MA5810: Introduction to Data Mining

Week 1; Collaborate Session 1; Intro

Martha Cooper, PhD

JCU Masters of Data Science

2019-22-10 (updated: 2021-07-06)

Housekeeping

- Collaborate 1 = Tuesday 6.45-6pm (Martha)
- Collaborate 2 = Thursdays 6.45-6pm (Martina)

For my Collaborate Sessions, you can get the slides & R code for each week here:

https://github.com/MarthaCooper/MA8510

Note: Weekly content will be updated on Tuesday each week

Subject: MA5810 Intro to Data Mining

MA5810 Learning Outcomes

1. Overview of Data Mining and Examples (Today)
2. Unsupervised data mining methods *e.g.* clustering and outlier detection;
3. Unsupervised and supervised techniques for dimensionality reduction *e.g.* PCA;
4. Supervised data mining methods for classification *e.g.* Naive Bayes, LDA;
5. Apply these concepts to real data sets using R.

Assignments

Time management is important!

Quiz 1 due 13/07/21 (No credit)

Assessment 1 due 25/07/21

Assessment 2 due 08/08/21

Assessment 3 (Capstone) due 18/08/21

Check the course outline for the Extension Policy and more information.

Today's Goals

- Understand the major roles of data mining within the broader scope of data science
- Classify the most common problems involved in data mining as:

predictive *vs* descriptive

unsupervised *vs* supervised tasks

- Learn RMarkdown

What is Data Mining?

The process of discovering useful...

Patterns

Information

Knowledge

Predictive models

...from large-scale data.

Data Mining Methods

Supervised Learning

What?

Find patterns in our data that explain a dependent variable, Y

Why?

Predict future values of the dependent variable, Y, using a set of independent variables, $X = X_1, \ldots, X_n$

How?

Regression, Classification

Unsupervised Learning

What?

Identify patterns in our data without defining a dependent variable, Y

Why?

Describe interesting patterns in the current set of independent variables, $X = X_1, \ldots, X_n$

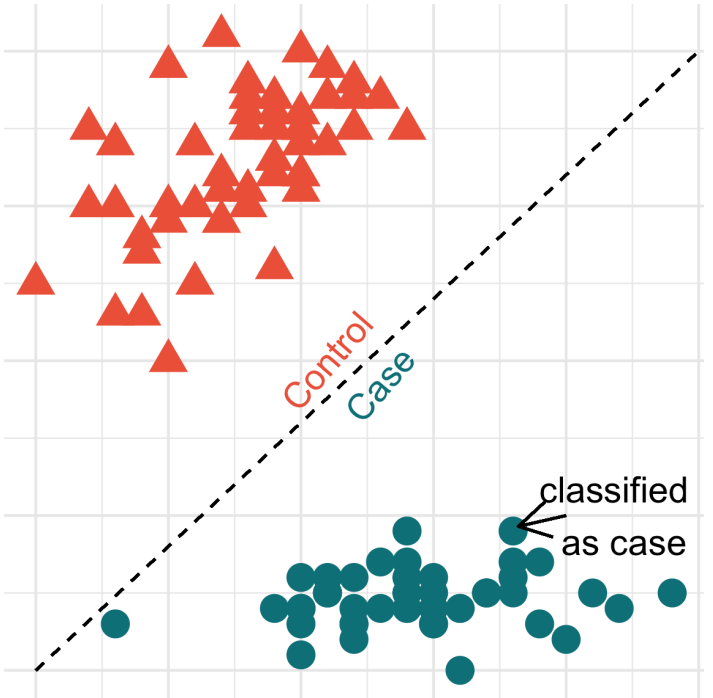
How?

Clustering, Outlier detection

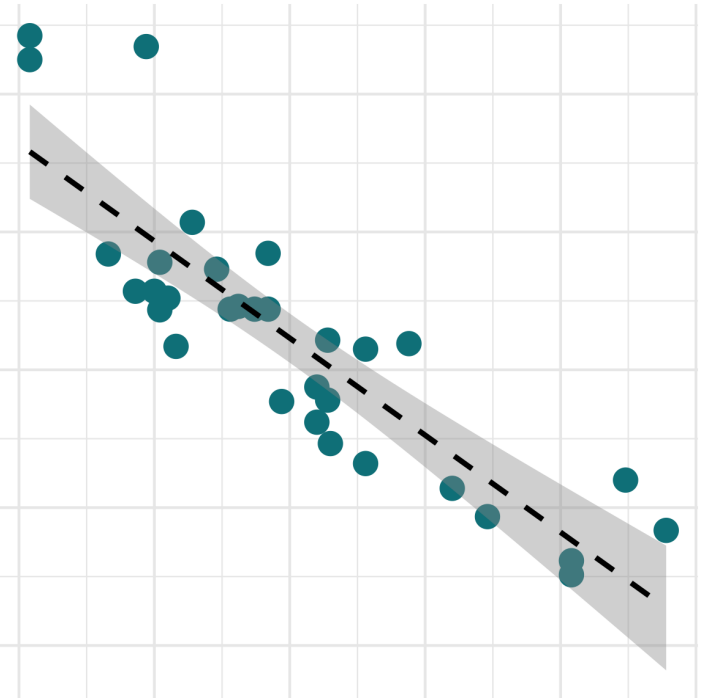
Supervised Learning

- The dependent variable, Y, is defined (data is "labelled")
- Used in predictive data mining tasks
- Training the model is called supervised learning

Classification

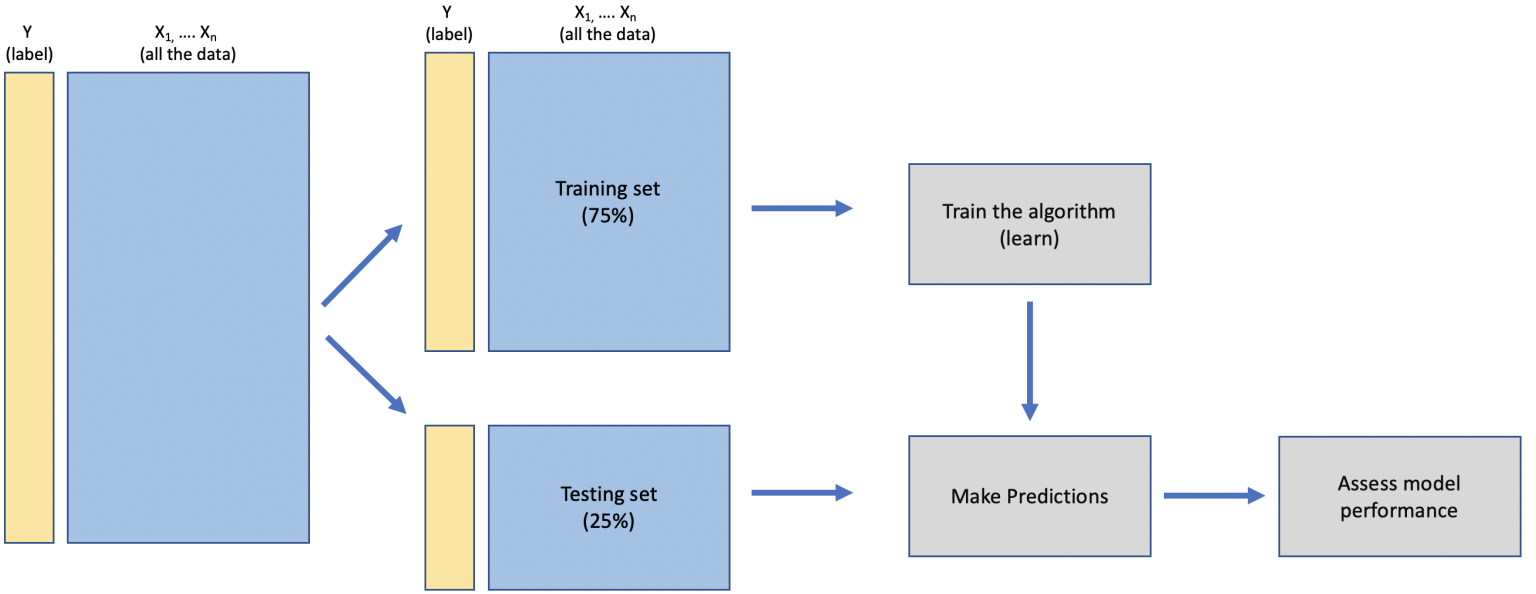


Regression



Supervised Learning

A supervised learning workflow:

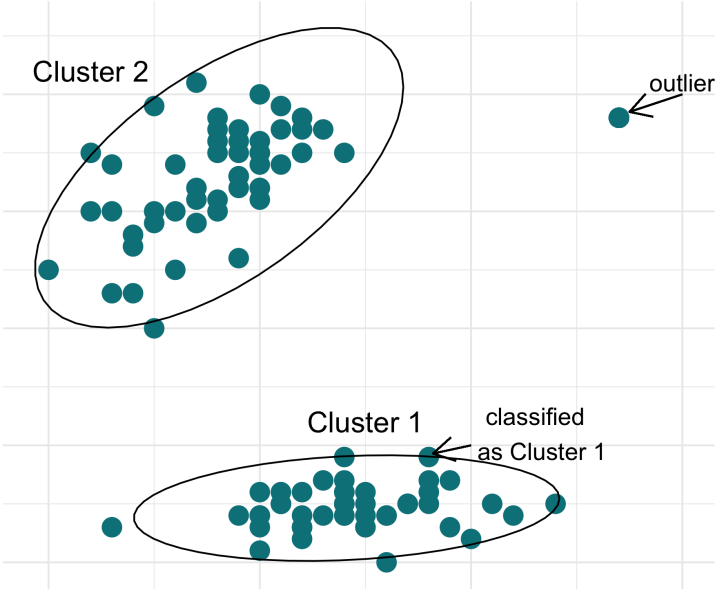


e.g. Naive Bayes Classifiers, Logistic Regression

Unsupervised Learning

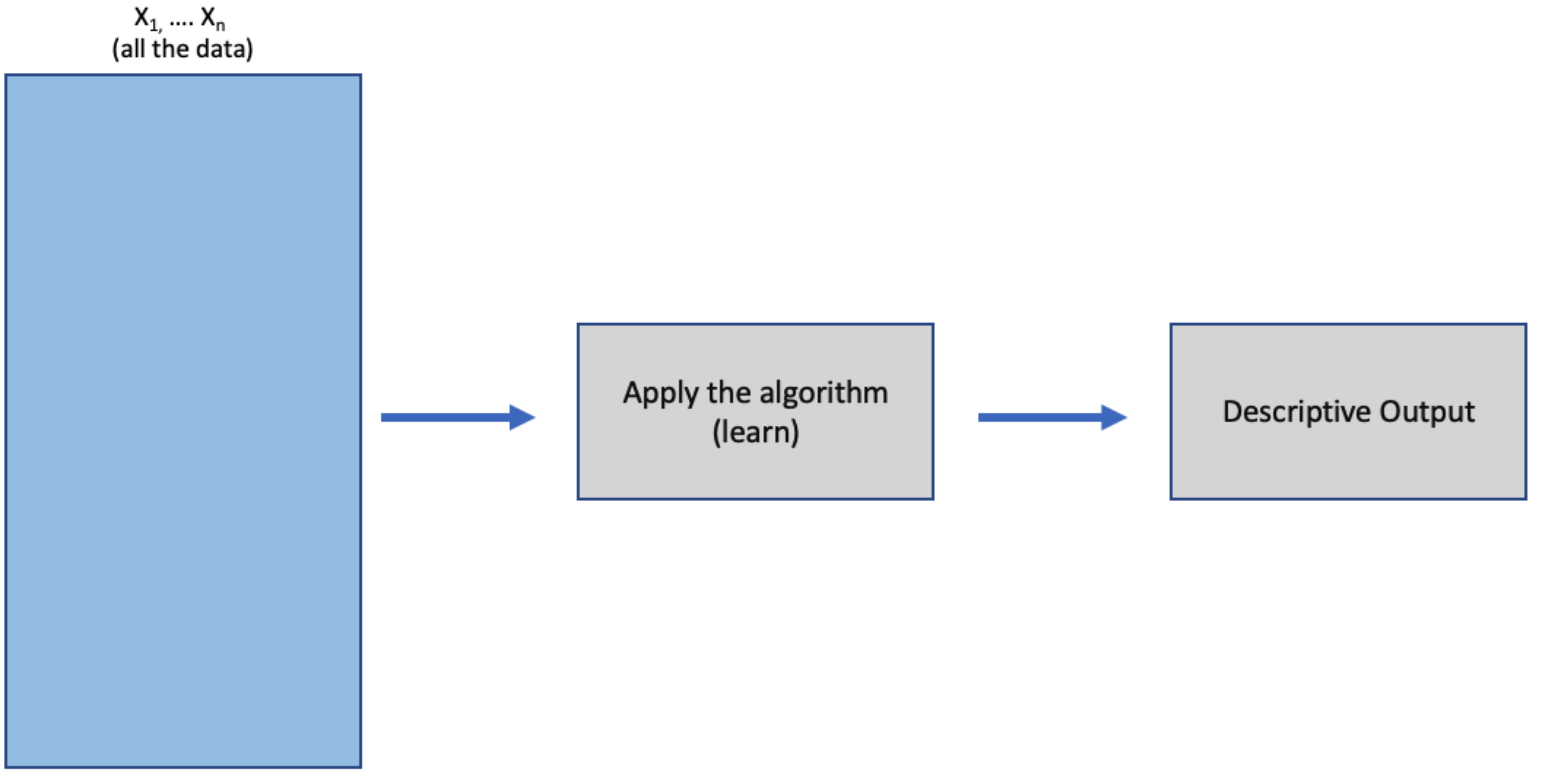
- We don't know (or define) a dependant variable (data is "unlabeled")
- Used in descriptive data mining tasks
- Training the model is called unsupervised learning

Clustering, Outlier Detection



Unsupervised Learning

An unsupervised learning workflow:



e.g. Principal Components Analysis (PCA), k-means clustering, hierarcical clustering

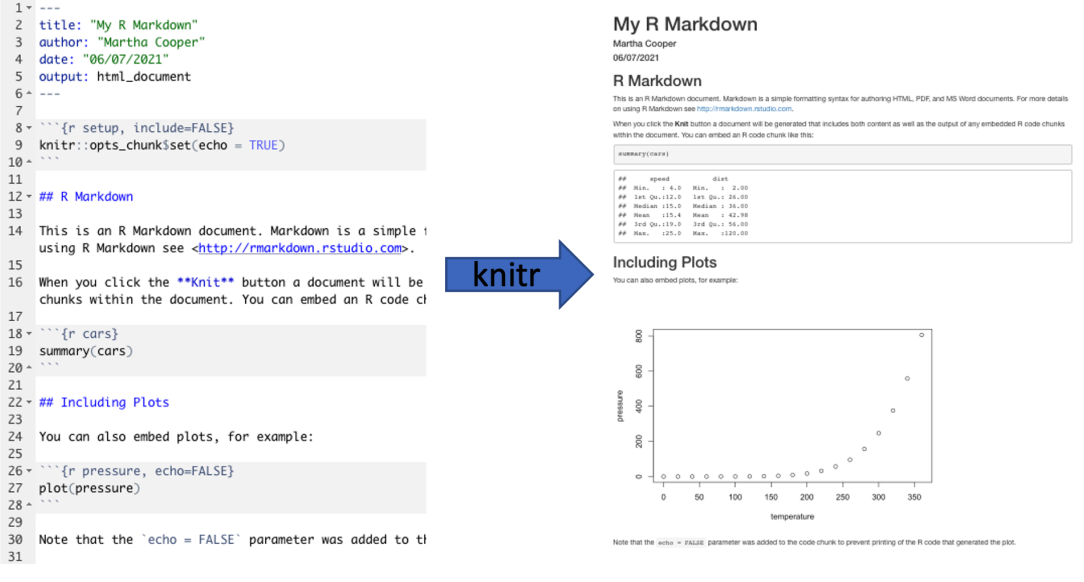
Task: Supervised *vs* Unsupervised?

1. Predictive Policing - forecasting when and where a crime will happen
2. Identifying subtypes of ovarian cancer based on genetic data
3. Automatic grading of students papers in some schools in china
4. A facial recognition system to identify gender
5. Dividing a set of photographs of people into piles containing each individual

R Markdown

R Markdown provides a notebook to:

1. Save and execute code
   - Use an R Markdown file to load data, run analyses, connect to databases
2. Generate high quality reports to share with an audience
   - Publish as a html, pdf, word file, slides, book, website etc...



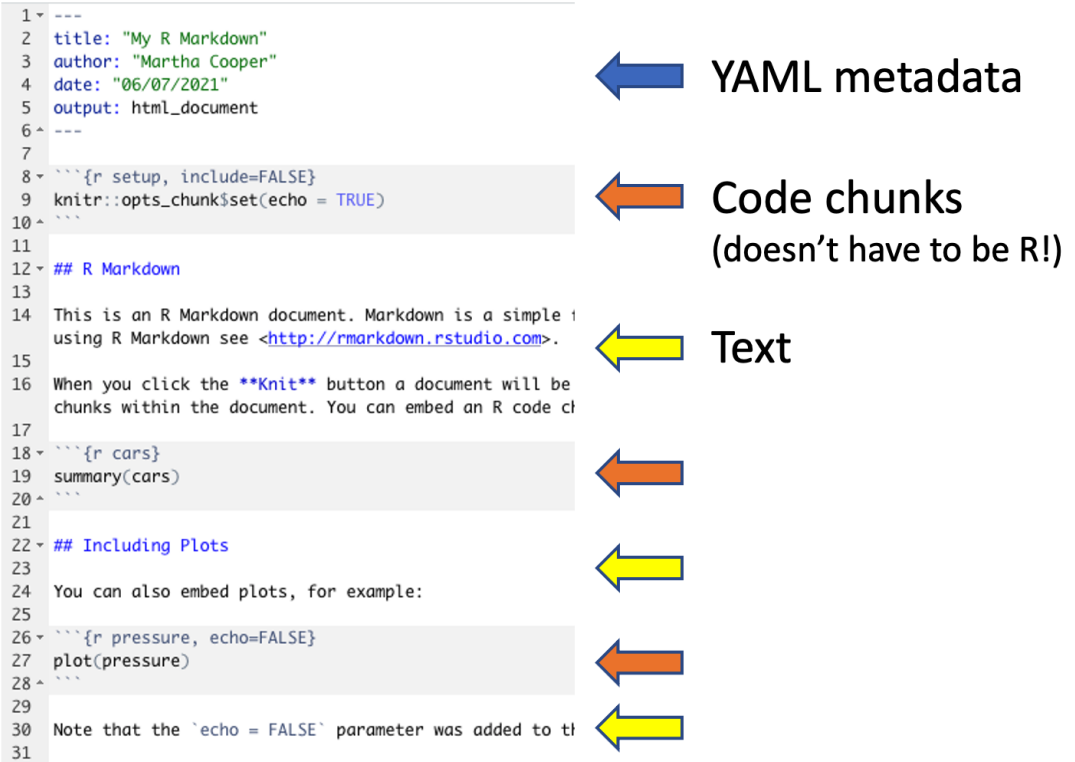
R Markdown

Why use R Markdown

- Reproducible
- Readable (contains text + code)
- Share-able
- Easy to use with version control (*e.g.* git)

R Markdown

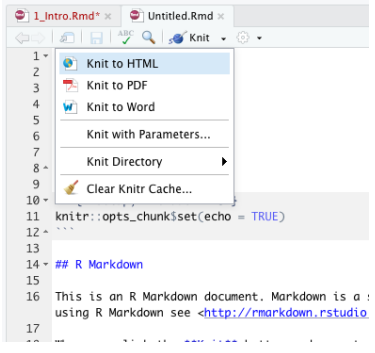
R markdown files have 3 types of content



R Markdown

Knitting with RStudio

Point and Click



In code

```
# From RStudio

rmarkdown::render("my_rmd.Rmd") #

pagedown::chrome_print(input = "m
```

Rmarkdown

How knitting works

R Markdown

Rendering text with Rmarkdown

```
# Heading 1
## Heading 2
### Heading 3

- Bullet pointed list
    - Sub-point list

1. Numbered list
    1. Sub numbered list

**Bold**, *Italic*

[link](www.mylink.com)
```

Heading 1

Heading 2

Heading 3

- Bullet pointed list
    - Sub-point list

1. Numbered list
    1. Sub numbered list

Bold, *Italic*

link

R Markdown

More information

- RStudio website
- R Markdown Cheatsheet

Extra reading/listening

- This stackoverflow thread
- This Guru99 tutorial
- Big Data Bioinformatics
- This Data Learner's podcast
-

References

R Markdown

- RStudio website

Slides

- xaringhan, xaringanthemer, remark.js, knitr, R Markdown