Introduction to Machine Learning and Data Mining

COMP5318 Machine Learning and Data Mining

semester 2, 2024, week 1a

Nguyen Hoang Tran

Based on slides prepared by Irena Koprinska (irena.koprinska@sydney.edu.au)

Reference: Witten ch.1, Tan ch. 1

THE UNIVERSITY OF SYDNEY

Acknowledgement of Country

I would like to acknowledge the Traditional Owners of Australia and recognise their continuing connection to land, water and culture. I am currently on the land of the Gadigal (Cadigal) people of the Eora Nation and pay my respects to their Elders, past, present and emerging.

I further acknowledge the Traditional Owners of the country on which you are on and pay respects to their Elders, past, present and future.

Outline

- Administrative matters
- Introduction to machine learning and data mining

Administrative matters

Welcome to COMP5318!

- There are more than 500 students currently enrolled in this course – this is a very big course!
- Local and international, from various degrees
- Welcome to everyone!

Teaching staff

- Unit coordinator and lecturer
  - A/Prof. Nguyen Hoang Tran
  - Computer Science Building (J12), room 428,
- Teaching assistants
  - Canvas
- Tutors
  - Canvas

Timetable

- Lectures
  - 2 hours weekly, 6-8pm, start in week 1
  - live-streamed Zoom, available to watch after the lecture time – see Canvas "Recorded lectures"
- Tutorials (also called labs or pracs)
  - 1 hour weekly
  - Start in week 2
  - You need to attend 1 tutorial only as per your timetable
  - Please attend your allocated tutorial – see the tutorial number in your timetable

Course website

- The main place for this course is the Canvas website; we will use it for:
  - all teaching materials (unit outline, lecture slides, tutorial notes, tutorial solutions, assignments)
  - posting marks
- All other relevant systems will be linked to the Canvas website:
  - discussion board (Ed)
- Important document on Canvas->Home: unit outline file – contains the most important information about this course

Course materials

- The lecture slides
- Tutorial solutions will be available on Friday evening after the last tutorial, which finishes at 9pm
- All on Canvas

Discussion board Piazza

- We will use Ed Discussion, it is linked to Canvas
  - You must have access to Ed, it will be the main communication channel for this course
  - When you click on Ed in Canvas, you should be able to see the posts for this course, otherwise you are either not enrolled in Ed or there is another problem
- Posting questions on Ed
  - Post your question on Ed instead of emailing them to us – this is beneficial for everyone
  - The question will be answered quicker
  - When it is answered, it is answered for everybody (and often many students have the same question)

Assessment overview

- 4 components:
  1. Assignment 1 – 15% (week 7)
  2. Assignment 2 – 25% (week 11)
  3. Exam – 60%

Assessment overview - assignments

- Two assignments
  - Programming assignments using Python and its machine learning libraries
  - Given a problem, you need to apply machine learning algorithms to solve it

  - Assignment 1 (15%) - due Friday week 7; in pairs (no more than 2 people are allowed)
    - Computer program and report
    - Submitted via Canvas (code and report)

  - Assignment 2 (25%) - due Friday week 11; group work (no more than 3 people are allowed)

    - Computer program and report
    - Submitted via Canvas (code and report)

Assessment overview – assignments (2)

- Assignments are due at 11.59pm
- Late submissions – allowed up to 5 days late
  - Late penalty of 5% per day will apply
  - Assignments submitted more than 5 days late will not be accepted

- Important: Start working on the assignments as soon as possible, do not delay them until a few days before the deadline!
  - Programming assignments require time; even the ones that look simple, almost always require much more time than expected!
  - Submit early to avoid last minute problems and busy systems

Assessment overview - exam

- Exam: 60% (individual), during the examination period
  - A minimum of 40% on the exam is required to pass the course – School of Computer Science policy.
  - More information about the exam will be provided later in the semester

Assumed knowledge

- You need to have programming skills for this course
- We expect that all students have a background in at least one programming language, preferably Python
- In this course we will use Python and its libraries, e.g. sklearn
- If you don't know Python or haven't used it recently, we recommend that you watch free online courses about Introduction to Python for Machine Learing/Data Science
  - https://learn.datacamp.com/courses/intro-to-python-for-data-science
  - https://www.edx.org/course/python-basics-for-data-science
  - https://www.coursera.org/learn/python-data-analysis
- We have also prepared a short Python refresher document – see Canvas

Tutorials – Jupyter and Colaboratoty

- We will use Jupyter Notebook during the tutorials for the Python part
  - See the document on Canvas on how to install it on your computer, and how to install some required packages, e.g. graphviz
- If you prefer, instead of Jupyter Notebook, you can use Colab – Google's Jupyter notebook environment

  https://colab.research.google.com/notebooks/welcome.ipynb

- In summary, there are some documents on Canvas related to the practical part of this course:
  - How to install Jupyter Notebook
  - Python refresher (short document)

Tutorials – structure

- For most of the weeks there will be 2 documents with tutorial exercise:
  1) Theoretical - involving paper-based exercises and calculations, testing your understanding of the algorithms
  2) Practical - using Python and its machine learning and neural network libraries (in Jupyter Notebook format .ipynb)
- Theoretical
  - We will do the first theoretical exercise either at the lecture or the beginning at the tutorial
  - The rest should be done at your own time. Make sure that you do all theoretical exercises as they are similar in style to the exam questions.
- Practical
  - The main focus of the tutorial. Sometimes it may not be possible to finish all. Please do this at home as this part is important for your assignments. We have prepared very detailed notes for the practical part, we hope you will find them useful.
- The solutions will be provided on Wednesday evening

Textbooks

- **Textbooks:**
  - Ian H. Witten, Eibe Frank, Mark Hall and Christopher J. Pal
  - *Data Mining - Practical Machine Learning Tools and Techniques*, 4th edition, Morgan Kaufmann, 2017 (You can also use the 3nd edition)
  - Pang-Ning Tan, Michael Steinbach, Anuj Karpathe and Vipin Kumar (2019). *Introduction to Data Mining*, 2nd edition. Pearson. (you can also use the previous edition)
- **Books for the practical part using Python:**
  - Andreas C. Mueller and Sarah Guido (2016). *Introduction to Machine Learning with Python: a Guide for Data Scientists*, O'Reilly.
  - Aurelien Geron (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow,* O'Reilly.
- All are available from the library as both hard copies and online, except Tan which is available as a hard copy only

Special Considerations (SC) due to Illness or Misadventure

- There is a centralized University system:
  http://sydney.edu.au/special-consideration
- Applications are submitted online, after login to “myUni”
- You are required to submit the SC form within 3 working days from the date when the assessment was due
- Applications are assessed by the University Student Administration Services (SAS) unit

Do you have a disability that impacts on your studies?

- You may not think of yourself as having a ‘disability’ but the definition under the Disability Discrimination Act (1992) is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities
- The types of disabilities we see include:
  - Autism, ADHD, Bipolar disorder, Broken bones, Cancer, Cerebral palsy, Chronic fatigue syndrome, Crohn’s disease, Cystic fibrosis, Depression Diabetes, Dyslexia, Epilepsy, Hearing impairment, Learning disability, Mobility impairment, Multiple sclerosis and many others
- In order to get assistance, students need to register with Disability Services. It is advisable to do this as early as possible. Please contact us or review our website to find out more.
- Disability Services Office, sydney.edu.au/disability, 02-8627-8422

Cheating and plagiarism

- Please read the University policy on Academic Honesty carefully: https://sydney.edu.au/students/academic-integrity.html
- All cases of academic dishonesty and plagiarism will be investigated
- There is a centralized University system and database
- Three types of offenses:
  - Plagiarism – when you copy from another student, website or other source. This includes copying the whole assignment/exam answer or only a part of it.
  - Academic dishonesty – when you make your work available to another student to copy (for assignments or exams). There are other examples of academic dishonesty.
  - Misconduct - when you engage another person to complete your assignment/exam (or a part of it), for payment or not. This is a very serious matter and the Policy requires that your case is forwarded to the University Registrar for investigation.

Plagiarism detection

- We will use the similarity detection software TurnItIn and MOSS to compare your assignments and exam with these of other students (current and previous) and the Internet
  - Turnitin is for text documents (Assignments 1 and 2 reports and exam)
  - MOSS is for programming code (Assignment 1 and Assignment 2)
- These tools are extremely good!
  - e.g. MOSS cannot be fooled by changing the names of the variables or changing the order of the conditions in if-else statements

Student excuses

- These are cases of plagiarism and academic dishonesty from our school
- The student excuses are not acceptable and both parties were penalized
- *I finished my assignment but my friend had family problems. I felt sorry for her, so I gave her my assignment as an example. She said she only wanted to have a look and promised not to copy it.*
- *The test has finished but the tutor hasn't collected the papers yet. I showed my answer to my friend. I didn't expect him to copy it.*
- *He is my best friend. I had no choice but to let him copy my assignment.*
- *I couldn't find a partner to work in pairs, so I joined their pair as they are my friends* (when only groups of 2 are allowed – illegitimate collaboration – academic dishonesty).

Cheating and plagiarism – key message

- Please do not confuse legitimate cooperation with cheating. In individual assignments, you can discuss the assignment with another student, this is a legitimate collaboration, but you cannot complete the assignment together – everyone must write their own code and report.

- Plagiarism and any form of academic dishonesty will be dealt with, and the penalties are severe
- We use plagiarism detection systems such as MOSS and TurnItIn that are extremely good. If you cheat, the chances you will be caught are very high.
- If someone asks you to see or copy your assignment or exam answers, or to complete the assignment or exam instead of them, just say: *I can't do this. This is against the University policy. I will not risk my reputation and future by doing this.*

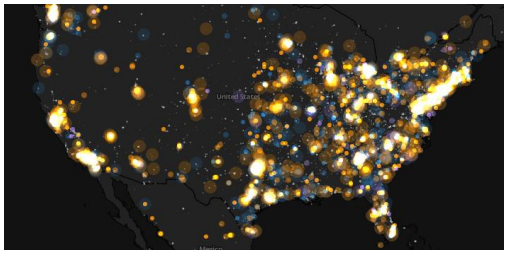
- **Be smart and don't risk your future by engaging in plagiarism and academic dishonesty!**

Introduction to Machine Learning and Data Mining

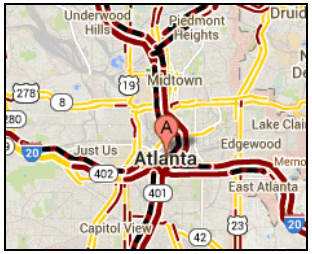
Data is everywhere

- Data explosion – society produces and stores huge amounts of data
  - Due to automated data collection tools and sensors, mature database technology, cheaper and more powerful computers
  - Sources: business, science, medicine, economics, environment, web, etc.
- Examples:
  - purchase data – supermarket, department stores, online stores – e.g. Amazon handles millions of visits a day
  - bank/credit card usage data
  - web data – Google, Facebook; other social networking sites
  - telephone call details, government statistics, traffic data

E-Commerce



Social Networking: Twitter



Traffic

amazon.com

Google

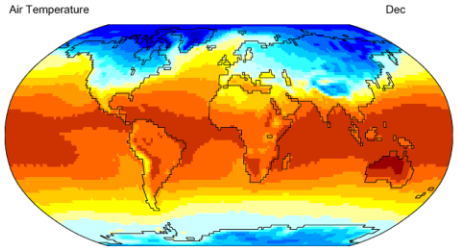
facebook

More examples

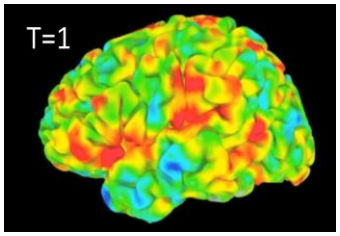
- Scientific data
  - telescopes scanning the skies
  - remote sensors on satellites
  - weather data

Sky survey data

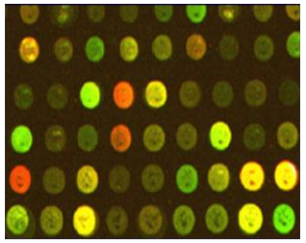
Sensor networks



Temperature



fMRI brain data



Gene expression data

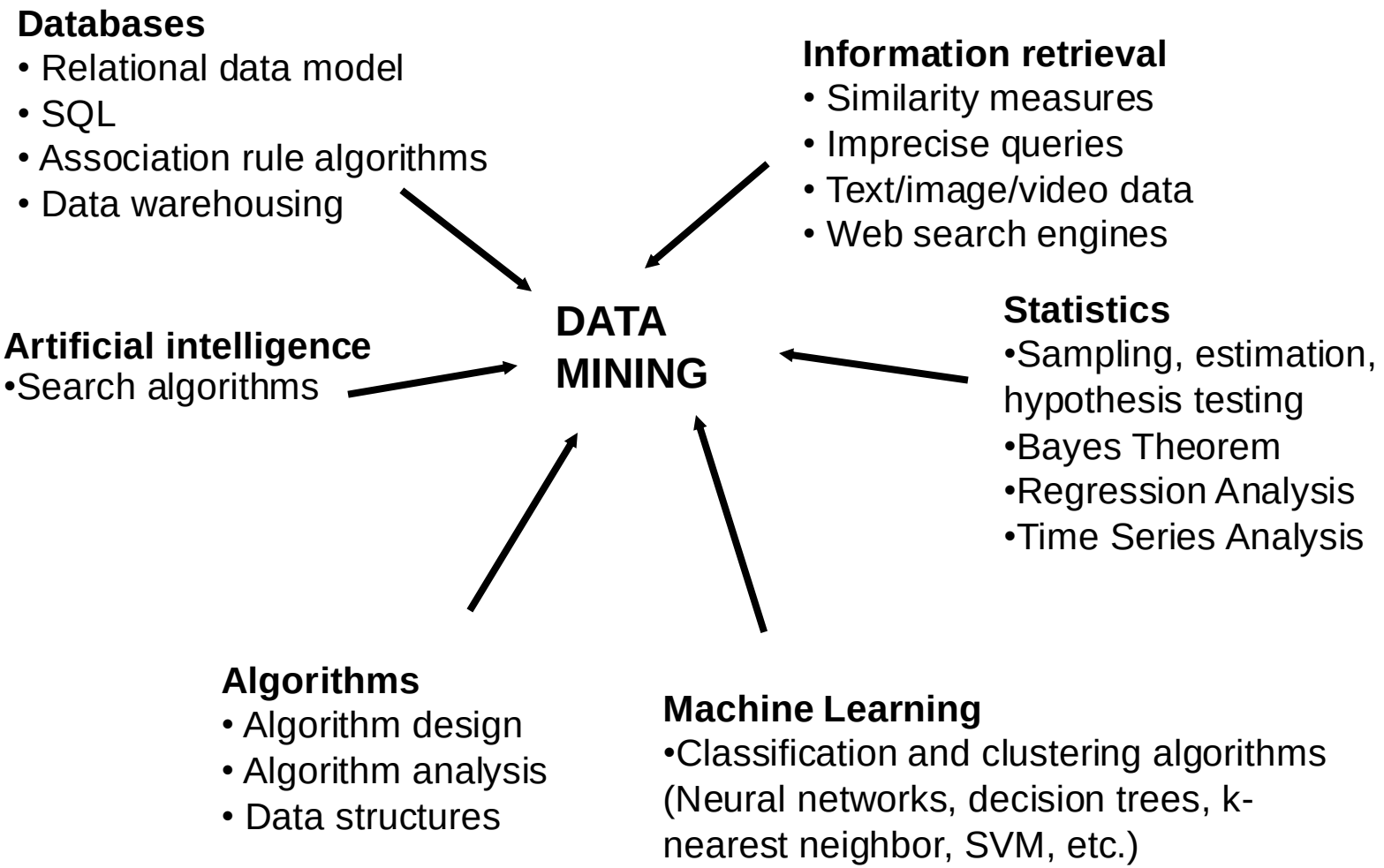
  - medical records and scans
  - biological data (high-throughput) – cytometry, gene expression

From data to knowledge – ML and DM

- Current trend: Gather whatever data you can, whenever and wherever possible!
- Expectation: it will be useful either for the purpose being collected or another purpose, not yet envisioned
- However, raw data is useless – need for methods to automatically extract knowledge (useful patterns) from it
- Machine Learning (ML) and Data Mining (DM) are concerned with finding patterns in data
  - These patterns should be meaningful, useful and actionable
  - The process is automatic or semi-automatic

- ML vs DM
  - ML is a core part of Artificial Intelligence
  - Most of the algorithms used for DM have been developed in ML
  - DM deals with large and multidimensional data, ML not necessary
  - DM can be seen as applied ML – we use ML algorithms to do DM

DM origins


Databases
• Relational data model
• SQL
• Association rule algorithms
• Data warehousing
Information retrieval
• Similarity measures
• Imprecise queries
• Text/image/video data
• Web search engines
Artificial intelligence
•Search algorithms
DATA MINING
Statistics
•Sampling, estimation, hypothesis testing
•Bayes Theorem
•Regression Analysis
•Time Series Analysis
Algorithms
• Algorithm design
• Algorithm analysis
• Data structures
Machine Learning
•Classification and clustering algorithms (Neural networks, decision trees, k-nearest neighbor, SVM, etc.)

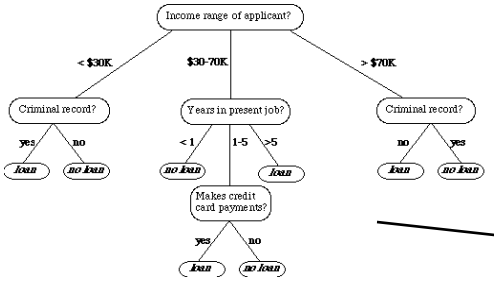
ML and DM tasks

Main tasks

- 2 main types of tasks:
  - Supervised learning - classification and regression
  - Unsupervised learning – clustering
- Other:
  - Association rule mining
  - Reinforcement learning
  - Outlier detection

- We will cover algorithms for supervised, unsupervised and reinforcement learning

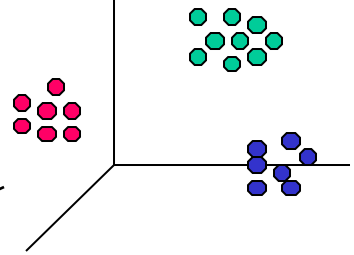
Main tasks - diagram



Supervised learning
(classification and regression)

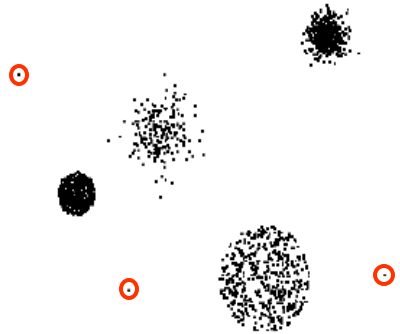
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



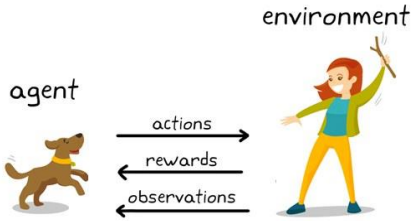
Unsupervised learning
(clustering)

Association rule mining



Anomaly
detection

Reinforcement
learning



Topics covered

Week	Topic
1	Administrative matters and course overview. Introduction to machine learning and data mining. Data: cleaning, pre-processing and similarity measures.
2	Nearest neighbour. Rule-based algorithms.
3	Linear regression. Logistic regression. Overfitting and regularization.
4	Naïve Bayes. Evaluating machine learning methods. Assignment 1 out
5	Decision trees. Ensembles.
6	Support vector machines. Kernels. Dimensionality reduction.
7	Neural networks - perceptrons and multilayer perceptrons. Assignment 1 due (9 Sep., 11.59pm) Assignment 2 out
	Mid-semester break
8	Deep neural networks I: CNN, RNN, and Transformer
9	Deep neural networks II: CNN, RNN, and Transformer
10	Clustering
11	Markov models. Assignment 2 due (14 Oct., 11.59pm)
12	Reinforcement learning.
13	Guest lecture. Revision.

Supervised learning

- Given: a set of pre-classified (labelled) examples {x,y}
  - x – input vector, y - target output
- Task: learn a function (classifier, model) that maps x->y and can be used predictively
  - i.e. to predict the value of y given the values of x for new, unseen examples
- Why is it called supervised?

- Two types of supervised learning
  - Classification: the variable to be predicted is categorical (i.e. its values belong to a pre-specified, finite set of possibilities)
  - Regression: the variable to be predicted is numeric

Classification - example

input vector, with 3 features

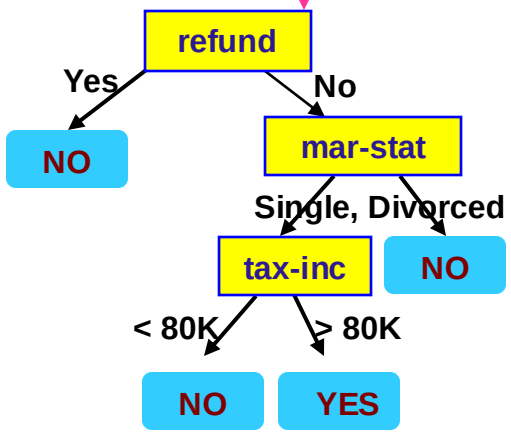
target class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

training data

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

new data



predict the class

Classifier

Step 1: Create the classifier
Step 2: Use it predictively on new data

More classification examples

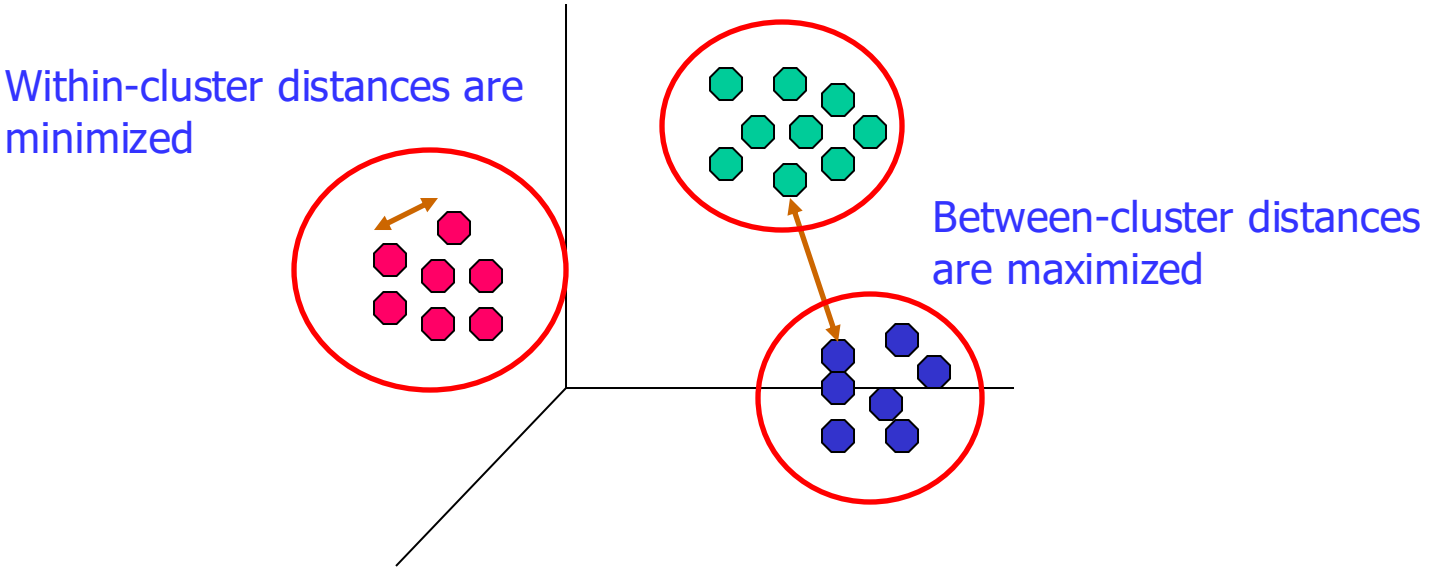
- Ex. 2: Fraud detection in credit card transactions
  - Data about customers and their transactions
    - previous credit card transactions
    - what they typically buy and when
    - demographic and socio-economic information - age, education, income, etc.
  - Label previous transactions as fraud or fair
  - Build a classifier to detect fraud transactions on new data (for new transactions of the same customer or for new customers)

Regression - examples

- Predict the electricity demand
  - Data: previous electricity demand, weather data and weather forecast data for the future days
  - Important to prevent blackouts and ensure reliable supply of electricity; also important for the economical and efficient operation of the electricity grid and for supporting the electricity market participants
  - Short and long term predictions - for the next few hours, next day, next week etc.; every 5 min, 30 min, 60 min, etc.
- Predict the exchange rate of AUD
  - Data from previous days, economical indicators, political events
- Predict retirement savings
  - Data: current savings and market indicators
- Predict the house prices in Sydney in 2030
- Predict the stock market index
- Predict wind velocity based on temperature, humidity, pressure

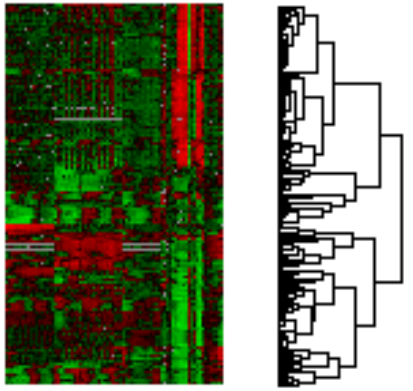
Unsupervised learning

- Given: a set of examples containing only input vectors x (no target outputs y)
- Task: group (cluster) the examples into a finite number of clusters, so that the examples
  - from each cluster are similar to each other
  - from different clusters are dissimilar to each other



Clustering - examples

- Ex.1: Targeted marketing
  - Segment customers into groups with distinct characteristics and use this knowledge to develop targeted marketing campaigns
  - (targeted campaigns are cheaper than mass-campaigns)
- Ex. 2: Customer loyalty
  - Analyse customer behavior and find groups of customer who are likely to defect, e.g. to another medical insurance, electricity or phone company
- Ex. 3: Gene clustering
  - Find genes with similar structure and functionality – important for understanding diseases and finding effective treatments
  - Data: microarray – from thousands of genes, analysed simultaneously



Clustering – examples (2)

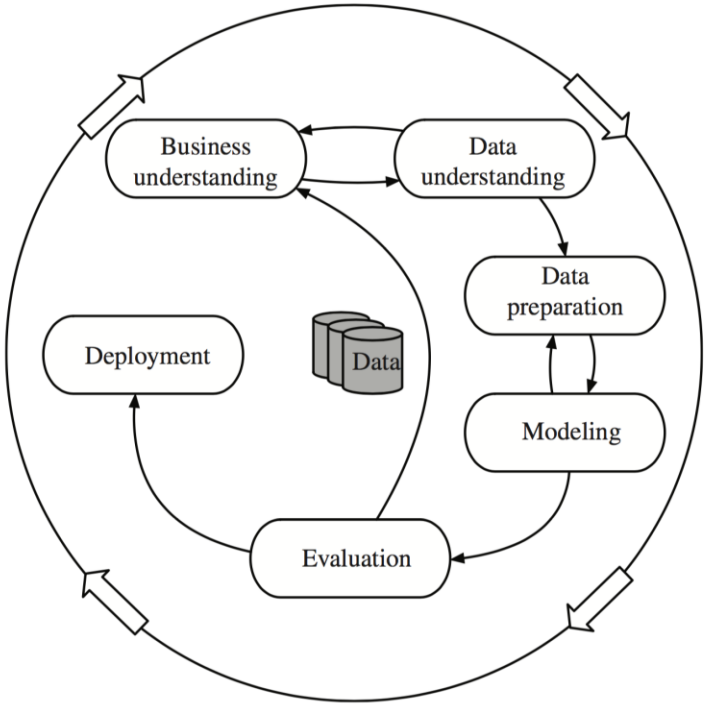
- Ex. 3: Document clustering
  - Find groups of documents that are similar to each other based on their content
  - Applications:
    - Patent documents assessment: group similar patent documents to make the evaluation of a new patent document easier
    - Personalized news recommendations

- Ex. 4: Clustering for understanding eating habits and dietary patterns of a particular cohorts (e.g. of young Australians)
  - Group 1: People who skip breakfast, care about weight, do not exercise regularly; eat high protein, low fat and high sugar diet; eat out because they enjoy the social aspect; snack after dinner
  - Group 2: …
  - Use this knowledge to promote good eating habits and changes in government policies

The DM process

1) Business understanding

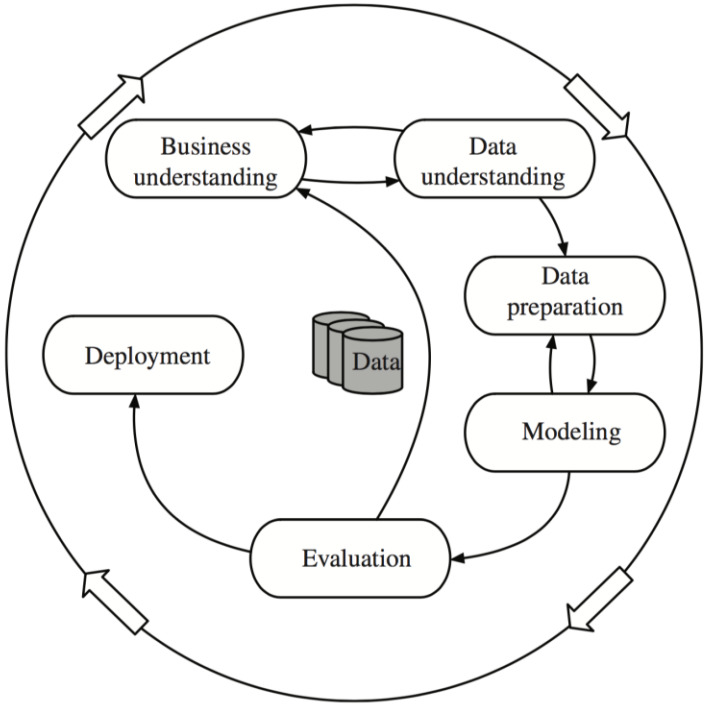
- Investigating the business objectives and requirements
- Deciding whether DM can be applied to meet them
- Determining what kind of data can be collected to build a deployable model



The DM process (2)

2) Data understanding

- Get an initial dataset; is it suitable for further processing?
- If the data quality is poor, collect more data based on more stringent criteria
- Gain insights from data and review the objective – can DM be applied?



The DM process (3)

3) Data preparation - preprocessing the data, so that ML algorithms can be applied. This involves cleaning and various transformations:

- Cleaning: data in real world is:
  - Incomplete, e.g. missing values
  - Noisy, e.g. containing errors or outliers
  - Inconsistent, e.g. in codes, names

  Fill in missing values, smooth noisy data, identify outliers and remove them, resolve inconsistencies
- Transformation – convert to common format; transform to new format; perform normalization, dimensionality reduction and feature selection

4) Modelling – building ML models, e.g. a prediction model

3) and 4) go hand and there are many iterations, e.g. the model informs the use of different preprocessing – e.g. use different feature selection and dimensionality reduction, build a model again

The DM process (4)

5) Evaluation – very important

- How good is the performance? E.g. accuracy, F1 measure, etc.
- Are the patterns meaningful and useful, or just reflecting spurious regularities?
- If the performance is poor, reconsider the project and return to step 1)
- If the performance is good -> deploy it in practice

6) Deployment

- Typically requires integration into a larger software system by software engineers
- May be necessary to re-implement the model in a different programming language