dm25s1

Topic 12 : Wrap Up

Part 01 : Review

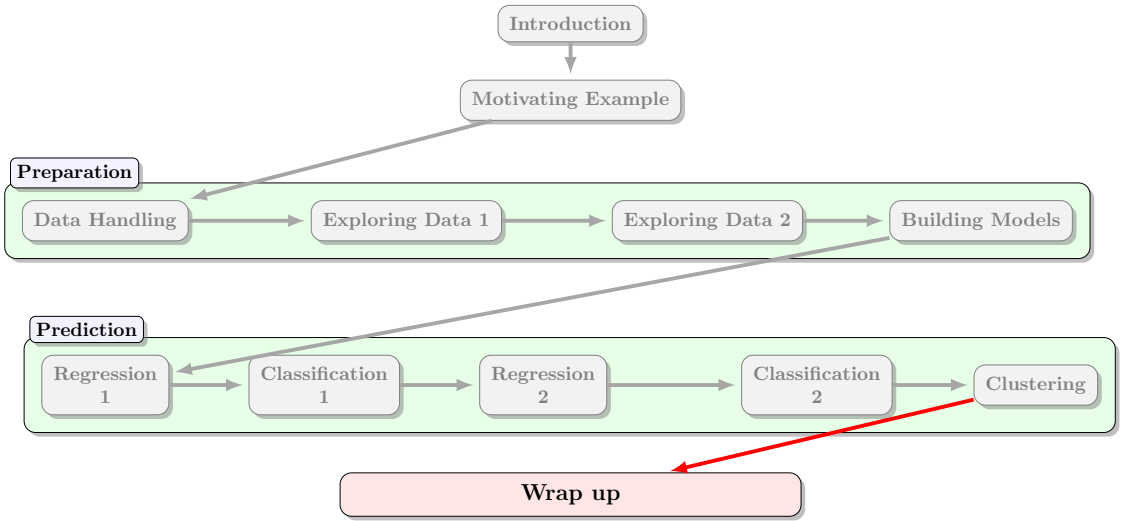
Dr Bernard Butler
Department of Computing and Mathematics, WIT.
(bernard.butler@setu.ie)

Autumn Semester, 2025

Outline

- Review of module

Data Mining (Week 12)



Outline

This Week's Aim

This week's aims are to

- Review options for feature encoding (an important step in feature engineering)
- Review the material covered in the module, especially the algorithms

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?
- When generative AI is added to the mix - what are the ethical concerns there?
- 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 (like DBMS and Excel) 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

- Python offers many data structures (lists, tuples, sets, dicts and operations on them)
- Numpy (borrows concepts from Matlab) module extends Python's mathematical processing and introduces arrays with extended semantics - masking, etc.
- Pandas (borrows concepts from R) introduces Series and Dataframes, with greater support for missing data types and performing operations like data cleaning and preparation
- Since approximately 80% of ML programmer effort is typically devoted to data preparation, cleaning and understanding (i.e., EDA), familiarisation with these libraries is critical.

04-Exploratory Data Analysis 1 and 2

- EDA requires time and care
- A 3-Phase process was described and students have used this in their CA attempts.
  - Phase 1 Read in the data, profile it, check (and fix) datatypes, column names, missing values, duplicates, ...
  - Phase 2 For each feature and the target, assess its distribution, impute missing values, check assumptions, ...
  - Phase 3 Examine correlations between columns, investigate feature-target and feature-feature relationships, identify possible column transformations, ...

06-Data Modelling

- This is the central focus of the module!
- Explain what a linear model is and how it can be extended (generally with more complex features)
- Training vs test data - why do we split the 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?
- Role of cross validation and metrics

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-Classification1

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

09-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, derived subset of the features
  - What are the advantages and disadvantages of each approach?
- Role of correlation: between features and between a feature and the target

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
- Algorithms and worked examples
- Choosing a 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
- Role of different linkages and interpretation of dendrograms for hierarchical clustering
- When would hierarchical clustering be used?
- EM algorithms: how they work
- Derivation of K-means and how it can be extended if needed
- How GMM relates to k-means (and, notionally, extends it)
- Motivation for density-based approaches and their pros and cons
- Derivation of the DBSCAN algorithm and how to tune it

And finally. . .

And finally...

Best of luck in your assessments!