dm22s1

Topic 01 : Module Overview

Part 04 : Review

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

Autumn Semester, 2022

Outline

- Selected questions

DM pipeline and Roles

- What are the main stages in the Data Mining pipeline?
  1. Obtaining data in the form of observations and attributes of those observations
  2. Preprocessing the data, also known as “exploratory data analysis”
  3. Using data-intensive machine learning procedures to derive insights
  4. Postprocessing the results, to answer questions
- Explain the roles played by the following in Data Mining
  1. Big Data
  2. Artificial Intelligence
  3. Machine Learning
  4. Deep Learning

DM terms

- Compare and contrast the following terms
  1. Artificial Intelligence vs. Deep Learning from notes
  2. Data Scientist vs. Data Engineer
     - from notes

Growth of Data Mining

- Give 3 reasons for the growth of data mining
  1. Needed to support moves towards automation (reducing costs, improving quality. . .)
  2. Growth of big data and the opportunity to monetise it
  3. Knowledge discovery and its potential to solve evolving problems

DM generations

- Contrast the first generation of data mining approaches with those used in 2022, under the following headings
    - Where data comes from
        - Transaction systems, to Unstructured web/social data
    - How the data is processed
        - In-database + Offline extracts, to pipelined streaming engines
    - How the results of that processing are used
        - operations and reporting, to knowledge discovery and integrated controls
    - Describe the technological trends in your answer.
        - see notes

Batch versus Streaming - advantages

- Compare batch and streaming approaches to data mining, giving 2 advantages and 1 disadvantage of each.
  - batch
    - advantages are easier control; consistent state
    - disadvantage is results lag the data (possibly by hours)
  - streaming
    - advantages are well suited to functional approaches; much reduced lag
    - disadvantage is repeatability is more difficult

Batch versus Streaming - scenarios

- Give two scenarios where each of the two appraches might be preferred (other factors being equal).
    - batch:
        - reports for 3rd parties: compliance, accounting
    - streaming:
        - fraud detection, textual stream classification

Expertise and Roles

- What types of expertise are needed to become an effective data scientist?
  - see notes, discuss Conway and Kolassa models…
- What is the role of *citizen data scientists* in Data Science teams?
  - original proposed to replace, now seen as adding to the data science team, with mentoring from senior data scientists

DM disadvantages

- List and describe 5 disadvantages of (increased adoption of) powerful data mining techniques, considering their effects on
  - individuals
    - making some jobs obsolete
    - biased/opaque decision making, e.g., for loan applications
  - groups of people
    - harmful effects on privacy
    - removing humans from decision making
  - the planet
    - energy usage in data centres
    - use of overly simplified models to represent complex systems

DM Ethics

- A colleague presents a proposal to use data mining to develop a case in support of a new business venture.
  - Identify at least 4 questions you might ask to decide whether the proposal is ethically sound.
    - How reliable is the data that has been obtained?
    - Is the data complete (in relation to the study aims)?
    - Does it go beyond what is necessary?
    - Does it relate to protected individuals or groups?
    - Has consent been obtained, and can the data subjects withdraw consent at any time?
    - What validation procedures and other controls are in place?
    - Can individuals be identified (even if pseudonymisation has been used)?
    - If the outputs affect individuals or groups, is there a procedure to review the findings?
    - Will the output be used to benefit the data processor at the expense of the data subjects or the wider community?

DIKW chain

- Decribe the main phases in the Data-to-Wisdom process, describing the transformations between each stage.
- see the lecture notes. Use examples for each phase and transformation.

CRISP-DM versus TDSP

- Compare and contrast CRISP-DM and Microsoft's Team Data Science Process (TDSP). In you answer, mention their
  - motivation
    - CRISP-DM was an attempt to introduce iterative development into waterfall-oriented organsiations
    - TDSP: link ML cycles into more modern devops practices
  - role of feedback loops
    - CRISP-DM: the whole process is a set of cycles, that drops deployment artefacts occasionally
    - TDSP: two separate cycles; deployment is itslf a process to be managed in parallel
  - integration with software engineering processes more generally
    - CRISP-DM: bridge waterfall and agile (particularlyfeature-oriented cyclical development) practices
    - TDSP: full agiles, integrates particularly with devops - resulting in *mlops*.

ML Tribes - Classification

- Classification is a common objective in data mining. For each of the following “tribes” of machine learning, identify a classification technique they might favour, and descibe why that might be the case.
  - Analogizers
    - Support Vector Machines; Decision Trees
  - Bayesians
    - Logistic Regression (and variants)
  - Connectionists
    - Feedforward artificial neural network

Regression versus Classification

- Regression and classification both learn from training data and can be used for prediction
    - How are they similar (at least 2 ways)?
        - learm from training data having both attributes and target
        - metrics can be computed to measure the quality of the learned parameters
        - in addition to parameters to be estimated, both need hyperparameters to be chosen
    - How are they different (2 ways)?
        - Target is numeric for regression and categorical for classification
        - Classification has a richer set of algorithms
        - Results analysis for classification is more complicated

Classification versus Clustering

- Identify two ways in which classification and clustering differ.
    - Classification learns from labeled data, clustering learns from unlabeled data.
    - Clustering is more often used for data exploration rather than prediction

Undesirable and Desirable Anomalies

- Give 2 examples of scenarios where anomalies are desirable and two where they are not. Justify your answer.
  - undesirable anomalies: identifying faults in manufacturing or fraudulent transactions in banking
  - desirable anomalies: discovering new medical treatments or subatomic particles (leading to new theories)

Regression versus Time Series

- Regression and Time Series Analysis both try to predict numeric values from data. Identify 2 indications from the data that might suggest the use of time series analysis in preference to simple regression.
    - time series data is equispaced (in space and/or time)
    - time series data is serially correlated (earlier data predicts later)

ARM and Recommendation Systems

- Describe at least two use cases where each of the following techniques might be useful
  - association rules mining (ARM)
    - grouping transactions by product choices
    - grouping documents by word combinations
  - recommender systems
    - cross-selling and up-selling in shopping sites
    - online dating