ETC3250/5250: Classification Trees

Semester 1, 2020

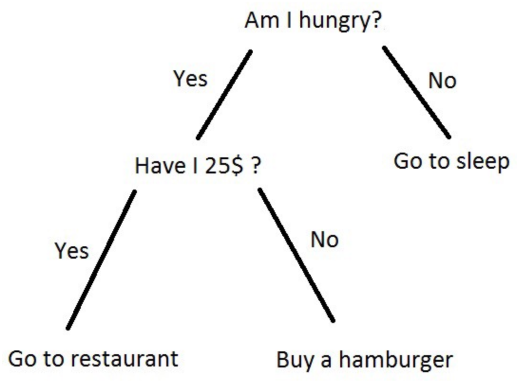
Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 6 (a)

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

In class exercise!

Everyone in the class line up from tallest to shortest.

Sorting algorithms

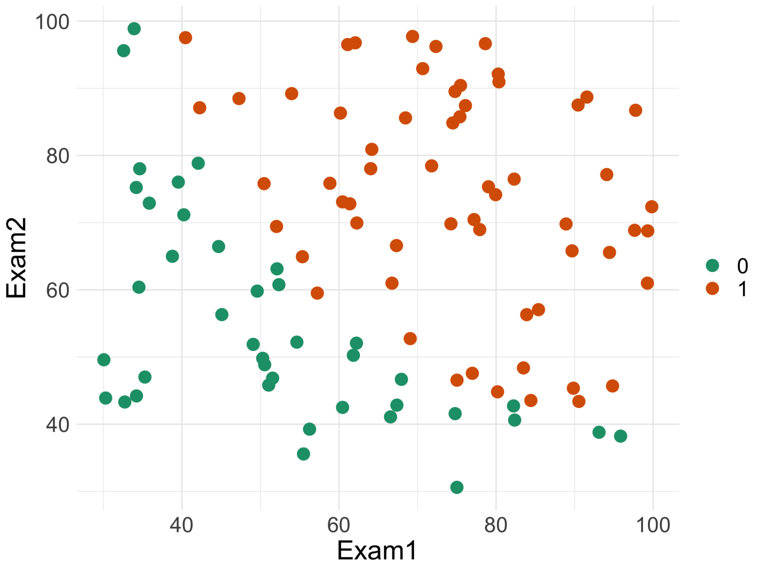
There are numerous sorting algorithms

The "speed" of classification trees depends on how quickly one can sort.
Source

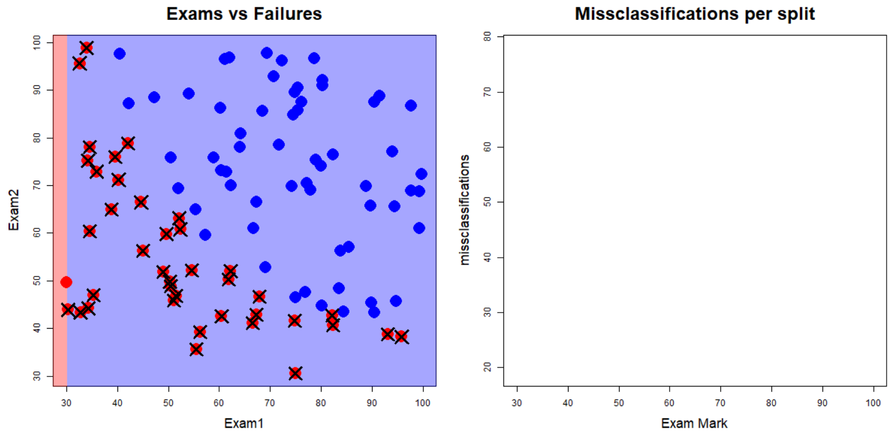
What about two dimensions ?

Consider the dataset `Exam` where two exam scores are given for each student, and a class `Label` represents whether they passed or failed the course.

```
##      Exam1    Exam2 Label
## 1 34.62366 78.02469     0
## 2 30.28671 43.89500     0
## 3 35.84741 72.90220     0
## 4 60.18260 86.30855     1
```



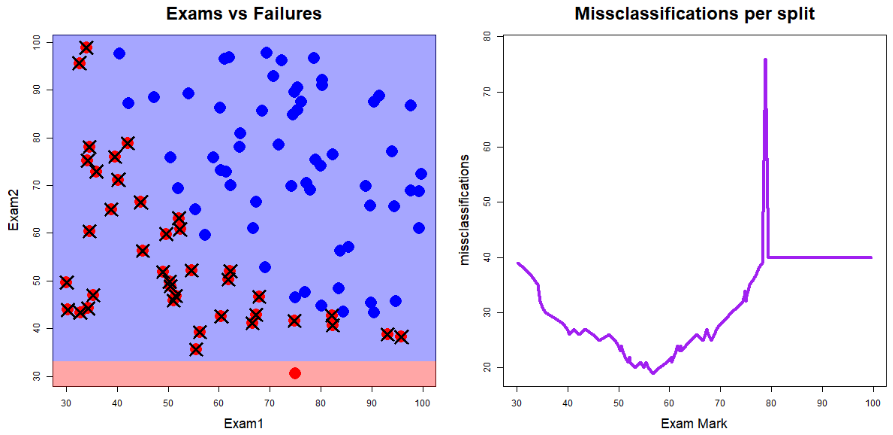
Calculate the number of misclassifications along all splits for `Exam1` classifying according to the majority class for the left and right splits



Red dots are "fails", blue dots are "passes", and crosses indicate misclassifications.

Source: John Ormerod, U.Syd

Calculate the number of misclassifications along all splits for `Exam2` classifying according to the majority class for the top and bottom splits



Red dots are "fails", blue dots are "passes", and crosses indicate misclassifications.

Source: John Ormerod, U.Syd

Combining the results from `Exam1` and `Exam2` splits

- The minimum number of misclassifications from using all possible splits of `Exam1` was 19 when the value of `Exam1` was **56.7**
- The minimum number of misclassifications from using all possible splits of `Exam2` was 23 when the value of `Exam2` was **52.5**

So we split on the best of these, i.e., split the data on `Exam1` at 56.7.

Split criteria - purity/impurity metrics

- The Gini index measures total variance across the K classes:

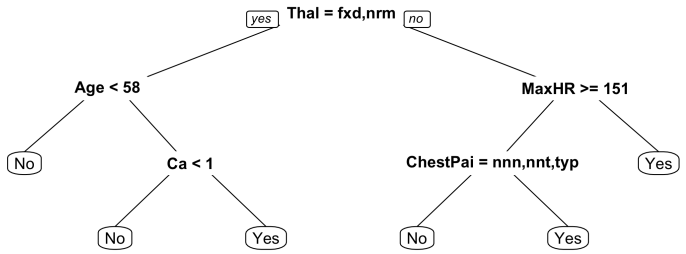
$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- Entropy is defined as

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

- If all $\hat{p}_{mk}$'s close to zero or one, G and D are small.

Example - predicting heart disease



Y: presence of heart disease (Yes/No)

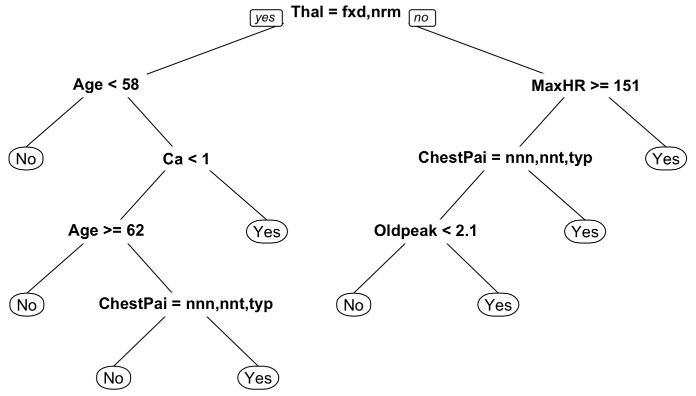
X: heart and lung function measurements

```
##  [1] "Age"       "Sex"       "ChestPain" "RestBP"    "Chol"      "Fbs"
##  [7] "RestECG"   "MaxHR"     "ExAng"     "Oldpeak"   "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.



Tabulate true vs predicted to make a confusion table.

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

- Accuracy: *(a+d)/(a+b+c+d)*
- Error: *(b+c)/(a+b+c+d)*
- Sensitivity: *a/(a+c)* (true positive, recall)
- Specificity: *d/(b+d)* (true negative)
- Balanced accuracy: *(sensitivity+specificity)/2*

Training confusion and error

```
##            Reference
## Prediction No Yes
##        No  75   5
##        Yes 11  58
```

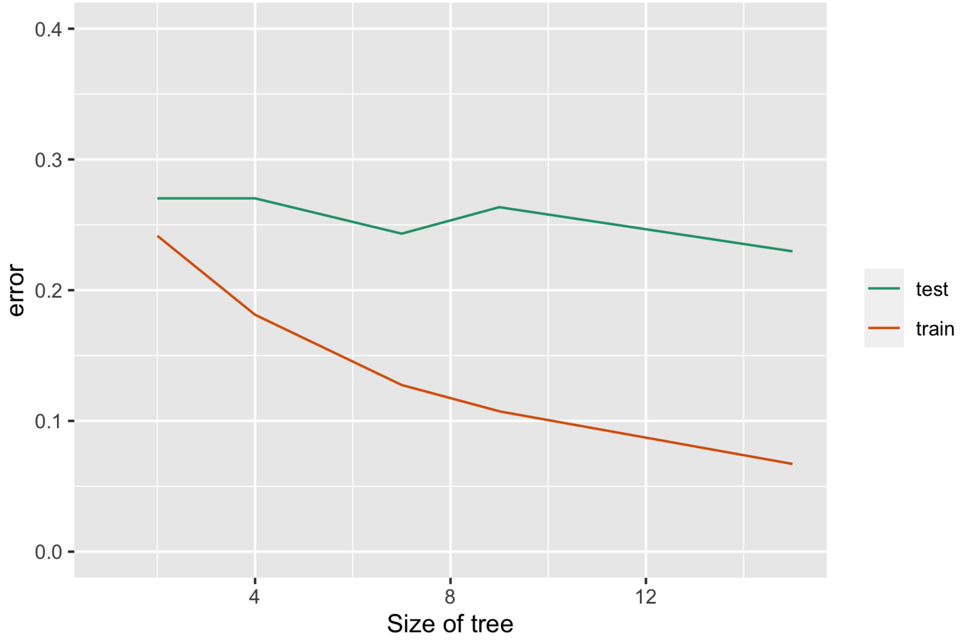
```
##  Accuracy
## 0.8926174
```

Test confusion and error

```
##            Reference
## Prediction No Yes
##        No  59  21
##        Yes 18  50
```

```
##  Accuracy
## 0.7364865
```

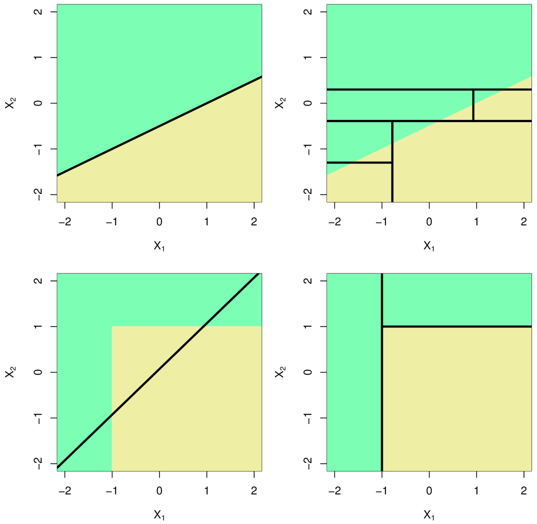
Training vs testing performance



Comparison with LDA

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

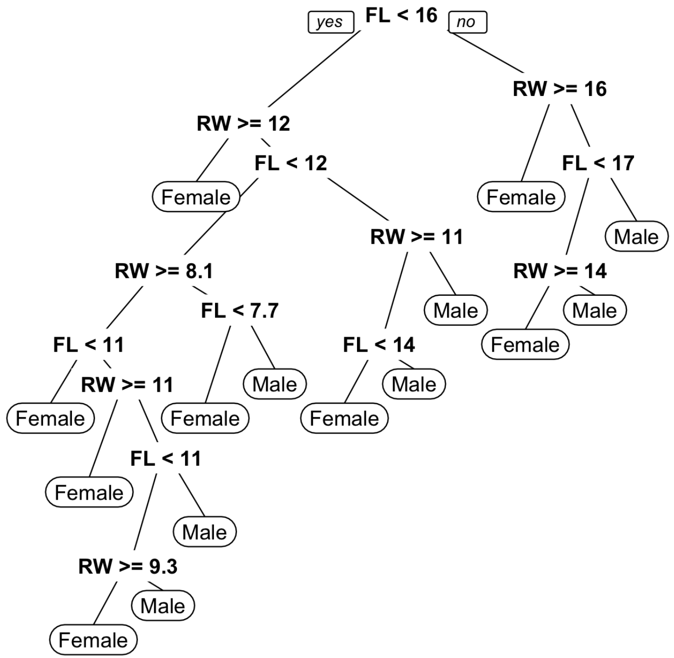
What characteristics determine which method is more appropriate?



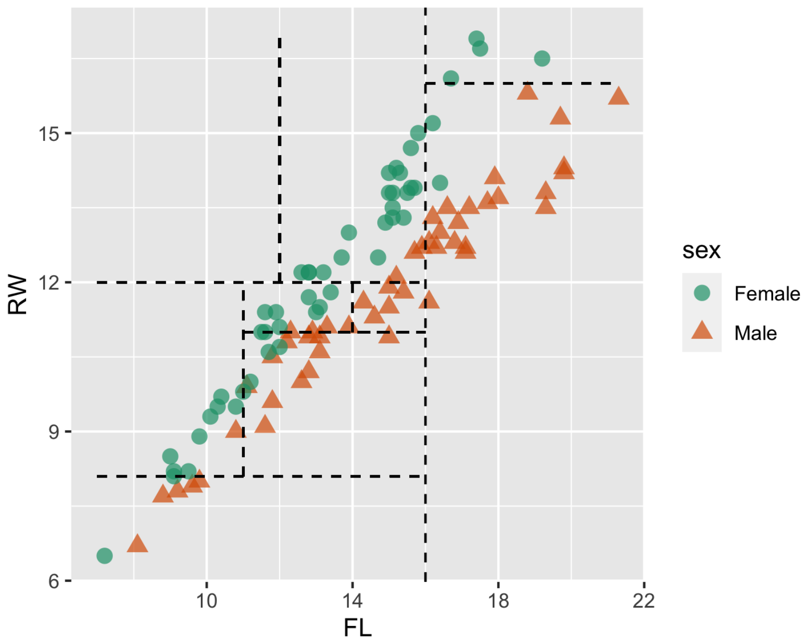
Example - Crabs

Physical measurements on WA crabs, males and females.

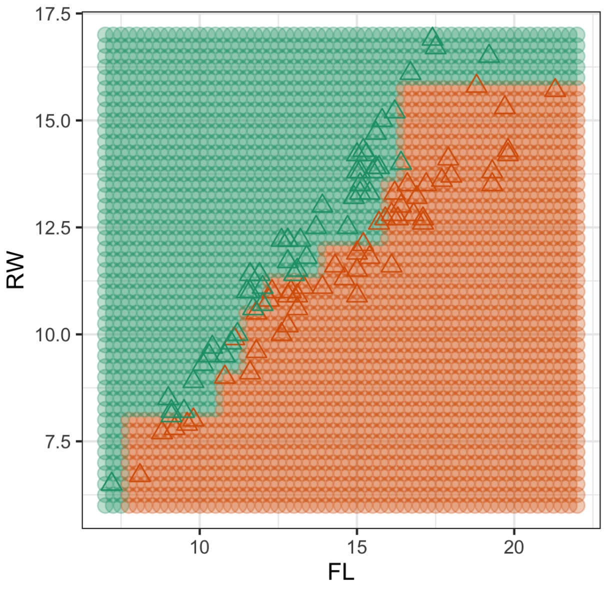
Data source: Campbell, N. A. & Mahon, R. J. (1974)



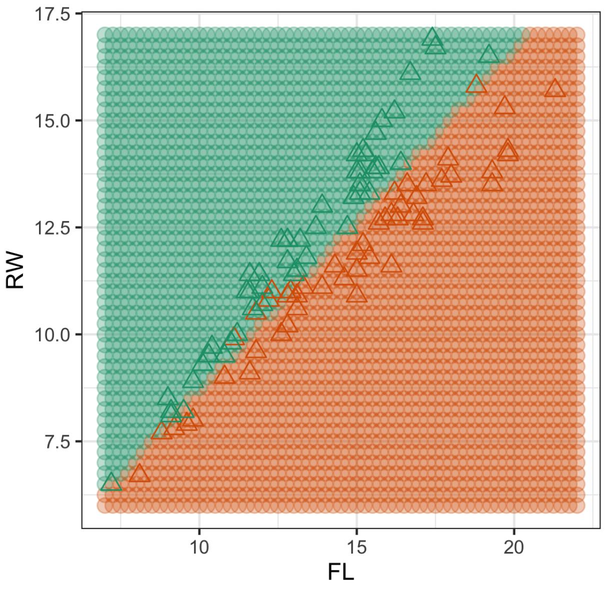
Example - Crabs



Classification tree



Linear discriminant classifier



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

👩‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.