dm22s1

Topic 12 : Wrap Up

Part 01 : Overview

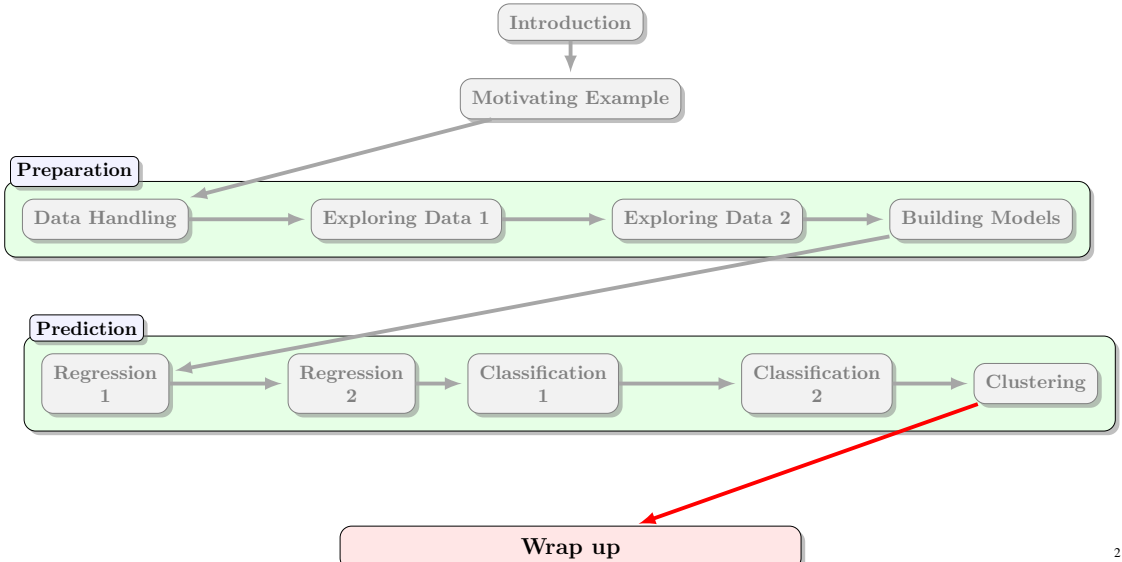
Dr Bernard Butler
Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Autumn Semester, 2022

Outline

- Review of module
- Example layout and tips

Data Mining (Week 12)



Overview — Summary

Outline

This Week's Aim

This week's aims are to

- Review the material covered in the module, especially the algorithms
- Outline the structure and marking of the sample exam

Outline

Introduction

What are the most important take-aways from this module

01-Data Mining History and Process

- Need to cut through hype and commentary by non-experts and those with (commercial) agendas
- Instead, focus on the key concepts and definitions of *big data*, *machine learning*, ...
- Data mining is the overarching *process*; what are its models and procedures?
- Understand how ICT advances enabled new applications, requiring new machine learning techniques, which enable...
- But is this a virtuous cycle? What about societal effects of unethical data mining?
- With the growing maturity of deep learning: how can we trust the ultimate "black box" of deep learning?
- In the lab we considered how unspecialised tools can be used for data analytics

02-Pandas and a simple classification example

- Pandas is the workhorse of data mining tools in python
- Used for data import/export, managing dataframes (naming, adding columns,...) and series
- More complex operations (filtering, aggregating, sorting) are also possible
- Used heavily and assessed as such in the CA programming assignments!
- Classification is one of the classic machine learning tasks: predicting a label, given data
- Introduced k nearest neighbours as a simple algorithm, based on voting, to identify the most likely label.
- What are its strengths and weaknesses? When would it be used?

03-Data handling: Storage and Computation

- Storage is used both for persistence and archiving: which is better for data mining?
- There are many data models and management systems: when to use each
- CAP theorem and its implications
- Big data requires a new approach, supporting out-of-memory computation
- Characteristics of Hadoop and Spark: how they work and how they differ
- Computational models and worked examples

04-Exploratory Data Analysis 1 and 2

- EDA requires time and care
- A 3-Phase process was described and students have used this in CA1 and CA2

06-Data Modelling

- This is the central focus of the module!
- Explain what a linear model is and how it can be extended (gneerally with mnore complex features)
- Training vs test data
- Objectives of modelling: explain a domain vs predict a result
- Components of error: bias, variance
- How can we control errors? How do we refine models and when do we stop?

07-Regression1

- What is regression and what types of features, targets are needed?
- What are the assumptions and what happens if they do not hold?
- How do we judge the success of the model?
- Distinguish between and statistical and machine learning metrics

08-Regression2

- What options do we have if vanilla regression models are not sufficient?
- Use of multivariate approaches to improve models
  - Regularisation: ridge regression to down-weight some features; lasso to drop them entirely
  - Dimensionality reduction: find a more economical subset of the features
  - What are the advantages and disadvantages?
- Role of correlation: between features and between a feature and the target

09-Classification1

- How it differs from regression
- conversely, how a variable tranformation can enable logistic regression to be used for classification
- Confusion matrices (true/false positives/negatives) and the derived ratios
- When to use a metric and how to interpret it in practice

10-Classification2

- Use of probability based classification techniques
- Naive Bayes: its derivation and worked examples
- Entropy in data mining and how it leads to the decision tree method
- Algorithm and worked examples
- Choosing classification technique for a given problem

11-Clustering

- How clustering differs from classification
- Partitional vs hierarchical clustering
- Role of distance metrics, their definition and calculation
- EM algorithms: how they work
- Derivation of K-means and how it can be extended if needed
- How GMM relates to k-means
- Motivation for density-based approaches and their pros and cons
- Derivation of the DBSCAN algorithm and how to tune it
- Role of different linkages and interpretation of dendrograms

Outline

Overview of Exam

ID	Section	Type	Marks
Q1	A	10 True/False	20
Q2	B	Parts/Essay	10
Q3	B	Parts/Essay	10
Q4	B	Essay/Essay	10
Q5	C	Parts/Worked	25
Q6	C	Parts/Worked	25

- All questions in Sections A, B and C are mandatory.
- There may be choice within a question, e.g., between essay topics

Exam - section A

- True/False questions can come from any part of the module
- They test detailed knowledge: read the statement carefully!
- Be careful of “some” versus “all” statements
- Validate your answer: try to come up with a counterexample, etc
- Note that negative marking does not apply, so students are encourage to try their “best” answer
- Questions can be drawn from anywhere in the module

Exam - section B

- There are three, mostly multipart, questions in this section, each worth 10 marks
- All questions should be attempted (worth 30 marks)
- You may be asked to derive one or more formulas; you will also generally be asked where the formula is used
- you may also be asked to interpret analysis results, e.g., to say how successful it was
- you may also be asked a more essay-style question. For example, it could be a compare-and-contrast question, so you should describe 5 major points of similarity or difference, to obtain the full 10 marks.
- you may also be asked quite specific questions arising from a more general concept
- General comments about reading the question, time management, etc., also apply here!

Exam - section C

- There are two multipart questions in this section, each worth 25 marks
- The early parts of each question are structured like the multi-part questions in Section B
- You may be expected to use one of the algorithms we covered in class to compute some quantity. The data sets will be small to make this possible in exam conditions. You should bring a scientific calculator with you.
- You may be presented with some results and asked to interpret/validate them.
- the questions in Section C are designed to test whether you can “do” data mining, not just remember facts, so they have a more applied feel than those in Sections A and B
- Time management is important here: Section C is worth half the marks, but you could get bogged down. Always write down anything you learned that is relevant.
- Questions may be drawn from anywhere in the module.