

ETC3250/5250: Introduction to Machine Learning

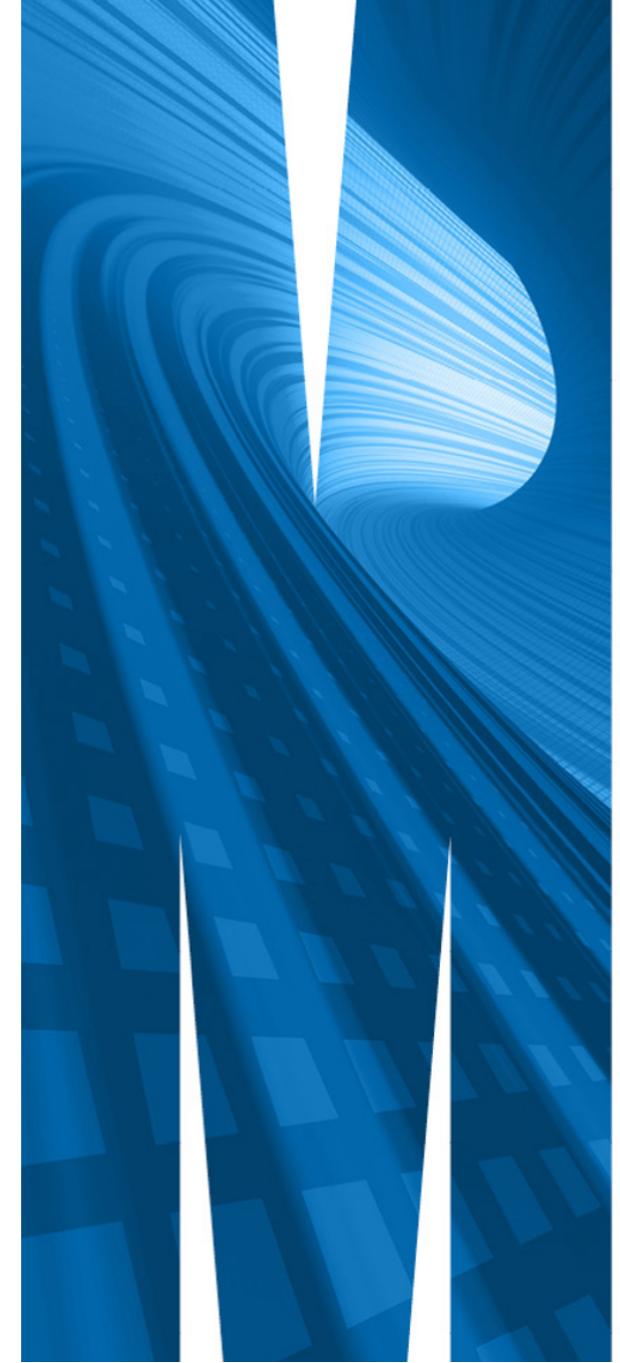
Classification Trees

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

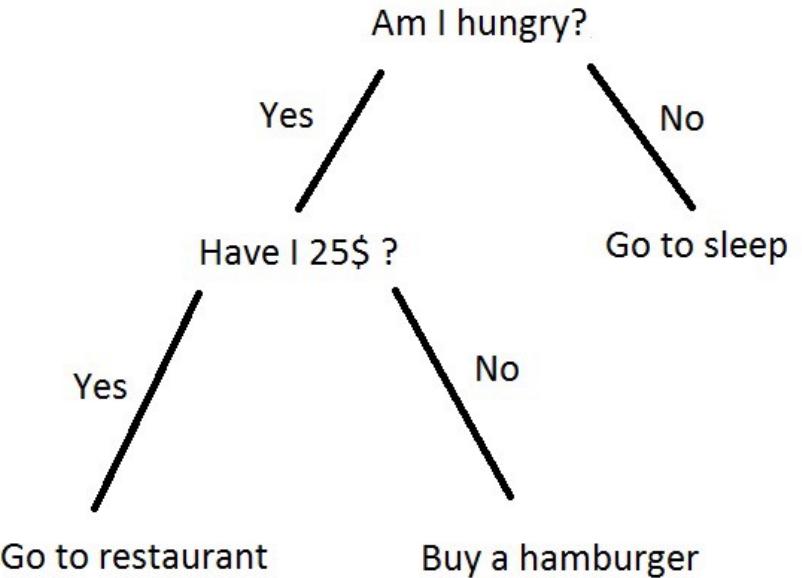
✉ ETC3250.Clayton-x@monash.edu

CALENDAR
Week 6a



What is a decision tree?

Tree based models consist of one or more of nested **if-then** statements for the predictors that partition the data. Within these partitions, a model is used to predict the outcome.



Source: [Egor Dezhic](#)

Classification trees

- A classification tree is used to predict a **categorical response** and regression tree is used to predict a quantitative response
- Use a recursive binary splitting to grow a classification tree. That is, sequentially break the data into two subsets, typically using a single variable each time.
- The predicted value for a new observation, x_0 , will be the **most commonly occurring class** of training observations in the sub-region in which x_0 falls

Algorithm: growing a tree

1. All observations in a single set
2. Sort values on first variable
3. Compute split criteria for all possible splits into two sets
4. Choose the best split on this variable
5. Repeat 2-4 for all other variables
6. Choose the best split among all variables. Your data is now in two sets.
7. Repeat 1-6 on each subset.
8. Stop when stopping rule is achieved.

Split criteria - purity/impurity metrics

For K classes, and subset m

- The **Gini index** measures is defined as:

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

- **Entropy** is defined as

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

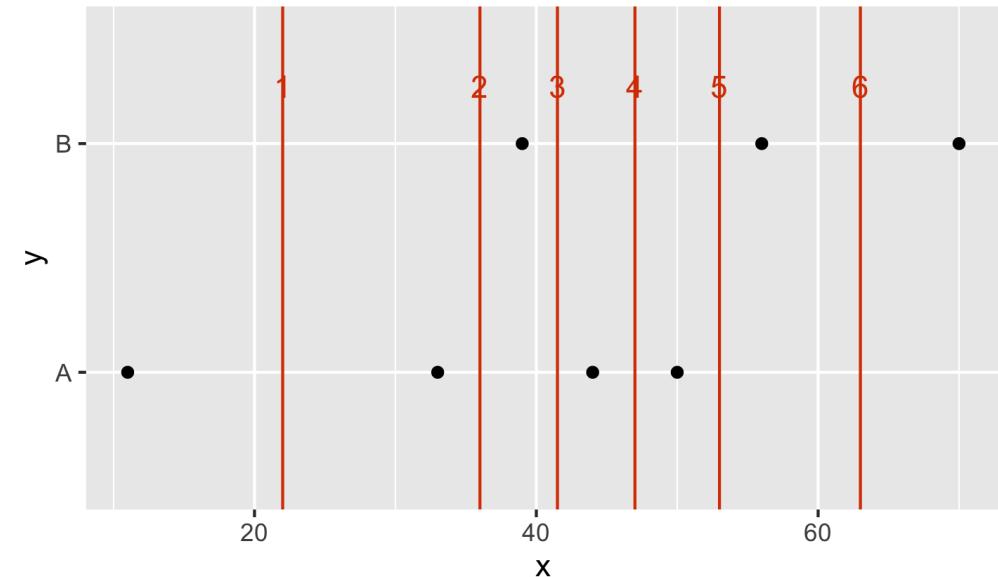
- If all \hat{p}_{mk} close to zero or one, **D** are small. **Lower is better!**

Stopping rules

- **Minimum split**: number of observations in a node, in order for a split to be made
- **Minimum bucket**: Minimum number of observations allowed in a terminal node
- **Complexity parameter**: minimum difference between impurity values required to continue splitting

Illustration for one variable

x	y
11	A
33	A
39	B
44	A
50	A
56	B
70	B



What do you think is the best split? 2, 3 or 5??

Note that x is sorted from lowest to highest!

Calculate the impurity for a split

Look at split 5.

The **left** bucket is

x	y
11	A
33	A
39	B
44	A
50	A

and the **right** bucket is

x	y
56	B
70	B

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

Left bucket:

$$\hat{p}_{LA} = 4/5, \hat{p}_{LB} = 1/5, p_L$$

$$G_L = 0.8(1 - 0.8) + 0.2(1 - 0.2)$$

Right bucket:

$$\hat{p}_{RA} = 0/2, \hat{p}_{RB} = 2/2, p_R$$

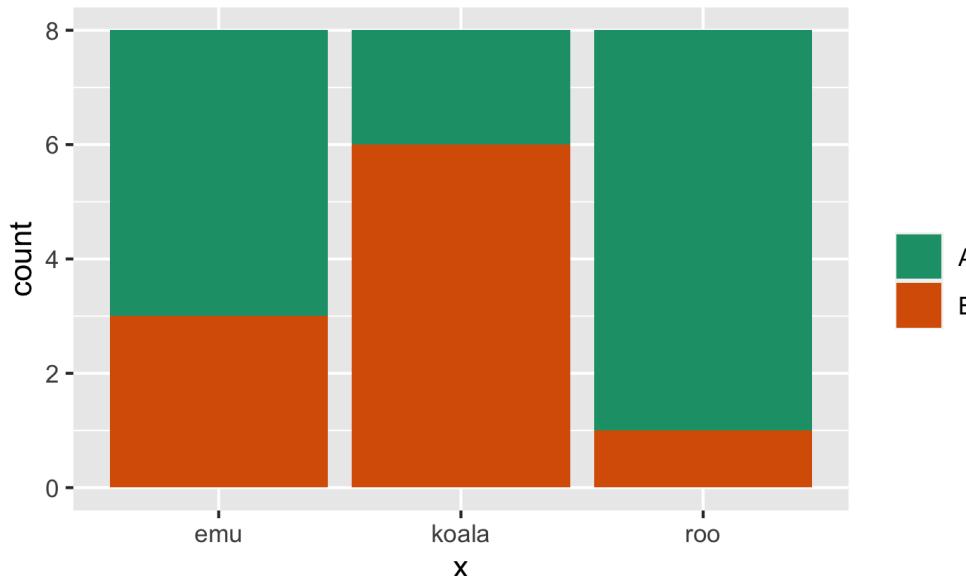
$$G_R = 0(1 - 0) + 1(1 - 1)$$

Combine with weighted sum to get **impurity for the split**:

$$5/7G_L + 2/7G_R = 0.32$$

Your turn: compute the impurity for split 2.

Splits on categorical variables



Split would be "if koala then assign to B else assign to A"

Handling missing values

x1	x2	x3	x4	y
19	-8	22	-24	A
NA	-10	26	-26	A
15	NA	32	-27	B
17	-6	27	-25	A
18	-5	NA	-23	A
13	-3	37	NA	B
12	-1	35	-30	B
11	-7	24	-31	B

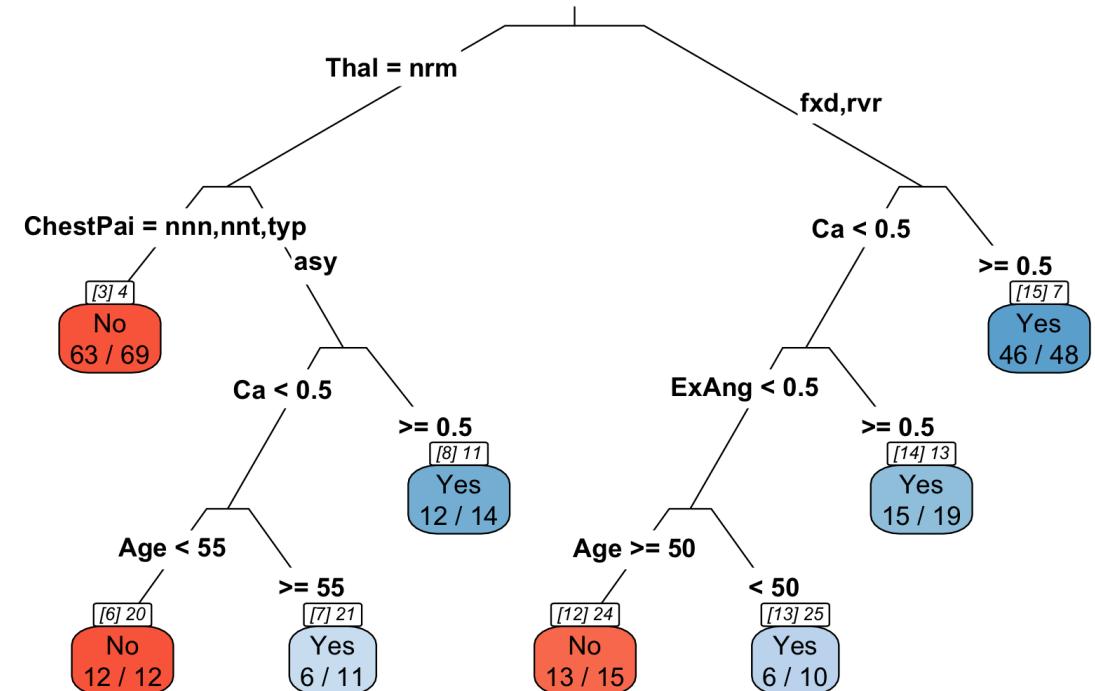
50% of cases have missing values, which causes most methods to falter. For trees missings only on a single variable are ignored.

Example - predicting heart disease

:YAHD, presence of heart disease (Yes/No)

:Xheart and lung function measurements

```
## [1] "Age"          "Sex"  
## [3] "ChestPain"   "RestBP"  
## [5] "Chol"         "Fbs"  
## [7] "RestECG"     "MaxHR"  
## [9] "ExAng"        "Oldpeak"  
## [11] "Slope"        "Ca"  
## [13] "Thal"         "AHD"
```

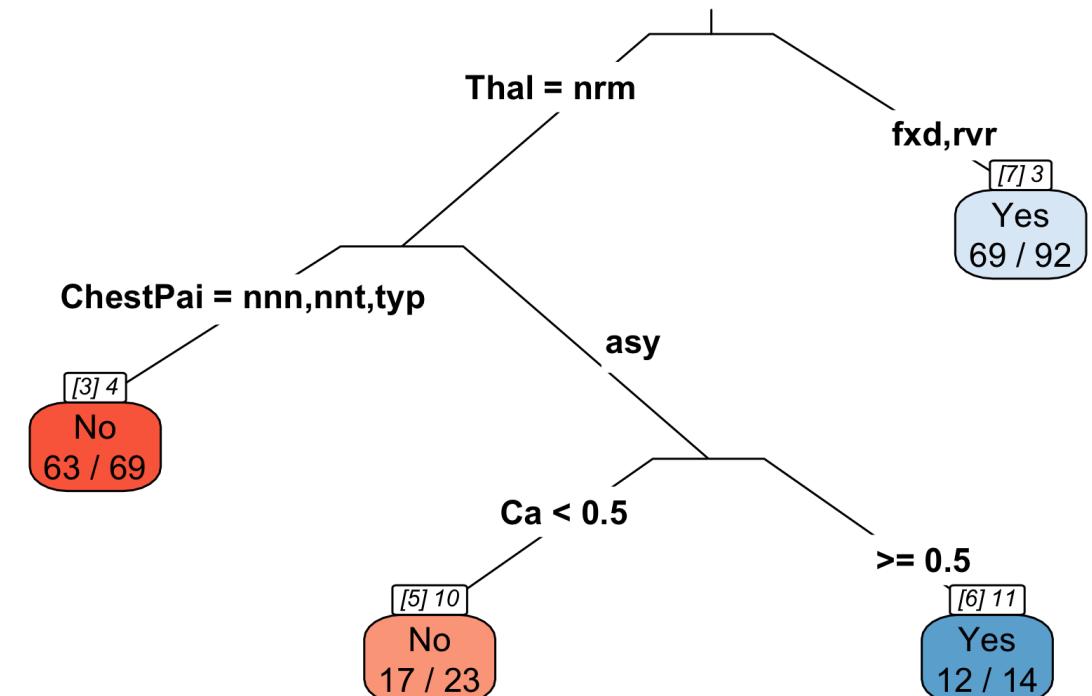


Deeper trees

Trees can be built deeper by:

- decreasing the value of the complexity parameter `cp`, which sets the difference between impurity values required to continue splitting.
- reducing the `minsplit` and `minbucket` parameters, which control the number of observations below splits are forbidden.

Larger complexity, simpler tree



Tabulate true vs predicted to make a **confusion table**.

		true	
		C1 (positive)	C2 (negative)
pred- icted	C1	a	b
	C2	c	d

- ➊ Accuracy: $(a+d)/(a+b+c+d)$
- ➋ Error: $(b+c)/(a+b+c+d)$
- ➌ Sensitivity: $a/(a+c)$ (true positive, recall)
- ➍ Specificity: $d/(b+d)$ (true negative)
- ➎ Balanced accuracy: $(\text{sensitivity}+\text{specificity})/2$

Training confusion and error

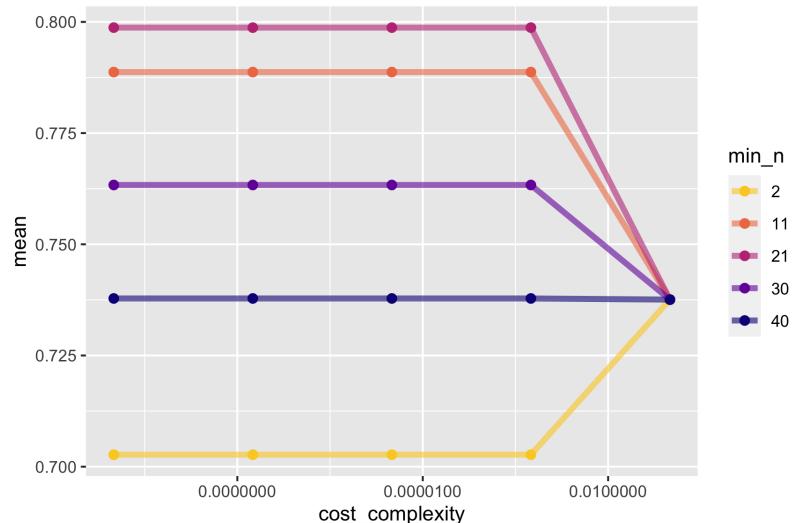
```
##          Truth  
## Prediction No Yes  
##      No  95  17  
##      Yes 11  75
```

Test confusion and error

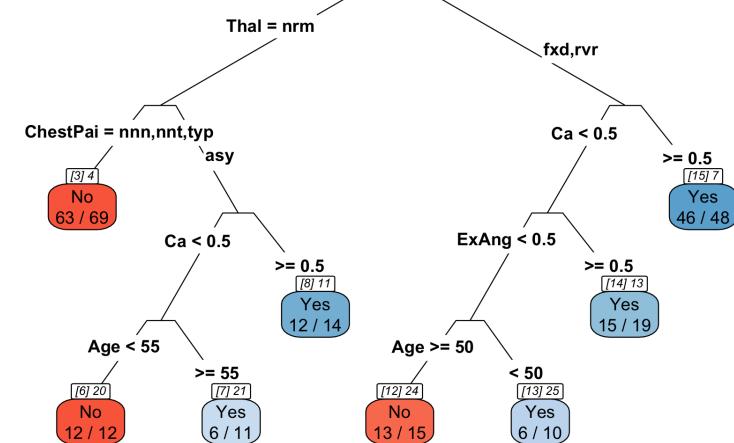
```
##          Truth  
## Prediction No Yes  
##      No  40  10  
##      Yes 14  35
```

Training vs testing performance

- Cross-validation, 5-fold
- Grid of values in complexity, and min split



```
## # A tibble: 1 × 3
##   cost_complexity min_n
##   <dbl> <int>
## 1       0.1      2
```



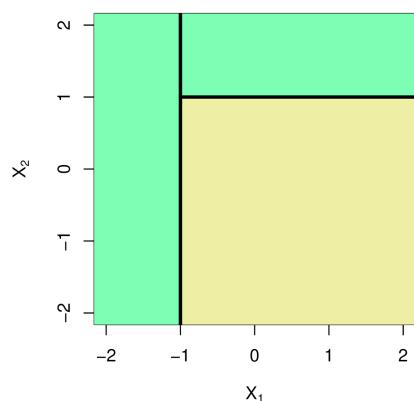
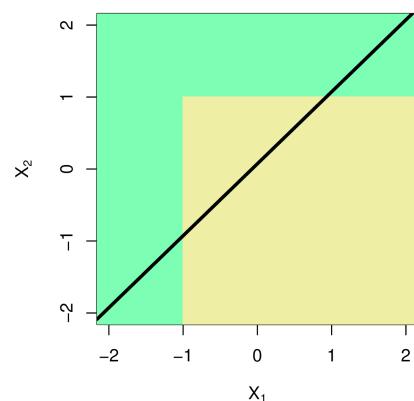
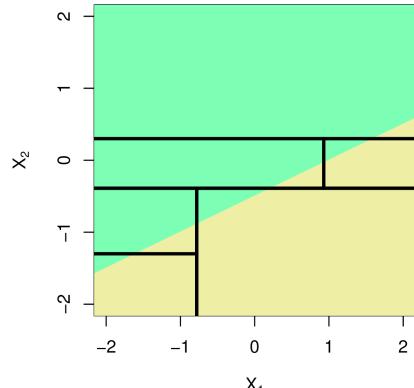
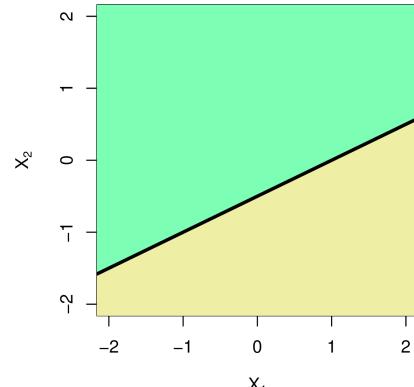
Test confusion matrix

		Truth	
		Prediction	No Yes
		No	41 12
##	##	Yes	13 33

Comparison with LDA

Look at the following classification problems and resultant decision boundaries for LDA (left) and CART (right).

What characteristics determine which method is more appropriate?

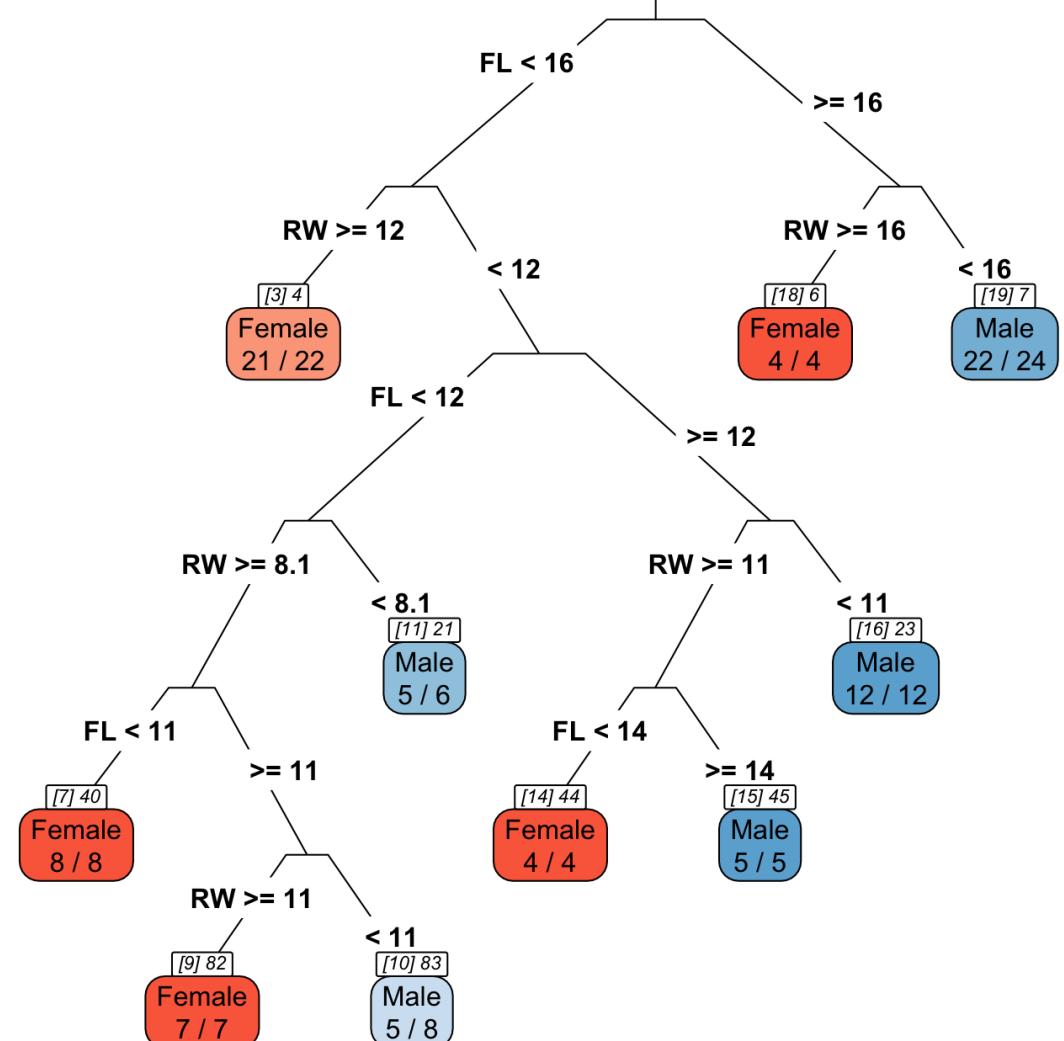


Example - Crabs

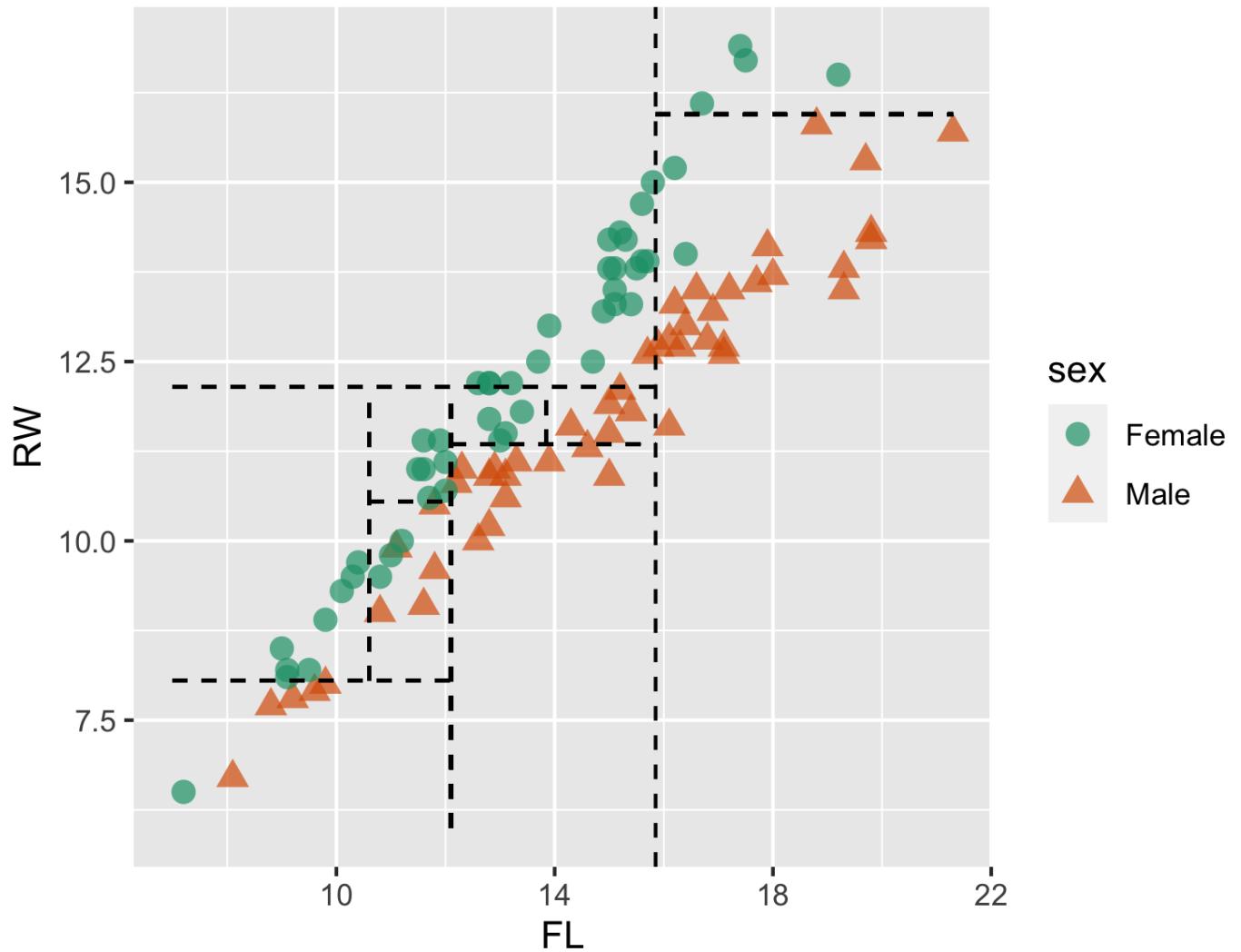
Physical measurements on WA crabs, males and females.

Data source: Campbell, N. A. & Mahon, R. J. (1974)

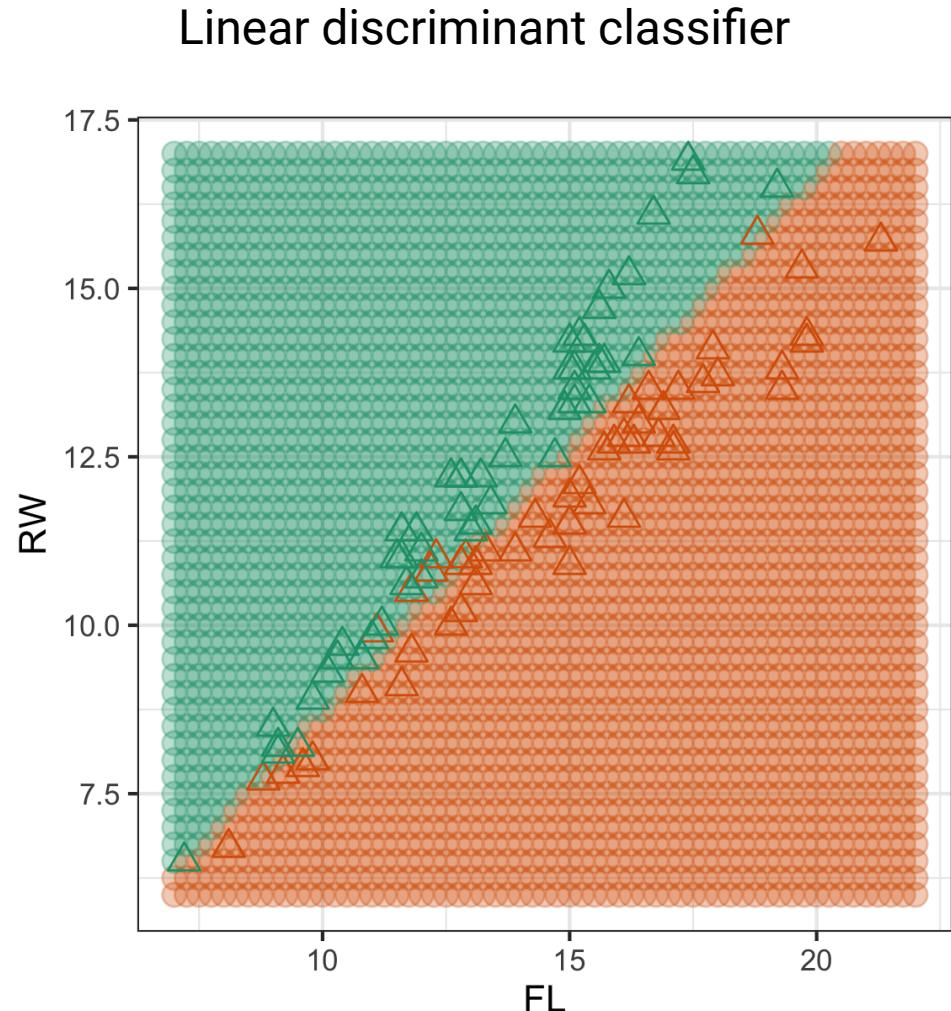
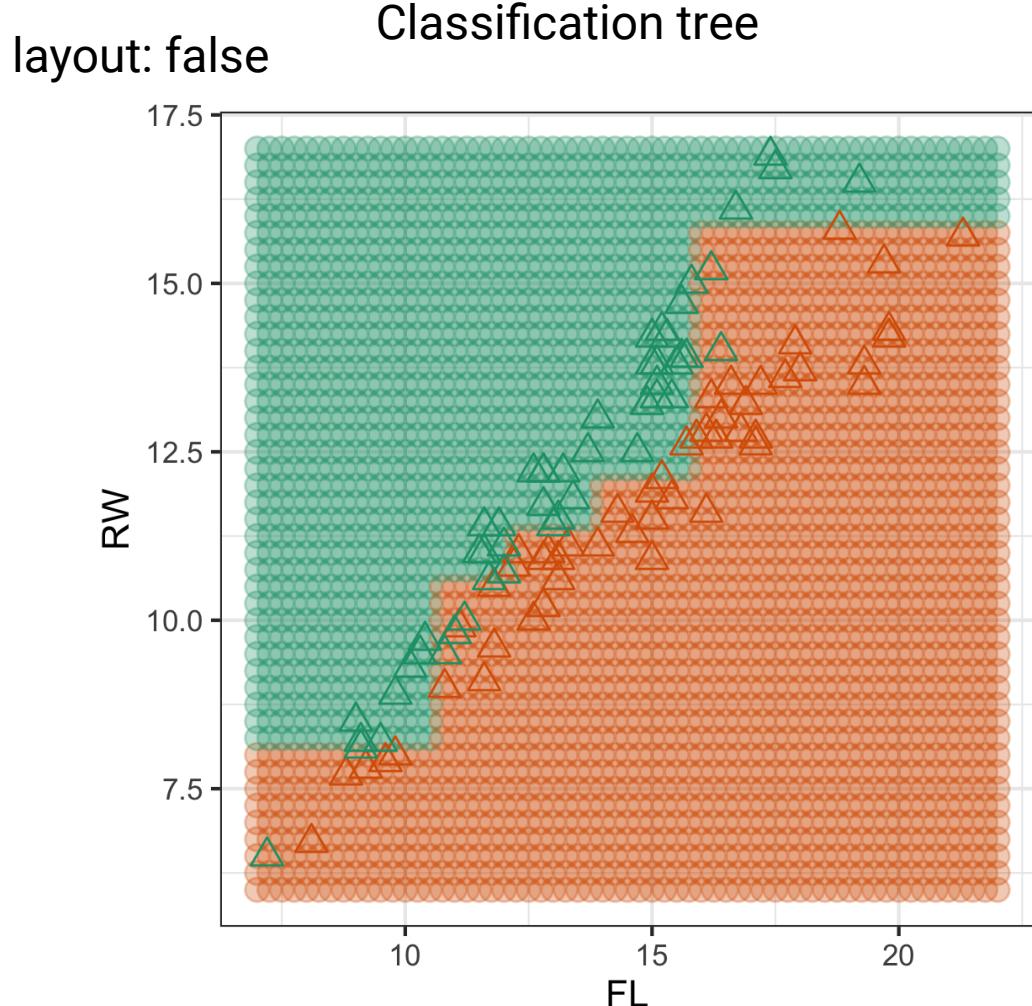
Decision tree parameters: minsplit=9. It's been forced to fit small subsets.



Example - Crabs



Boundaries induced by different models



Pros and cons

- The decision rules provided by trees are very easy to explain, and follow. A simple classification model.
- Trees can handle a mix of predictor types, categorical and quantitative.
- Trees efficiently operate when there are missing values in the predictors.
- Algorithm is greedy, a better final solution might be obtained by taking a second best split earlier.
- When separation is in linear combinations of variables trees struggle to provide a good classification



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: Professor Di Cook

Department of Econometrics and Business Statistics

✉ ETC3250.Clayton-x@monash.edu

CALENDAR
Week 6a

