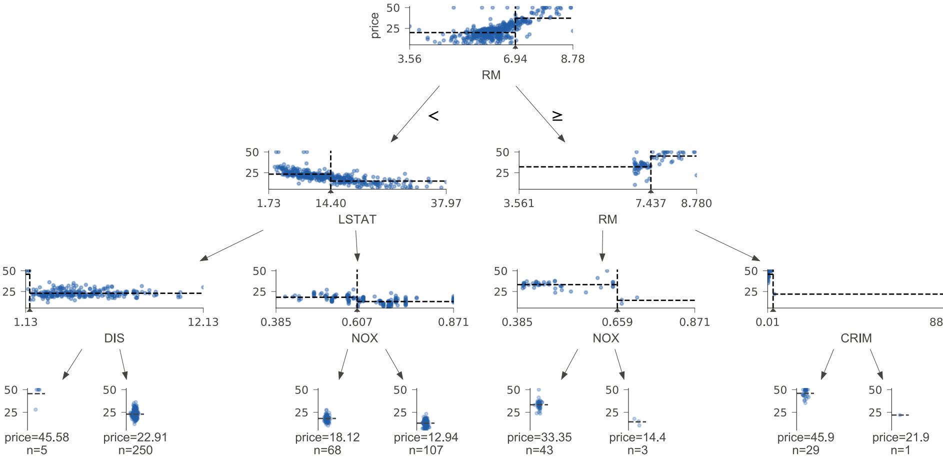


## Regression Tree: Interpretation

- $R_1 = \{X \mid RM < 6.94, LSTAT < 14.40, DIS < 1.13\}$
- $R_2 = \{X \mid RM < 6.94, LSTAT < 14.40, DIS \geq 1.13\}$
- $R_3 = \{X \mid RM < 6.94, LSTAT \geq 14.40, NOX < 0.607\}$
- $R_4 = \{X \mid RM < 6.94, LSTAT \geq 14.40, NOX \geq 0.607\}$
- $R_5 = \{X \mid RM \geq 6.94, RM < 7.437, NOX < 0.659\}$
- $R_6 = \{X \mid RM \geq 6.94, RM < 7.437, NOX \geq 0.659\}$
- $R_7 = \{X \mid RM \geq 6.94, RM \geq 7.437, CRIM < 0.01\}$
- $R_8 = \{X \mid RM \geq 6.94, RM \geq 7.437, CRIM \geq 0.01\}$

## Regression Tree: Interpretation

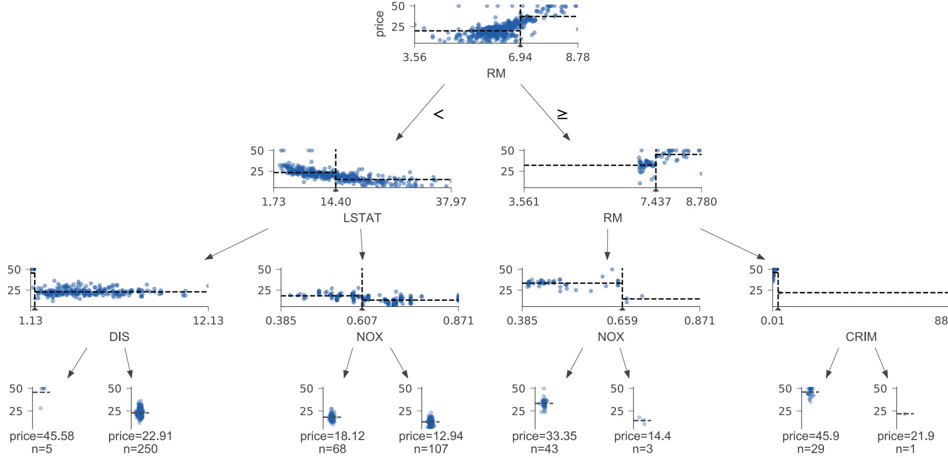
Discussion: What features that have large impact on the house price?



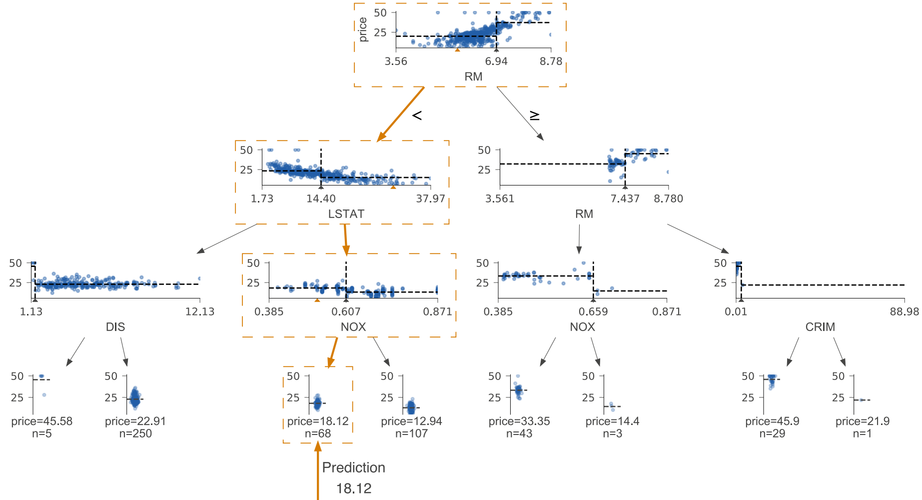
# Regression Tree: Prediction

Given a new house, can we predict its house price?

| CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.21 | 12.50 | 7.87 | 0.00 | 0.52 | 5.63 | 100.00 | 6.08 | 5.00 | 311.00 | 15.20 | 386.63 | 29.93 |



# Regression Tree: Prediction




| CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.21 | 12.50 | 7.87 | 0.00 | 0.52 | 5.63 | 100.00 | 6.08 | 5.00 | 311.00 | 15.20 | 386.63 | 29.93 |

# How to build a decision tree?

MONASH University

**Hint: Think about what the criteria we use to train a linear regression model taught in Week 3**

## How to build a decision tree?

**Hint: Think about what the criteria we use to train a linear regression model taught in Week 3**

- The idea: divide the predictor space into high-dimensional rectangles, or boxes.
- The goal: find boxes $R_1, ..., R_J$ that minimise the RSS:

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

  where $\hat{y}_{R_j}$: the mean response for the training observations within the $j$th box.
- Discussion: What is the problem?

## Recursive binary splitting

1. Select the predictor $X_j$ and the cutpoint $s$ such that splitting the predictor space into the regions $X \mid X_j < s$ and $X \mid X_j \geq s$ leads to the greatest possible reduction in RSS.

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

Minimize

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

## Recursive binary splitting

MONASH University

1. Select the predictor $X_j$ and the cutpoint $s$ such that splitting the predictor space into the regions $X \mid X_j < s$ and $X \mid X_j \geq s$ leads to the greatest possible reduction in RSS.

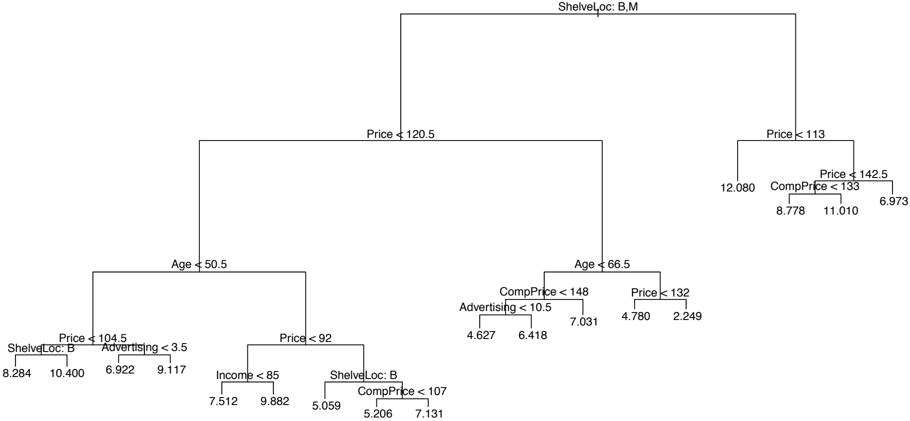
$$R_1(j, s) = \{X \mid X_j < s\} \text{ and } R_1(j, s) = \{X \mid X_j \geq s\}$$

   Minimize

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

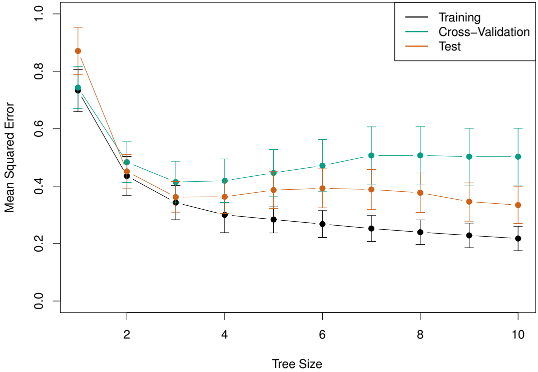
2. Repeat the process, looking for the best predictor and best cutpoint in order to split the data further so as to minimise the RSS within each of the resulting regions.

## Tree Pruning




Discussion: Bias and variance trade-off (**Hint: use the knowledge learned in Week 1**)

# Tree Pruning




- Bias and variance trade-off
  - A larger tree is likely to be overfitted, leading to poor test set performance
  - A smaller tree with fewer splits might lead to **lower variance** and better interpretation at the cost of a little **bias**.

## Tree Pruning: Cost Complexity Pruning

- A sequence of trees is indexed by a nonnegative tuning parameter $\alpha$.
- For each value of $\alpha$ there corresponds a subtree $T \subset T_0$ such that

$$\min_{T} \sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

  where
  - $|T|$: the number of terminal nodes of the tree $T$.
  - $R_m$: the rectangle corresponding to the mth terminal node.
  - $\hat{y}_{R_m}$: the mean of the training observations in $R_m$.
- Discussion:
  - How does $\alpha$ control the depth of the tree?
    (**hint: think about the regularization method taught in week 6**)
  - How can we choose $\alpha$?
    (**hint: think about the re-sampling methods taught in week 5**)

**How do decision trees work for real-word problems?**

- Make a prediction
- Interpret the prediction
- Visualise the trees

**How can decision trees be built from data?**

- Recursive binary splitting method
- The criterion used to make the splits

**Tree Pruning**

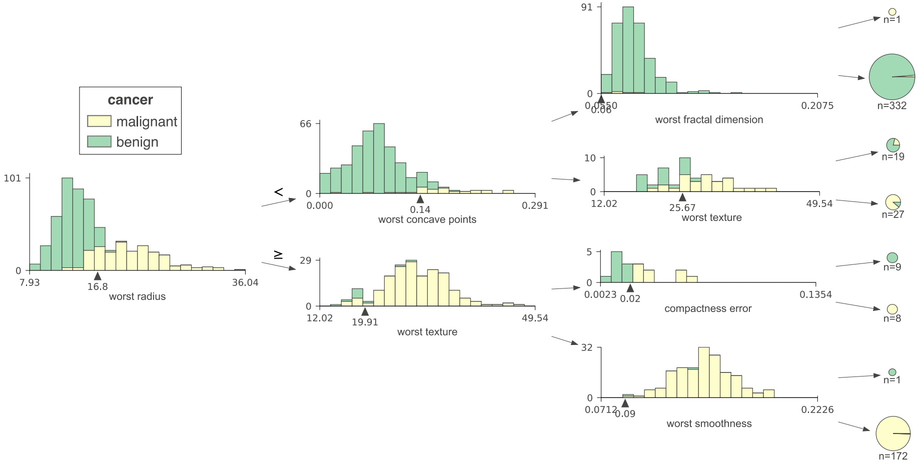
- Bias-Variance trade off
- Cost complexity pruning

## Outline

1. The Basic Decision Trees
   - Regression Tree
   - Classification Tree
2. Advanced Tree-based Methods

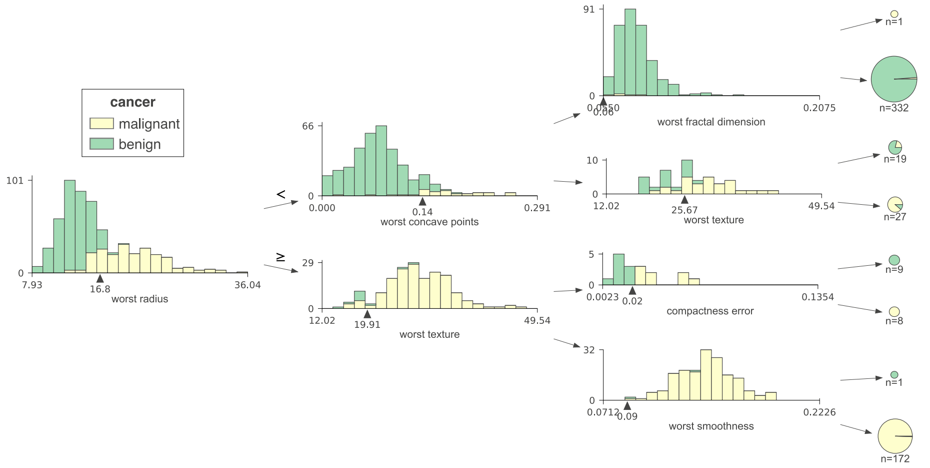
# Classification Tree

- Task: Predict whether the cancer is benign or malignant?
- Data: 569 samples from the clinical study at University of Wisconsin Hospitals



# Classification Tree

MONASH University

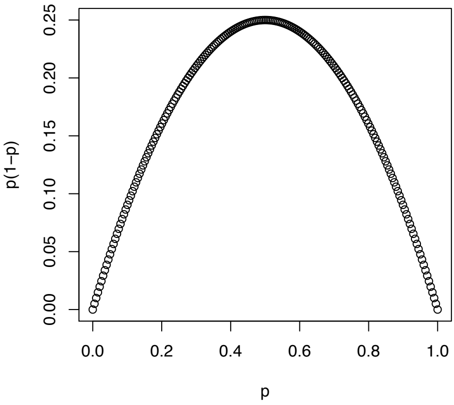



Discussion: How to prediction the qualitative response in a subtree? If we have an observation with "worst radius" = 15.4, "worst concave points" = 0.265, "worst texture" = 17.33, is the tumour malignant or benign?

## Classification Tree: Criteria for making binary splits

- Gini index: a measure of total variance across the K classes

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$




- The Gini index takes on a small value if all of the $\hat{p}_{mk}$'s are close to zero or one.
- A measure of node purity: a small value indicates that a node contains predominantly observations from a single class.

## Classification Tree: Criteria for making binary splits

- Entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk})$$

  - The entropy will take on a small value if the m-th node is pure

$$(0.9999316, 6.838223e-05, 1.418525e-13)$$

The Shannon's Entropy is 0.001.

$$(0.3090459, 0.279941, 0.4110131)$$

The Shannon's Entropy is 1.565.

## Classification tree: Hand on a toy dataset

MONASH University

Consider the following dataset:

| Obs. | X1 | X2 | X3 | Y |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 0 | 1 | 0 | -1 |
| 3 | 1 | 0 | 1 | -1 |
| 4 | 1 | 0 | 0 | 1 |

The three predictors are all categorical variables, taking binary values. Let us train a classification tree with the dataset, and call the tree $T_1$. If we fully train $T_1$ until each terminal node has data points of the same output label. Now, sketch the classification tree.

## Advantages and disadvantages of Trees

Discussion:

- Advantages
  -
  -
  -
- Disadvantages
  -
  -

# Advanced Tree-based Methods

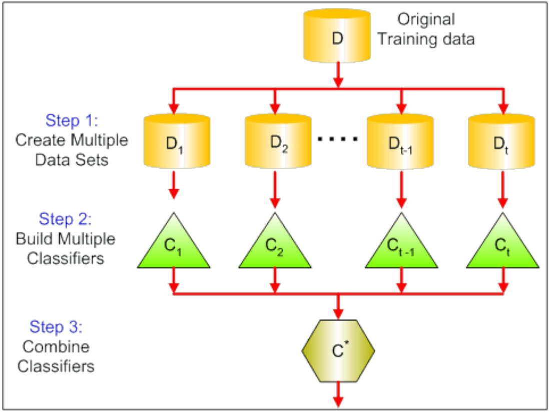
# Bagging

- A single decision tree: high variance
- Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method.
  - Given a set of n independent observations $Z_1, \ldots, Z_n$, each with variance $\sigma^2$, the variance of the mean $\bar{Z}$ of the observations is

$$\begin{aligned} \mathrm{var}(\bar{Z}) &= \mathrm{var}\left(\frac{1}{n}\sum_i X_i\right) \\ &= \frac{1}{n^2}\mathrm{var}(\sum_i X_i) = \frac{1}{n^2}\sum_i \mathrm{var}(X_i) \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

which means averaging a set of observations reduces variance.

# Bagging: bootstrapped training data[2]




$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

[2]The figure is adopted from datacamp

- Discussion:
  - When does Bagging make sense? (**Hint: Think about the bootsrap method discussed in week 5**)
  - What are the disadvantages?

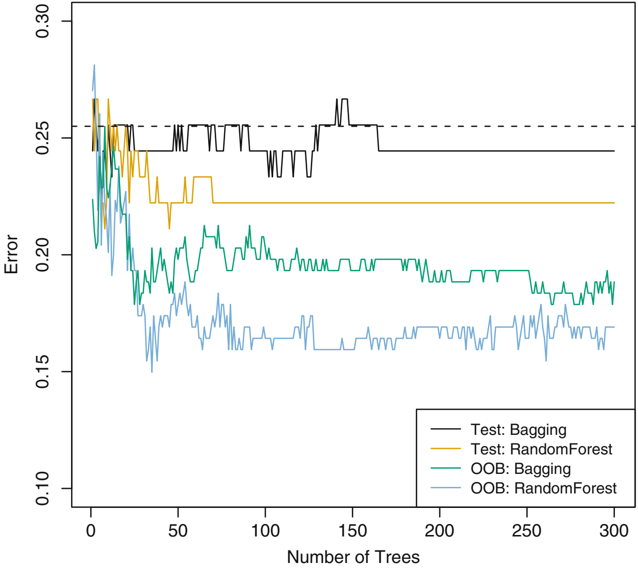
## OOB: Out-of-Bag Error

- No need to perform cross-validation.
- Recall that the key to bagging is that trees are repeatedly fit to bootstrapped subsets of the observations. Each bagged tree makes use of around two-thirds of the observations.

$$1 - (1 - 1/303)^{303} = 0.6327$$

- The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations.

# OOB Error on the Heart data set




The green traces show the OOB error of bagging, which in this case is considerably lower

## Random Forests



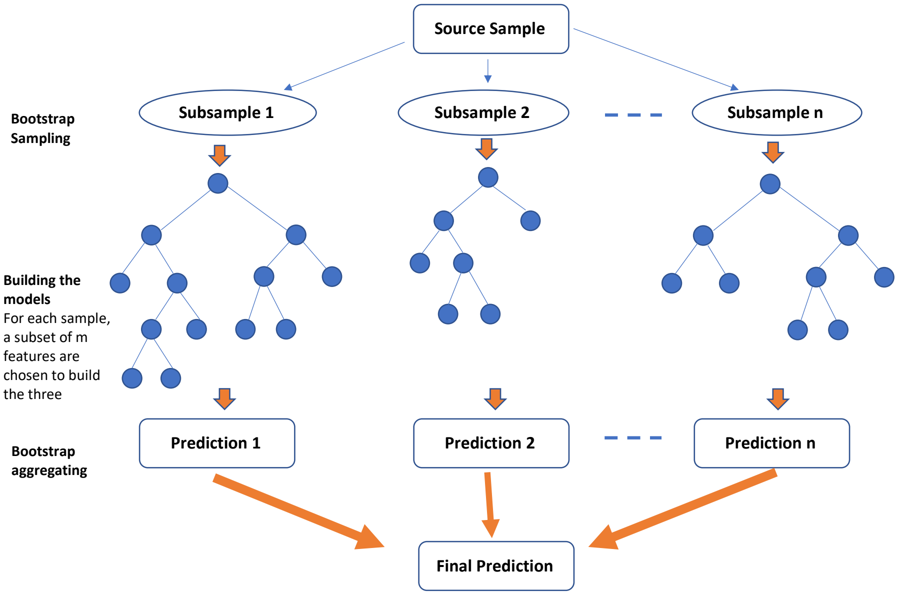
- In Bagging, trees can be correlated.
  - Most or all of the trees will use this strong predictor in the top split.
  - Averaging many highly correlated quantities does not lead to as large of a reduction in variance as averaging many uncorrelated quantities.
- How to de-correlates the trees?
  - A random sample of m predictors is chosen as split candidates from the full set of $p$ predictors.
  - Typically,

$$m \approx \sqrt{p}$$

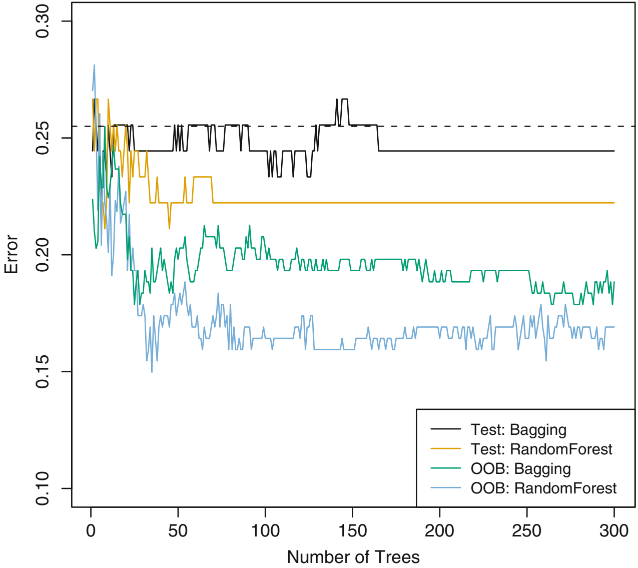
  - On average $(p - m)/p$ of the splits will not even consider the strong predictor.
- The main difference between bagging and random forests: the choice of predictor subset size $m$.

## Random Forests

MONASH University

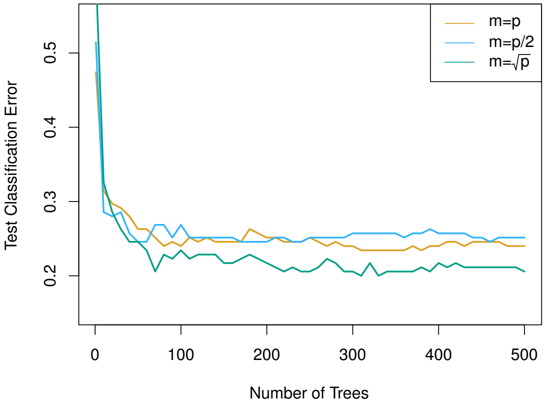


# Random forests on the Heart data set



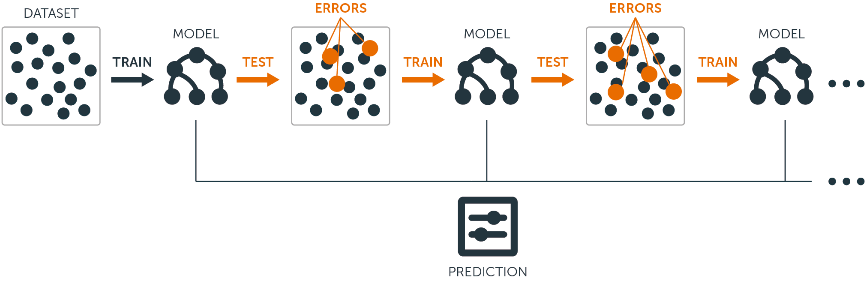

A reduction in both test error and OOB error over bagging

# Random forests on a high-dimensional biological data set




- A high-dimensional biological data set consisting of expression measurements of 4,718 genes measured on tissue samples from 349 patients.
- Each of the patient samples has a qualitative label with 15 different levels: either normal or 1 of 14 different types of cancer.
- We use random forests to predict cancer type based on the 500 genes that have the largest variance in the training set.

# Boosting




$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x)$$

## Boosting on gene expression data

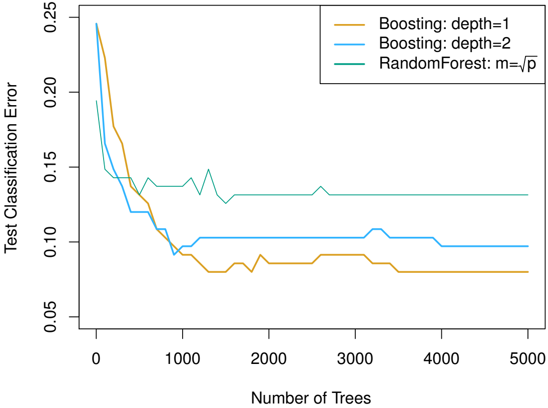
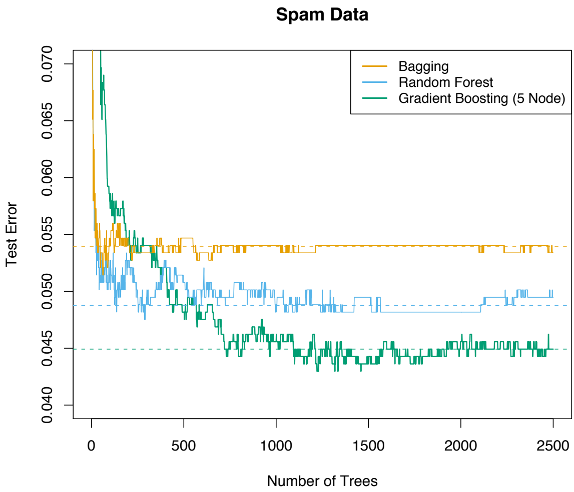



Figure: We applied boosting to the 15-class cancer gene expression data set, to develop a classifier that can distinguish the normal class from the 14 cancer classes.

- The test error is displayed as a function of the number of trees.
- For the two boosted models, $\lambda = 0.01$.
- The test error rate for a single tree is 24%.
- The standard errors are around 0.02, making none of these differences significant.

## Random Forests vs Boosting




from *Elements of Statistical Learning, chapter 15.*

## Summary

- Decision trees are simple and interpretable models for regression and classification
- However they are often not competitive with other methods in terms of prediction accuracy
- Bagging, random forests and boosting are good methods for improving the prediction accuracy of trees. They work by growing many trees on the training data and then combining the predictions of the resulting ensemble of trees.
- The latter two methods— random forests and boosting— are among the state-of-the-art methods for supervised learning. However their results can be difficult to interpret.

## Summary

- Regression/Classification trees
- Bagging, Random forest and boosting
- Reading materials:
  - "Tree-Based method", Chapter 8 of "Introduction to Statistical Learning", 6th edition
- **Optional** reading materials:
  - Elements of Statistical Learning, chapters 10 and 15
- Acknowledgement:
  - Figures in this presentation were taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
  - Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani