FIT1043 Introduction to Data Science

Week 7: Classification

Ts. Dr. Sicily Ting

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Week 6 Coverage

Model Fitting

Bias & Variance

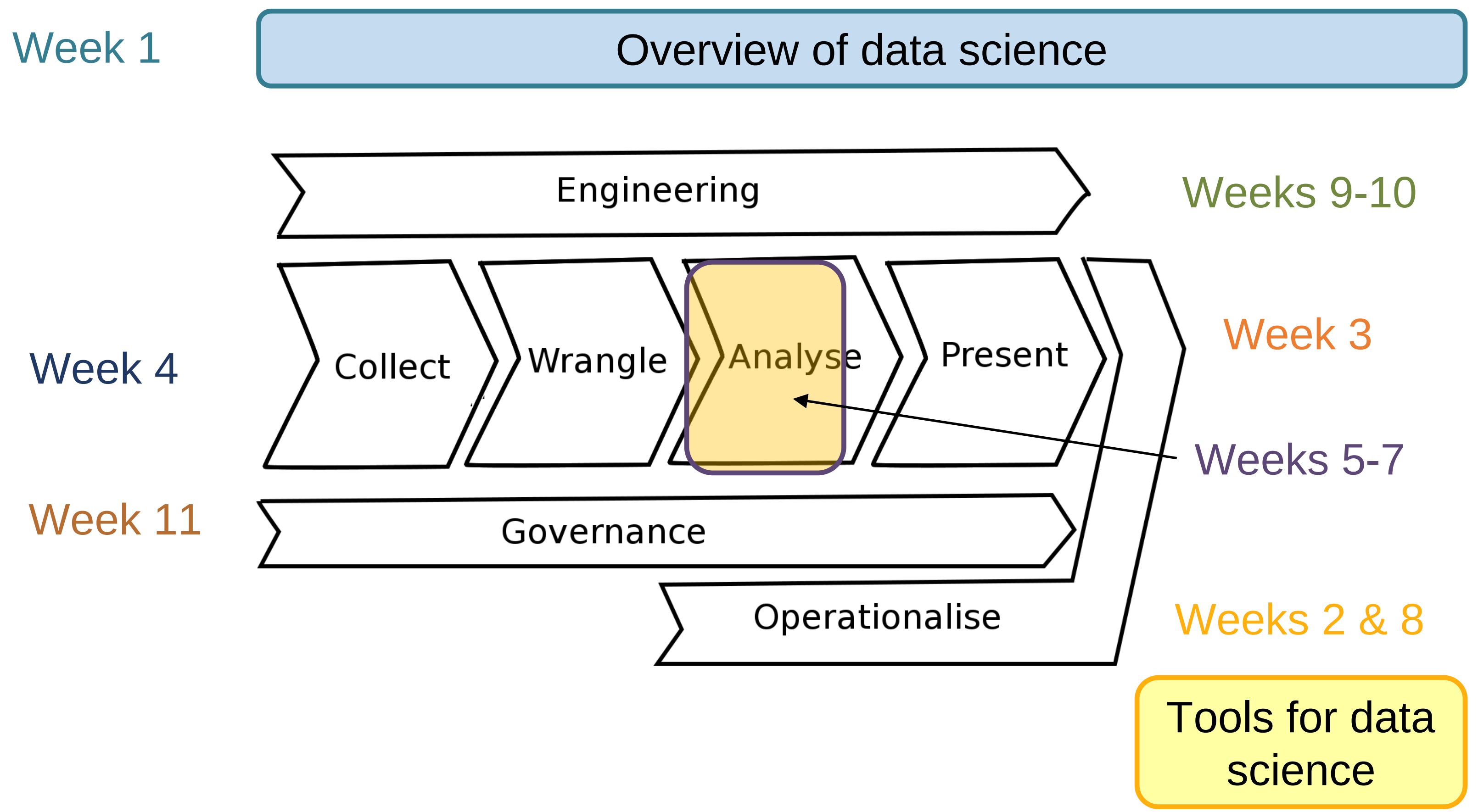
Ensembles

MONASH University

Introduction to Data Analysis

Week 6 Outline

- Linear regression terminology
- How to calculate model parameters
- Underfitting vs Overfitting
- Bias and Variance
  - No free lunch theorem
- Ensemble models
- Descriptive vs Predictive Data Analysis


Week 1
Overview of data science
Engineering
Weeks 9-10
Week 4
Collect
Wrangle
Analyse
Present
Week 3
Weeks 5-7
Week 11
Governance
Operationalise
Weeks 2 & 8
Tools for data science

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	Assignment 2
9	Characterising data and "big" data	
10	Big data processing	
11	Issues in data management	Assignment 3
12	Industry guest lecture (tentative)	

MONASH University

Classification and Clustering

Week 7 Outline

- Classification
  - How to evaluate
  - Classification metrics
  - Decision trees
- Regression
  - Regression trees
- Ensemble learning
  - Random forest
- Clustering
  - *k*-means

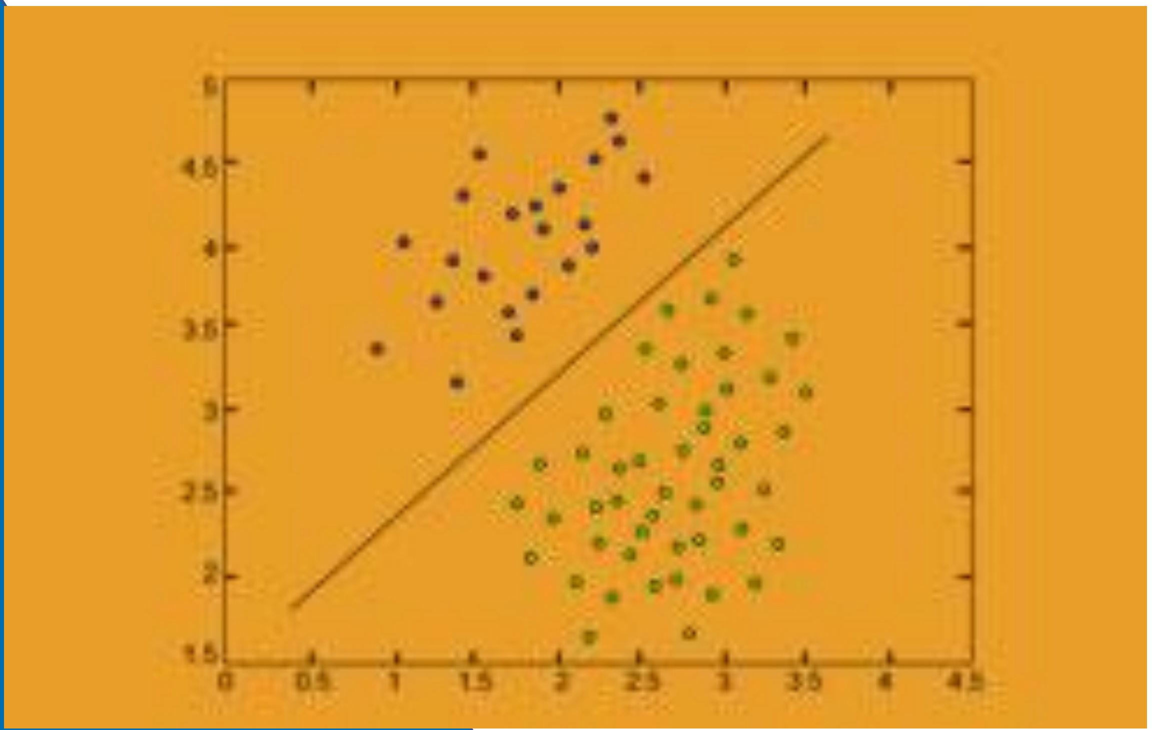
Learning Outcomes

Week 7

By the end of this week you should be able to:

- Differentiate between classification and regression models
- Explain how decision trees and regression trees work
- Explain how random forest works
- Explain how *k*-means clustering works
- Analyse confusion matrix and how to calculate prediction accuracy
- Differentiate between different classification metrics

Data Analysis Classification Algorithms



Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

MONASH University

Classification

Cat **Dog**

?

If we are to ask the computer whether this image provided is a cat or a dog.
What would the computer need?
What would an infant need?

Classification

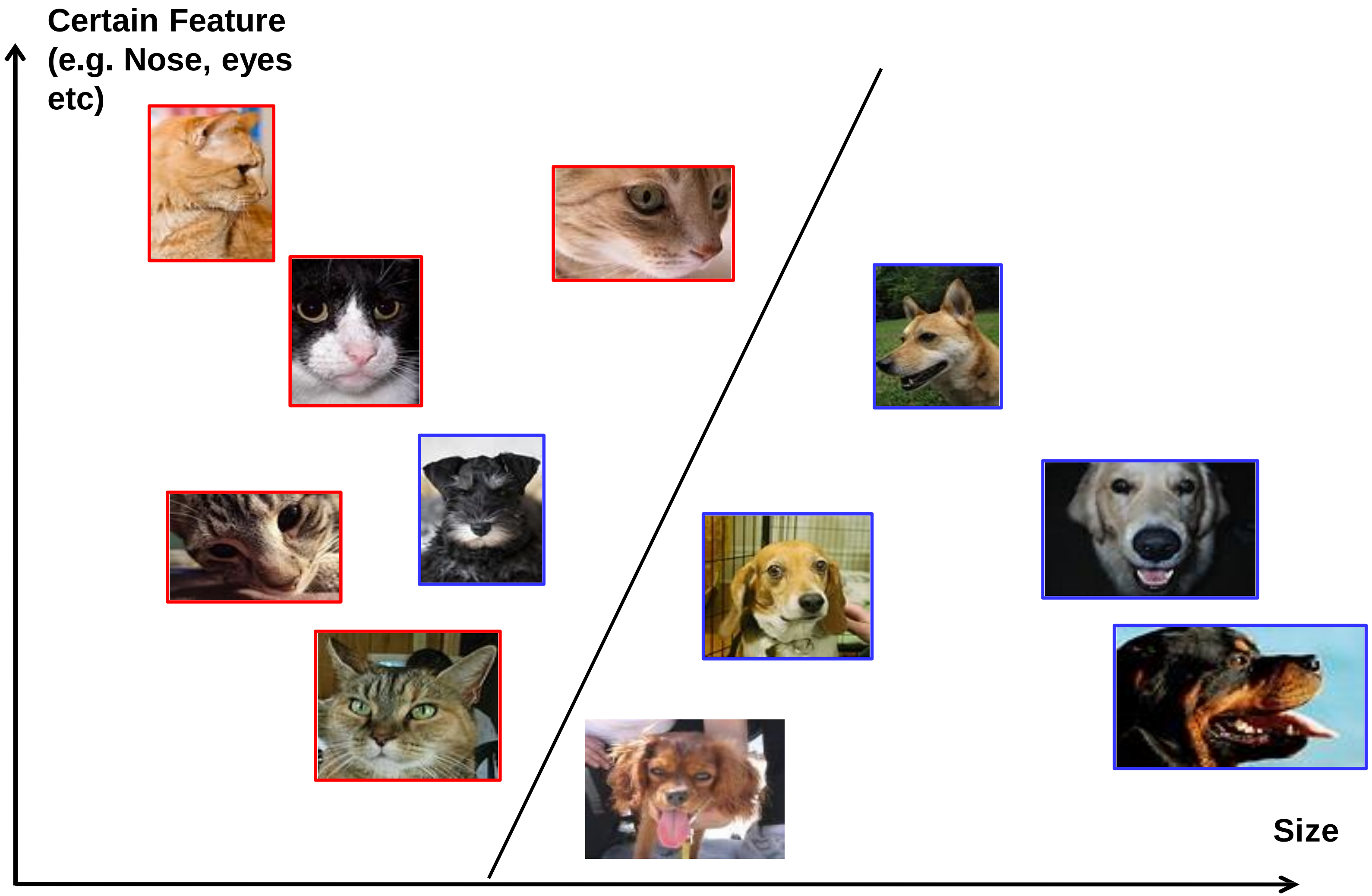
Cat

Dog

?

First thing is that the computer will need to learn about cats and dogs from the past (based on data).

Classification



- if you look closely at what I have put up there, there is a dog that may be segmented and identified as a cat.

Classification

Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

Classification

- **Question**: Can we predict the diabetes status of a patient given their health measurements (i.e., 'pregnant', 'insulin', 'BMI', 'age')?

    - *Dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database*

- *Do go through the following resources and as additional materials and focus on the confusion matrix. It is quite comprehensive but we will be focusing on confusion matrix.*

    - *Code: https://github.com/justmarkham/scikit-learn-videos/blob/master/09_classification_metrics.ipynb*
    - *Video (55 mins on evaluating a classification model): https://www.youtube.com/watch?v=85dtiMz9tSo&list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A&index=9*

MONASH University

How do we evaluate the prediction accuracy?

- Percentage of correct predictions by comparing the **actual** with the **predicted** response values

- The simplest is just to determine how many are correctly predicted.
  - What’s correct?

MONASH University

Confusion Matrix

A tool to measure performance for classification

		Predicted Values	
		Positive(1)	Negative(0)
Actual Values	Positive(1)	True Positive (TP)	False Negative (FN)
	Negative(0)	False Positive (FP)	True Negative (TN)

Diabetes detection

True Positives (TP): we correctly predicted that they do have diabetes
True Negatives (TN): we correctly predicted that they don't have diabetes
False Positives (FP): we incorrectly predicted that they do have diabetes
False Negatives (FN): we incorrectly predicted that they don't have diabetes

Confusion Matrix

Predicted Class

		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

From: “[What is Confusion Matrix and Advanced Classification Metrics?”]

Classification Metrics

- **Accuracy**: Overall, how often is the prediction correct?
- **Sensitivity** (Recall): When the actual value is positive, how often is the prediction correct?
- **Specificity:** When the actual value is negative, how often is the prediction correct?
- **False Positive Rate:** When the actual value is negative, how often is the prediction incorrect?
- **Precision:** When a positive value is predicted, how often is the prediction correct?

Why is this important?

You should be able to understand the following …

PUTRAJAYA (Bernama): Malaysia will receive a new antigen rapid test kit from South Korea on Monday (April 6), says Health director-general Datuk Dr Noor Hisham Abdullah.

He said once received, it would be verified by the Ministry of Health (MOH) on its accuracy and sensitivity.

"We are hoping that this new kit's sensitivity, specificity and accuracy rate will be over 80% so that we can improve the number of tests conducted daily to 16,500 from 11,500 tests currently," he told a daily news conference here Sunday (April 5).

MONASH University

Which Metrics Should be Used?

It depends …

Spam filter: Optimise **precision** or **specificity**

- False negatives (spam goes to the inbox) are more acceptable than false positives (non-spam is caught by the spam filter)

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Which Metrics Should be Used?

It depends …

Fraudulent transaction detector: Optimise **sensitivity**

- False positives (normal transactions that are flagged as possible fraud) are more acceptable than false negatives (fraudulent transactions that are not detected)

Predicted Class

		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Actual Class	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Which Metrics Should be Used?

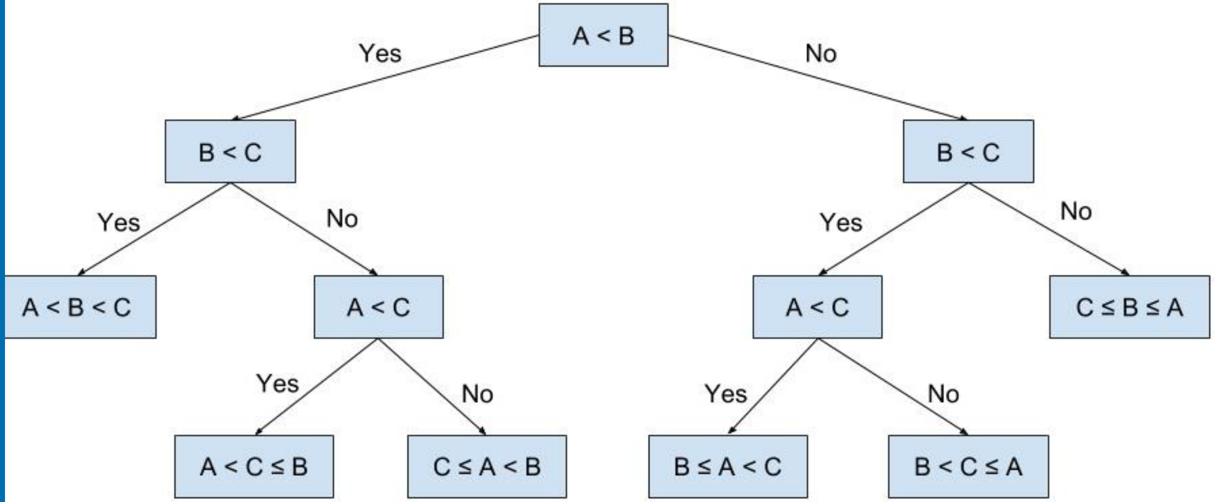
It depends …

Examples we covered

Please watch the lecture video on this for examples to help you understand why the different measures are important in different situations.

- As a Covid test kit effectiveness measure
- As a spam filter measure
- Fraudulent transaction filter

Decision Tree and Regression Tree Algorithms



MONASH University

Decision Trees and Regression Trees

What is Decision Trees?

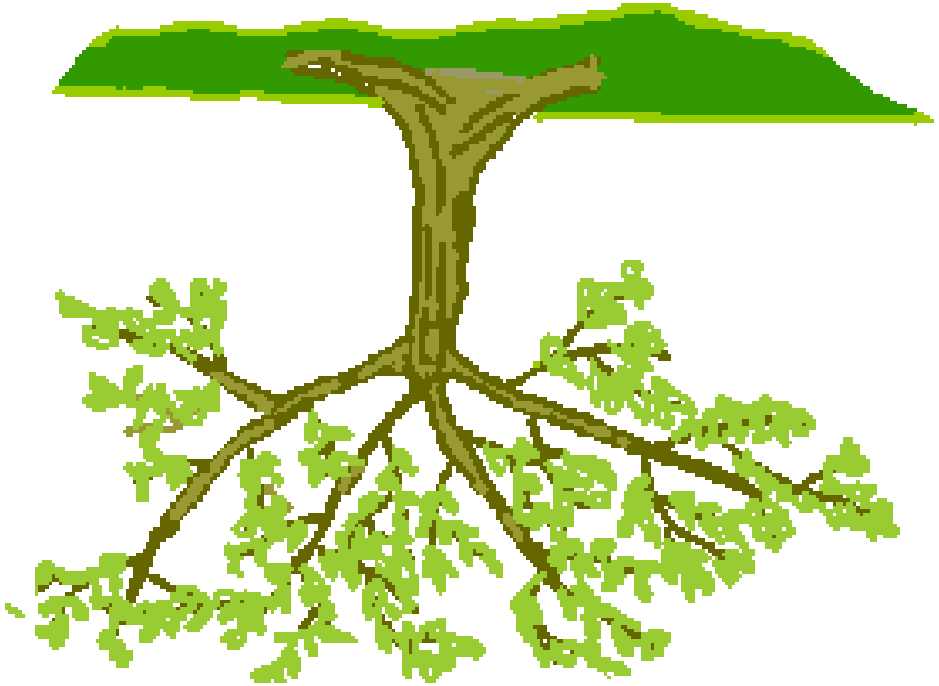
- Predict binary or multi-class (categorical) outcomes

What is Regression Trees?

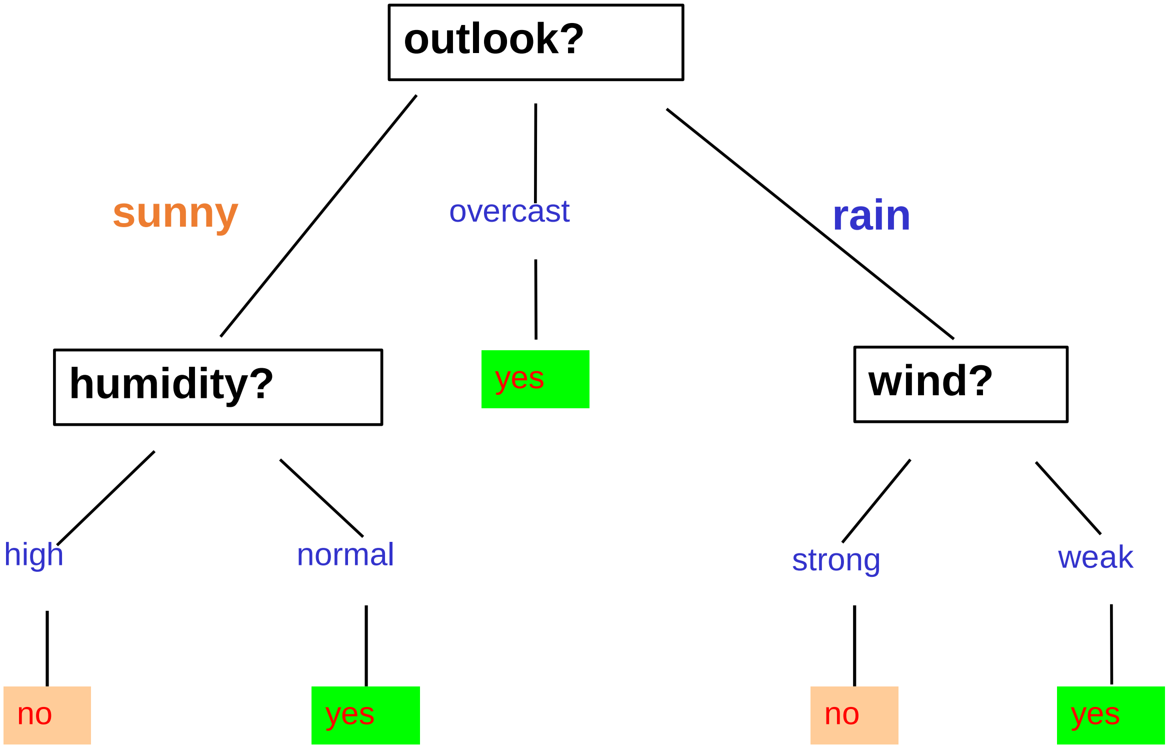
- Predict continuous (i.e. real) values

What are Trees in Computing?

We can start at the root and at we can decide which branch to take.

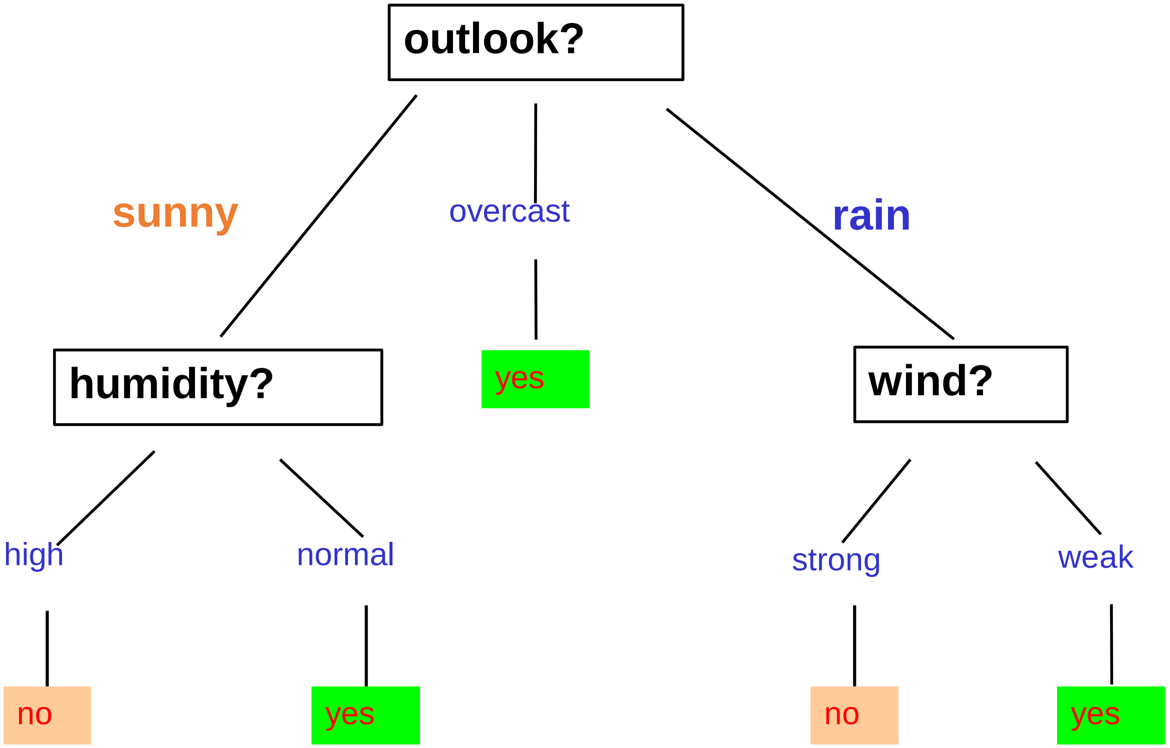


Decision Tree Example



What do you think this decision tree is about?

Decision Tree Example

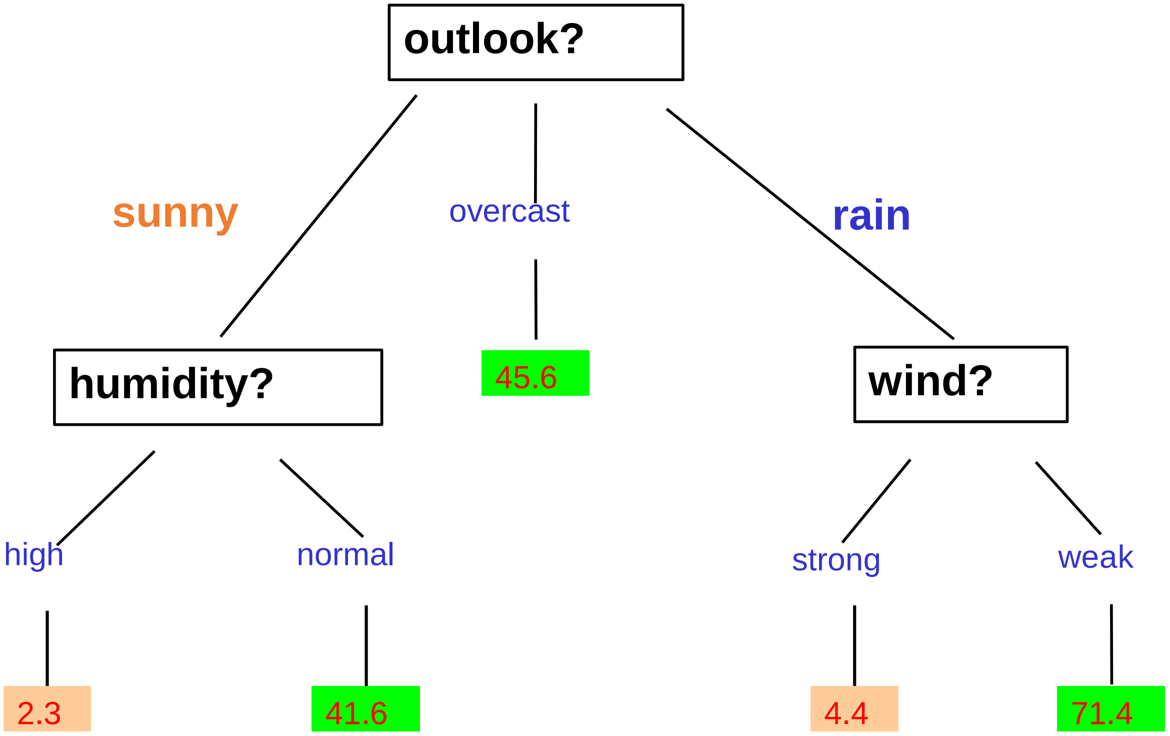


Set of rules:

Good day to play tennis ⇔ (Sunny and Normal) or Overcast or (Rain and Weak)

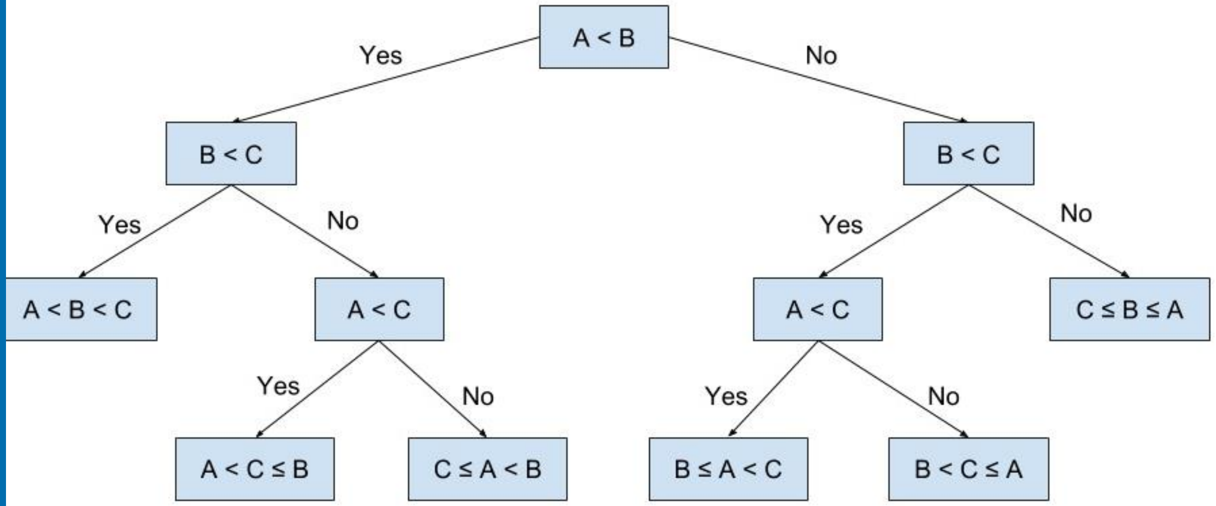
Bad day to play tennis ⇔ ?

Regression Tree Example



What if we are looking at “satisfaction” level in terms of playing tennis instead of a binary, yes or no?

Building a Regression or Decision Tree

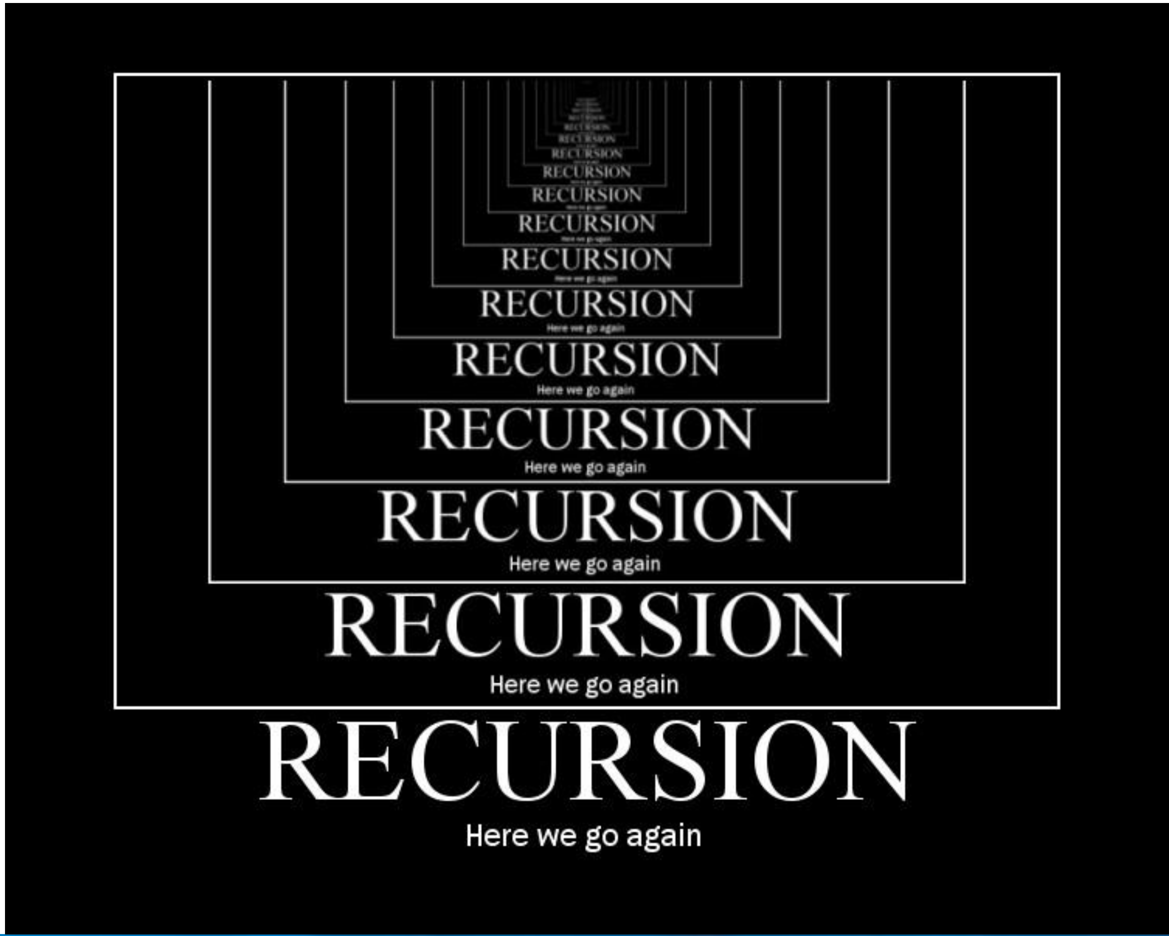


MONASH University

How to Build Regression and Decision Trees?

Recursively partition (divide up) the feature space into regions

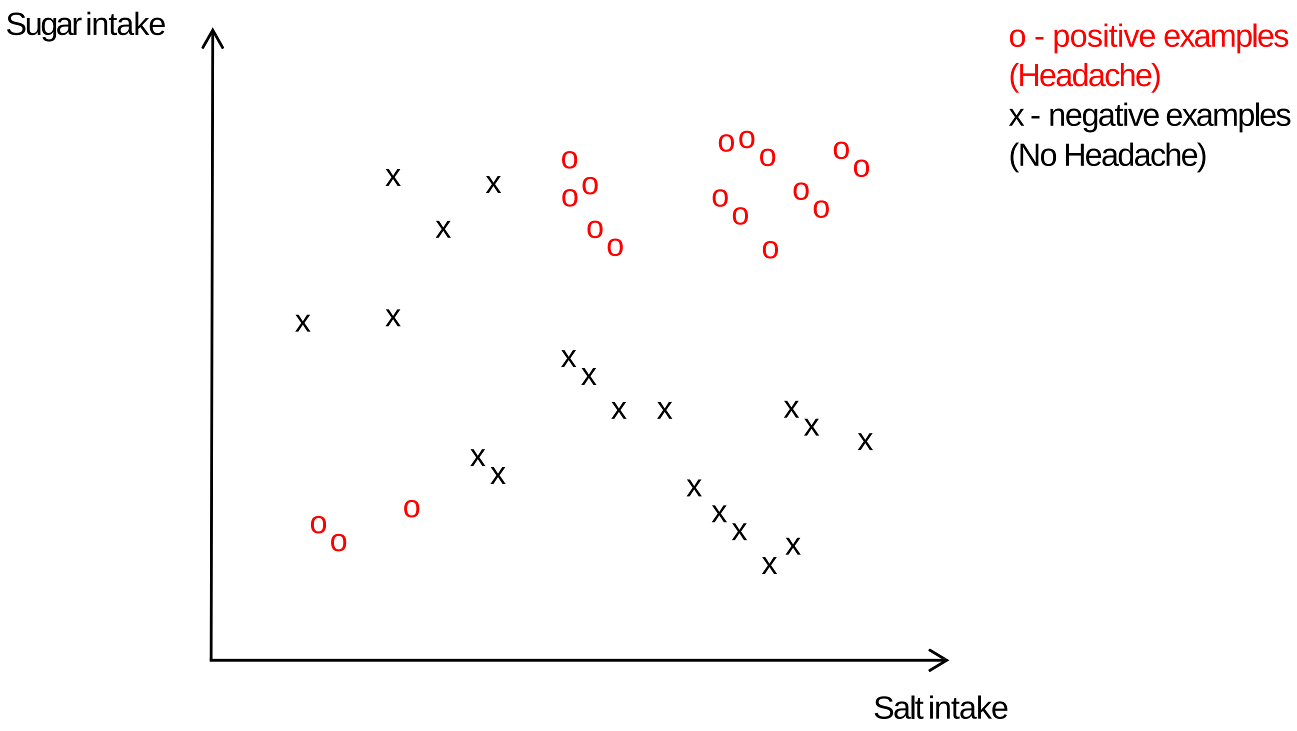
While grouping similar instances together



MONASH University

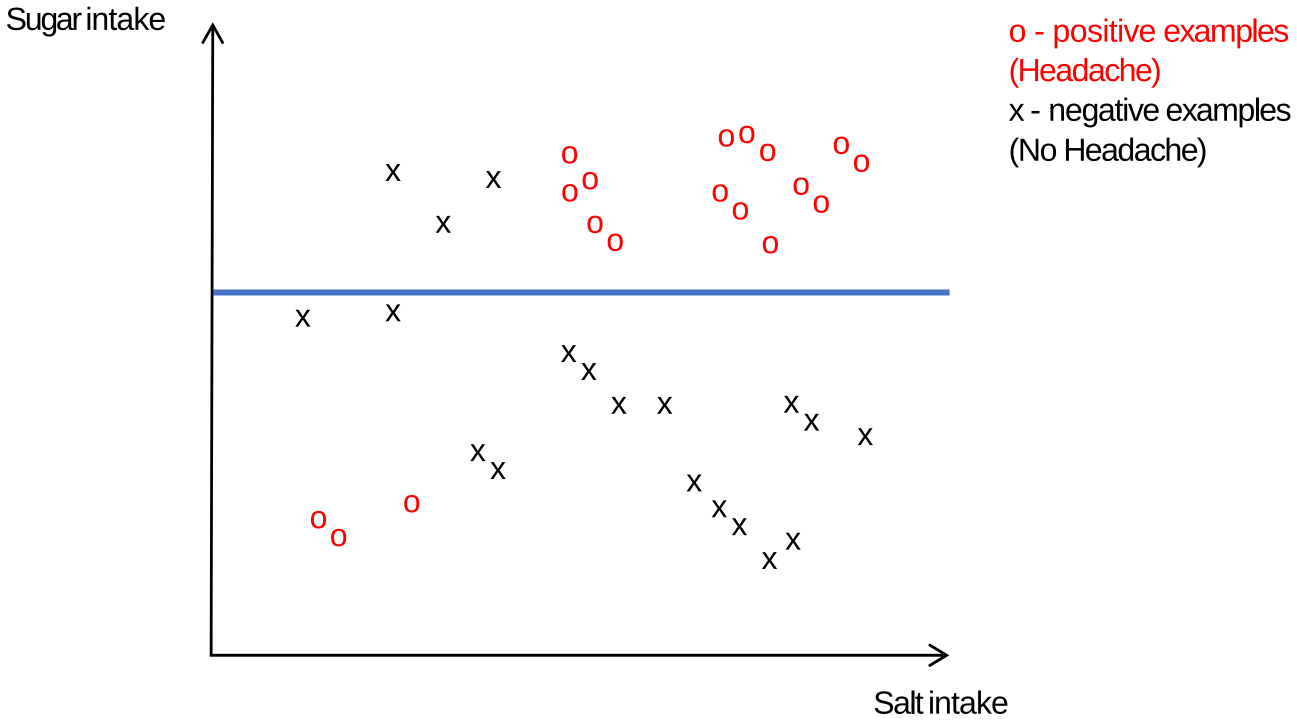
Recursive Partitioning

At each iteration, we divide the data to group similar instances together



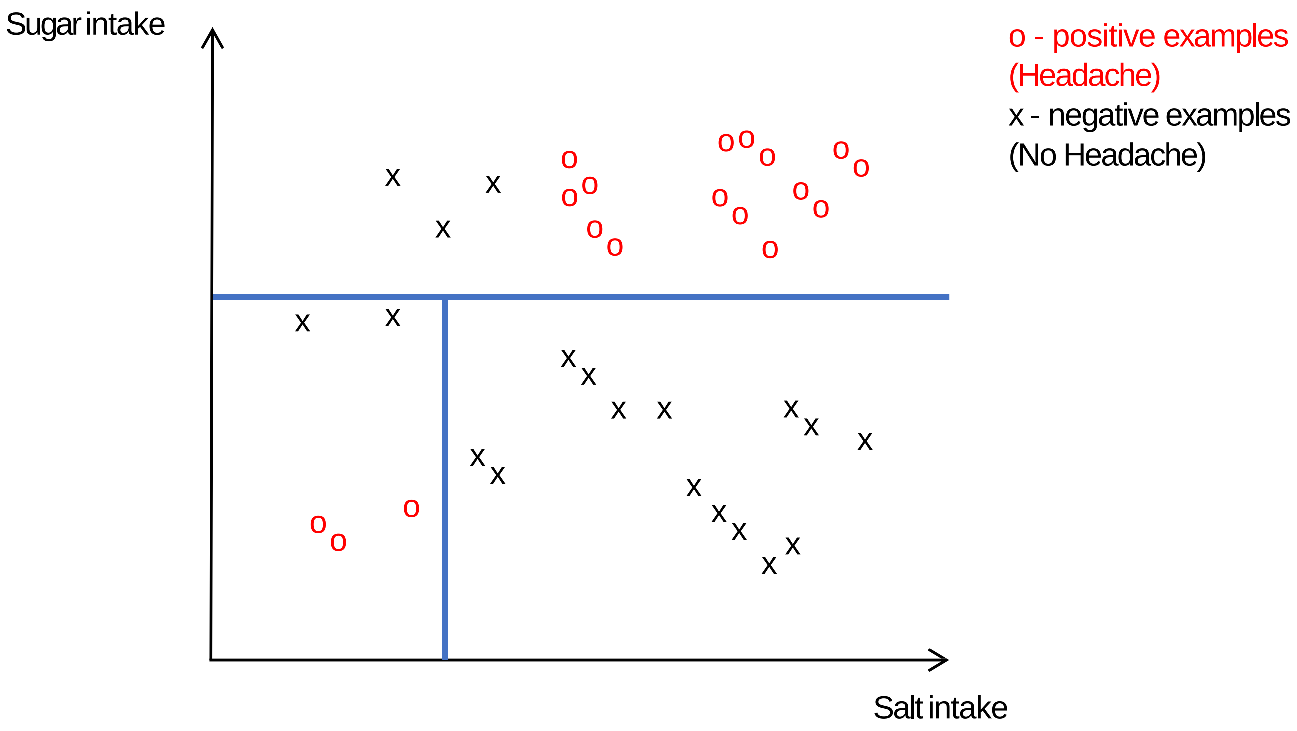
Recursive Partitioning

At each iteration, we divide the data to group similar instances together



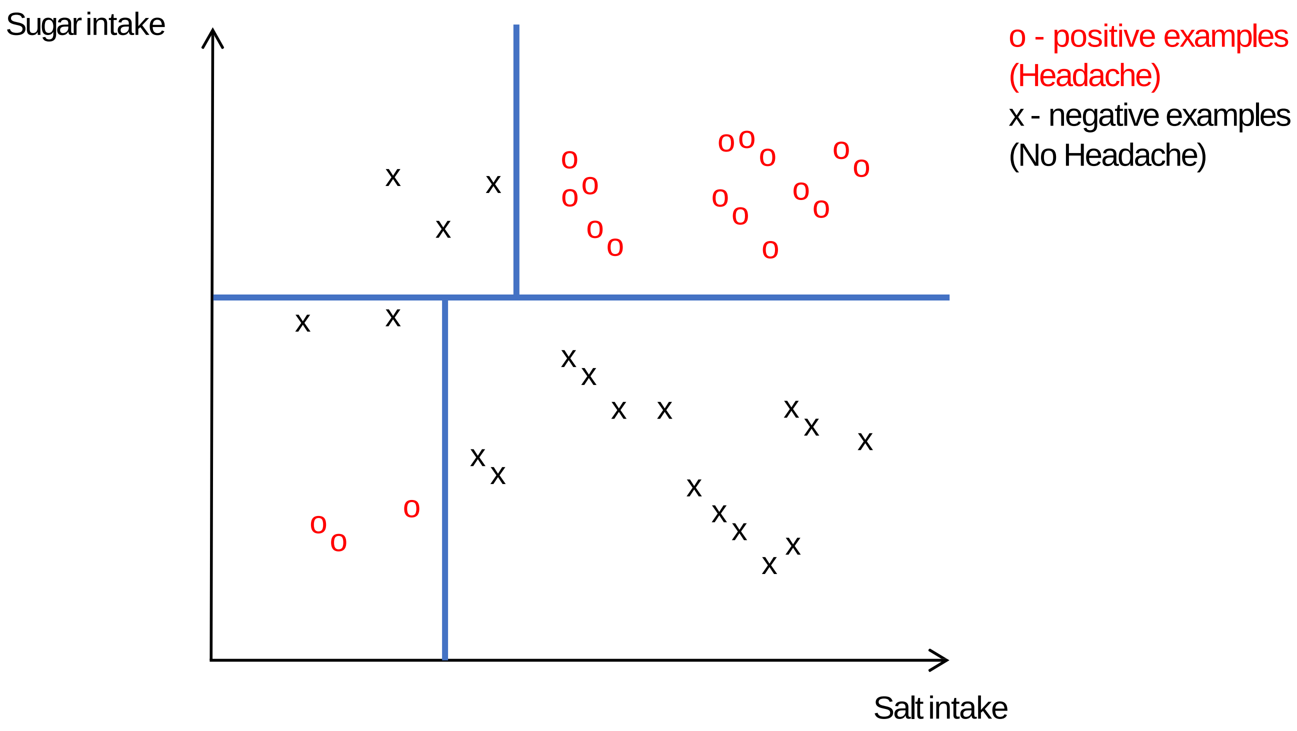
Recursive Partitioning

At each iteration, we divide the data to group similar instances together



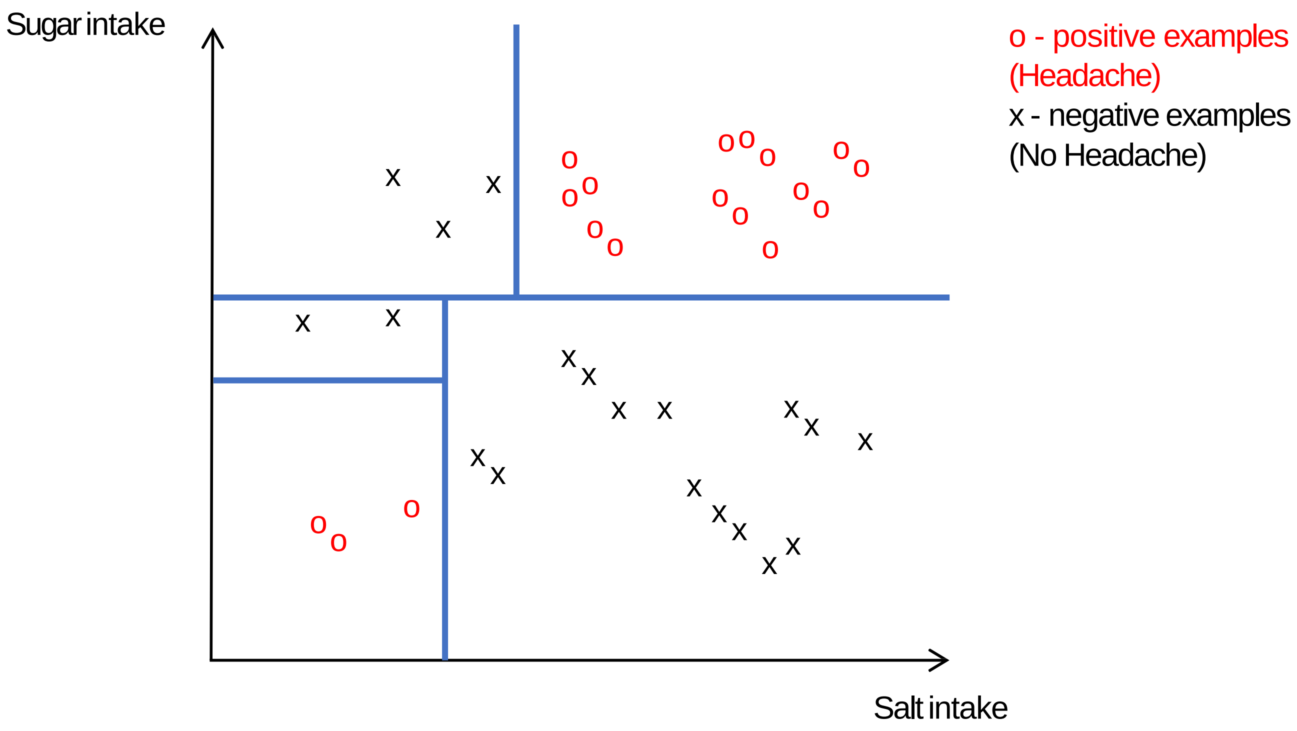
Recursive Partitioning

At each iteration, we divide the data to group similar instances together



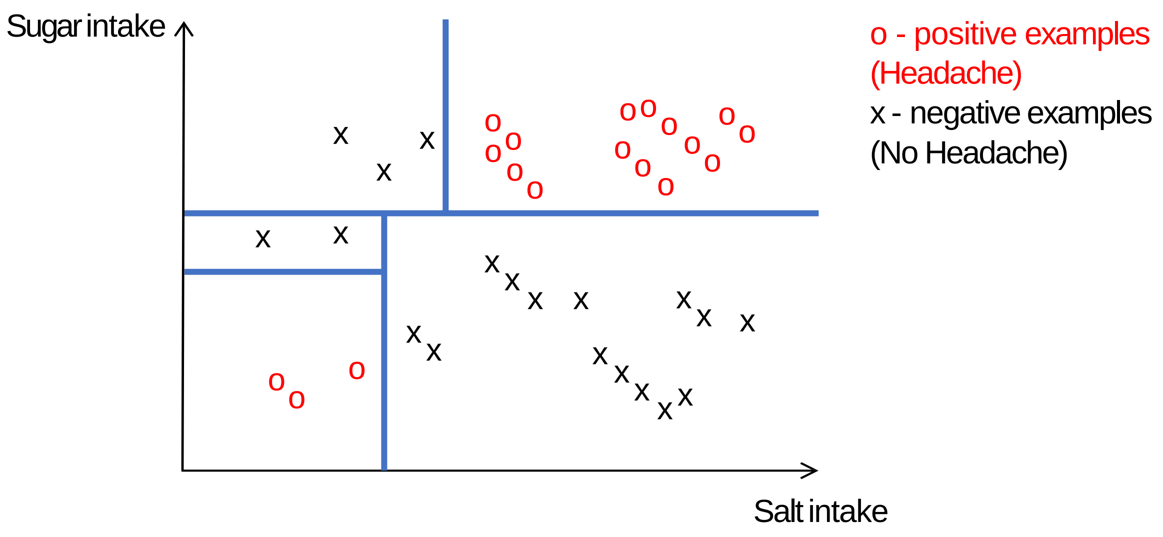
Recursive Partitioning

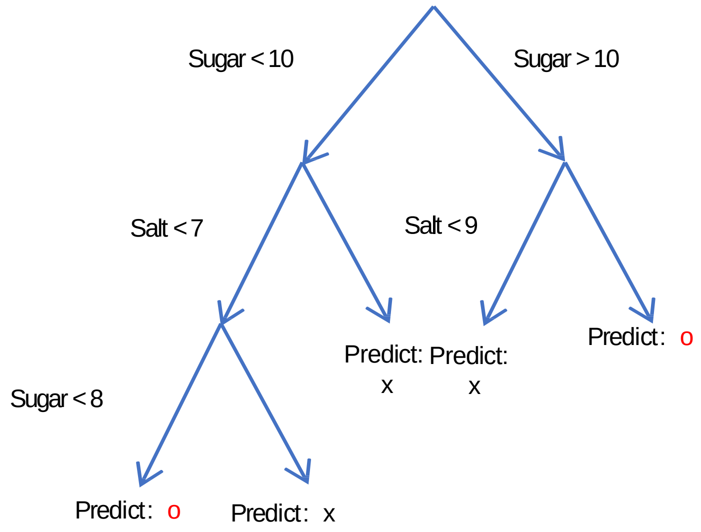
At each iteration, we divide the data to group similar instances together



Prediction Model is a Tree

This model learnt can be represented as a tree with predictions at the leaves:





Prediction in Decision and Regression Trees

Decision Trees:

- Prediction is the <u>most common values</u> in each region

Regression Trees:

- Prediction is usually <u>the average value</u> in each region

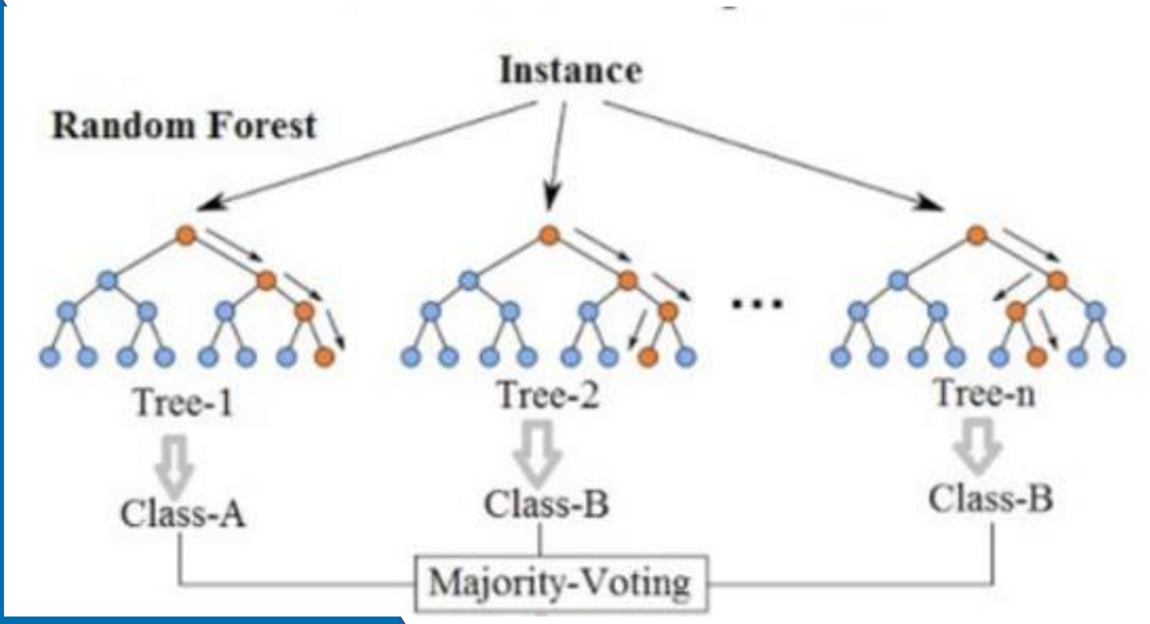
MONASH University

Decision and Regression Trees

More Information

- Algorithms for building Decision & Regression trees differ on the criteria (e.g., Entropy) used to:
    - Decide on which feature to split on in each iteration
    - Decide when to stop splitting

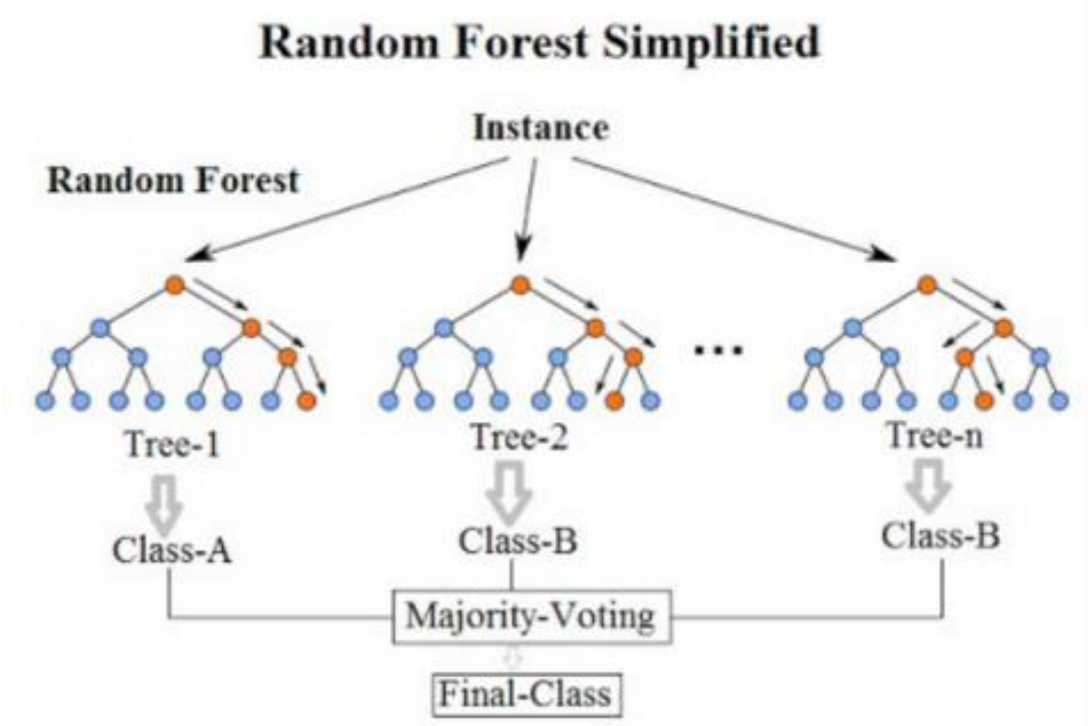
Random Forest Algorithm



MONASH University

What is a Random Forest?

Ensemble learning method that operate by constructing a number of decision trees



It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.