

MA5810: Introduction to Data Mining

Week 1; Collaborate Session 1; Intro

Martha Cooper, PhD

JCU Masters of Data Science

2019-22-10 (updated: 2020-11-03)

Housekeeping

- Collaborate 1 = [Wednesdays 6-7pm](#) (Martha)
- Collaborate 2 = [Thursdays 7-8pm](#) (Hongbin)

For my Collaborate Sessions, you can get the [slides & R code](#) for each week here:

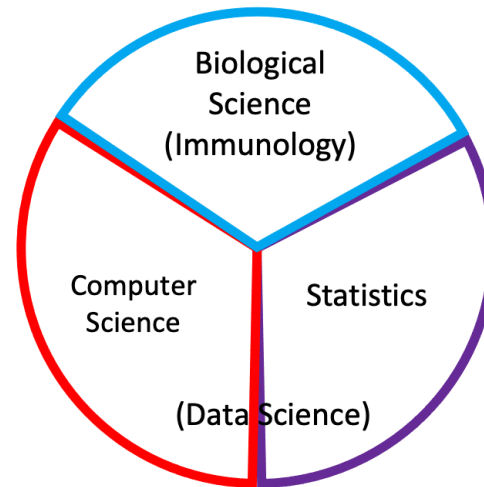
<https://github.com/MarthaCooper/MA8510>



Introduction

Dr. Martha Cooper

- I'm a research scientist at the [Australian Institute of Tropical Health and Medicine, JCU](#)
- Immunology & Bioinformatics
- I use Data Science to understand how people's immune systems respond to parasite infections.



Email: martha.cooper@jcu.edu.au

MA8510 Discussion board: [Saturday & Sunday](#)

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;
4. Supervised data mining methods for pattern classification;
5. Apply these concepts to real data sets using R.

Assignments

Time management is important!

Assignment 1 due Sunday Week 3 (30%)

Assignment 2 due Sunday Week 5 (30%)

Assignment 3 (Capstone) due Wednesday Week 7 (40%)

The **Extension Policy** has been updated. Check the course outline for 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

- Understand the main challenges for data mining in the context of Big Data analytics

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, \mathbf{Y}

Why?

Predict **future** values of the dependent variable, \mathbf{Y} , using a set of independent variables, $\mathbf{X} = \mathbf{X}_1, \dots, \mathbf{X}_n$

How?

Regression, Classification

Unsupervised Learning

What?

Identify patterns in our data without defining a dependent variable, \mathbf{Y}

Why?

Describe interesting patterns in the **current** set of independent variables, $\mathbf{X} = \mathbf{X}_1, \dots, \mathbf{X}_n$

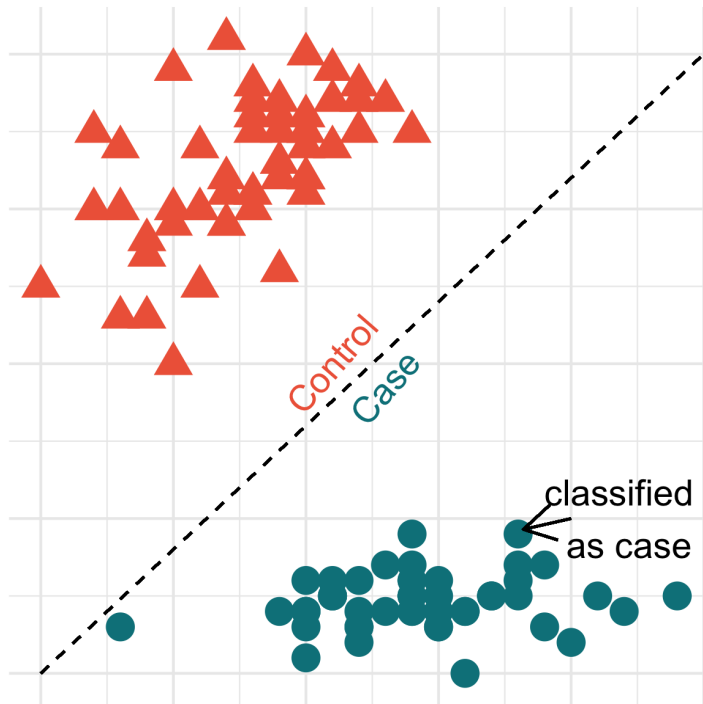
How?

Clustering, Outlier detection, Frequent Pattern Mining

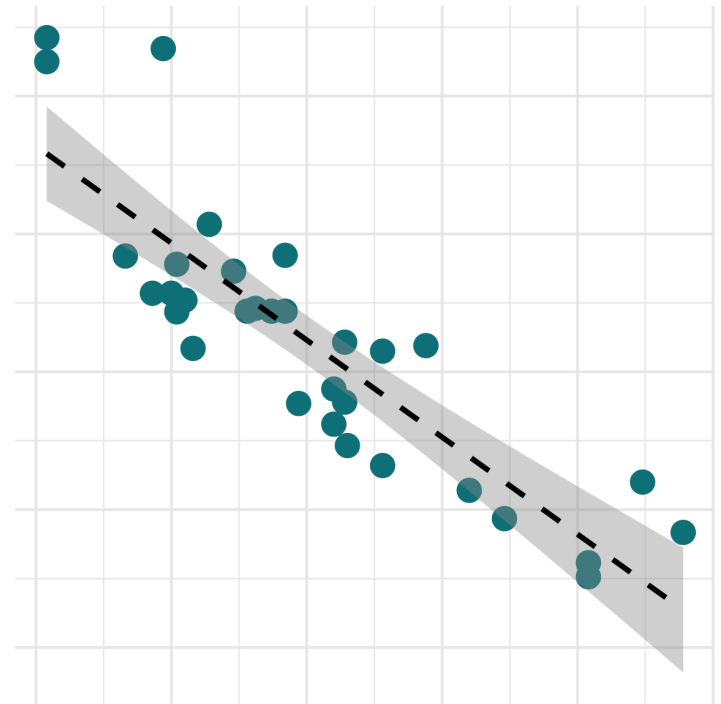
Supervised Learning

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

Classification

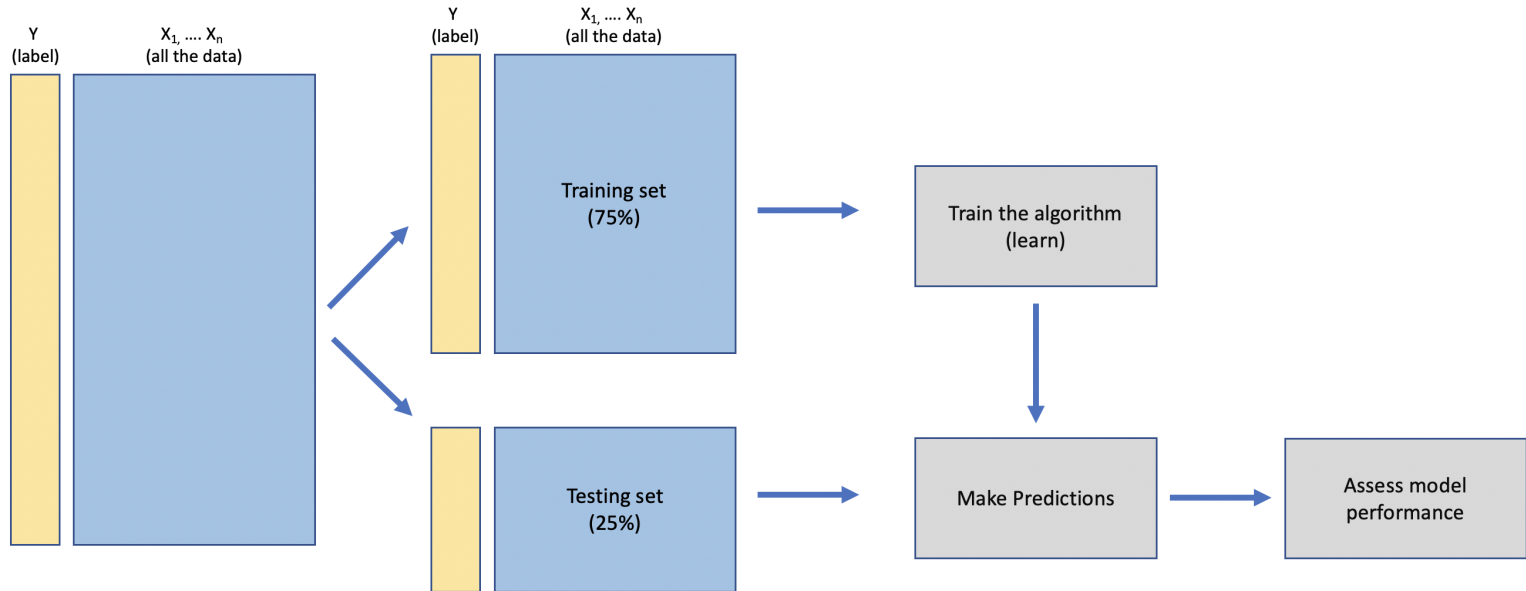


Regression



Supervised Learning

A supervised learning workflow:

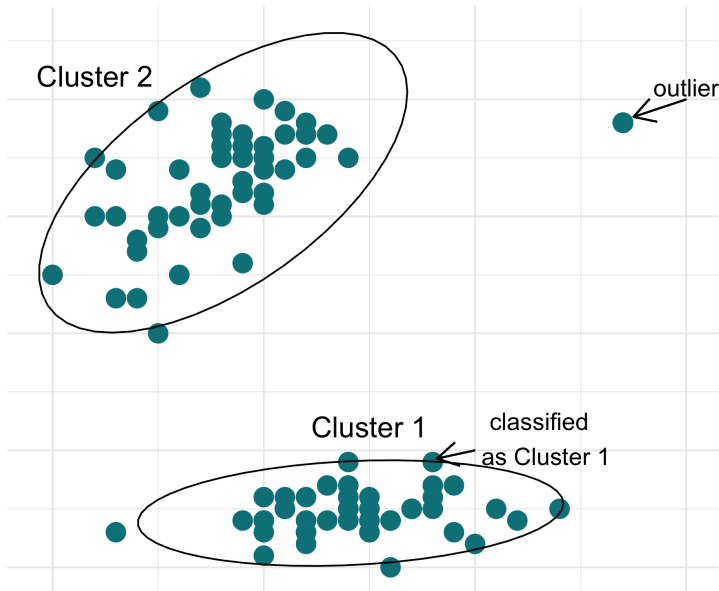


e.g. Naive Bayes Classifiers, Logistic Regression

Unsupervised Learning

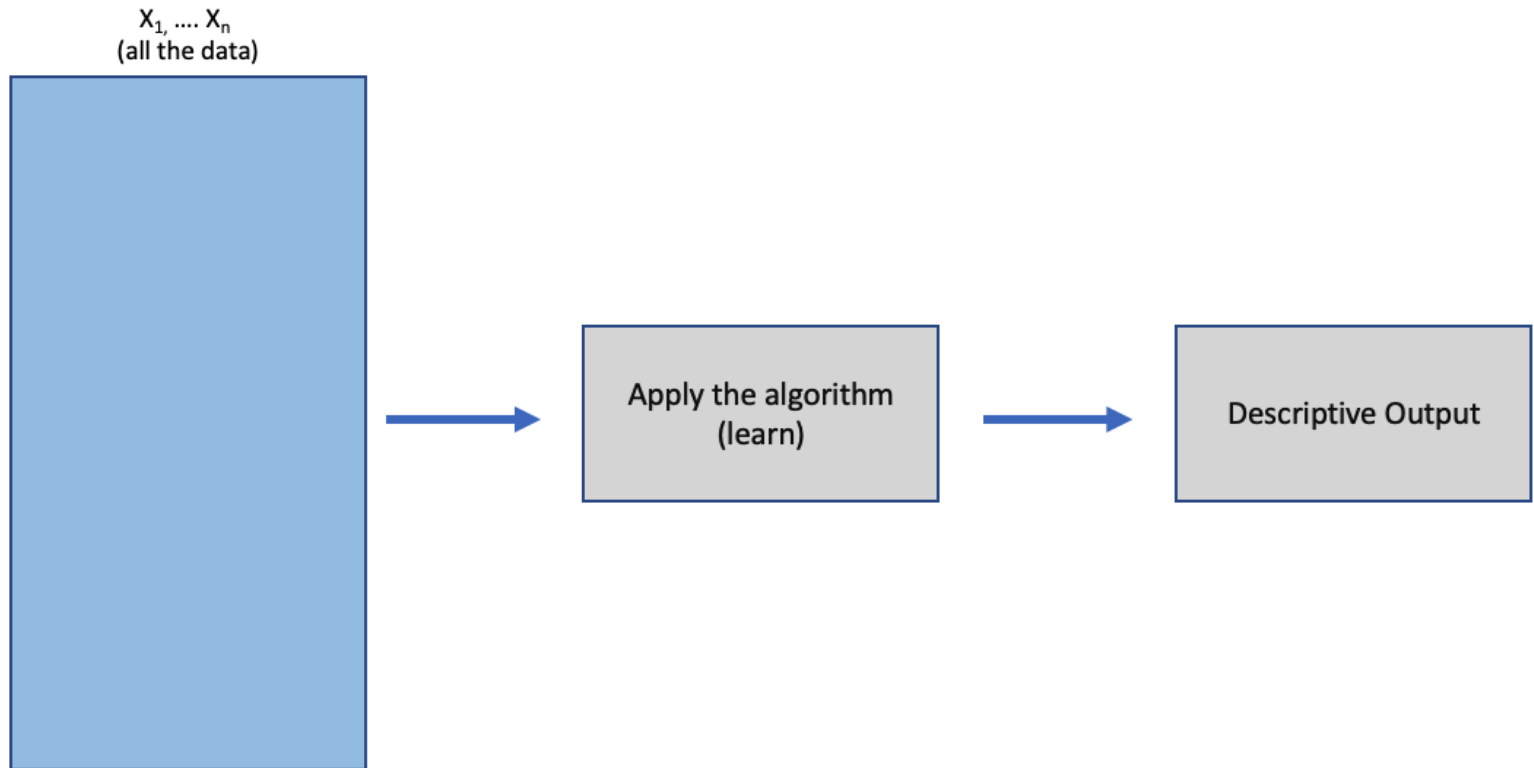
- We don't know (or define) a dependant variable (data is "unlabeled")
- Used in **descriptive** data mining tasks
- Training the model is called unsupervised learning

Clustering, Outlier Detection



Unsupervised Learning

An unsupervised learning workflow:



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

Task 1: 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 [Chinese schools](#)
4. A facial recognition system to [identify gender](#)
5. Dividing a set of photographs of people into piles containing each individual

Task 2: Challenges for data mining in the context of Big Data

Any ideas?

Task 2: Challenges for data mining in the context of Big Data

- Heterogeneity
- Complexity
- Data Privacy and Security
- Storage
- Computation Issues

Extra reading/listening

Get used to using stackoverflow:

- This [stackoverflow thread](#)

Still stuck? Go here:

- This [Guru99 tutorial](#)

Want a challenge? Go here:

- [Big Data Bioinformatics](#)

Just for fun:

- This [Data Learner's podcast](#)

References

Slides

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