Introduction to Machine Learning and Data Mining

COMP5318 Machine Learning and Data Mining

semester 1, 2021, week 1a

Irena Koprinska

Reference: Witten ch.1, Tan ch. 1

Outline

- Administrative matters
- Introduction to machine learning and data mining

Administrative matters

Welcome to COMP5318!

- There are 386 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 weeks (1-6 and 13)
  - Associate Professor Irena Koprinska
  - Computer Science Building, room 450, irena.koprinska@sydney.edu.au
- Lecturer (weeks 7-12)
  - Dr Chang Xu
  - Computer Science Building, room 316, c.xu@sydney.edu.au
- Teaching assistants
  - Henry Weld and Givanna Putri
- Tutors
  - Henry Weld, Stephen McCloskey, Chen Chen, Nicholas Rhodes, Claire Hardgrove, Mashud Rana, Gio Picones and Thomas Selvaraj

Timetable

- Lectures
  - 2 hours weekly, 6-8pm, start in week 1
  - pre-recorded (not live-streamed), available to watch during the lecture time – see Canvas “Recorded lectures”
- Tutorials (also called labs or pracs)
  - 1 hour weekly on Monday, Tuesday or Wednesday
  - Start in week 2
  - You need to attend 1 tutorial only as per your timetable
  - There are 2 types: online (RE) and face-to-face (CC) - you chose one of them when you enrolled
    - Online: via Zoom, live-streamed
    - Face-to-face: in the School of Computer Science labs, level 1
    - One online tutorial will be recorded every week and made available on Canvas in “Recorded lectures”
  - Please attend your allocated tutorial – see the tutorial number in your timetable

Tutorial timetable

Tutorial	Time	Tutor
RE (remote = online)		
1 Zoom	Monday 20-21	Henry
2 Zoom	Monday 20-21	Stephen
3 Zoom	Monday 20-21	Nicholas
4 Zoom	Monday 20-21	Mashud
5 Zoom	Monday 20-21	Chen
6 Zoom	Monday 20-21	Gio
7 Zoom	Tuesday 17-18	Claire
8 Zoom	Tuesday 17-18	Nicholas
9 Zoom	Tuesday 17-18	Stephen
10 Zoom	Wednesday 17-18	Gio
11 Zoom	Wednesday 17-18	Claire
12 Zoom	Wednesday 17-18	Mashud
CC (on campus = face-to-face)		
1 face-to-face	Tuesday 17-18, CSC Lab 115	Mashud
2 face-to-face	Tuesday 17-18, CSC Lab 114	Chen
3 face-to-face	Tuesday 17-18, CSC Lab 130A	Thomas
4 face-to-face	Wednesday 17-18, SCS Lab 130A	Stephen
5 face-to-face	Wednesday 17-18, SCS Lab 118	Thomas

Face-to-face tutorials – keep attendance

- For the face-to-face tutorials we need to take attendance
- This will only be used for COVID tracing (required by the University), not for any other purpose
- We don't want to impose attendance as a class requirement – we believe in academic freedom – you are mature and responsible individuals and should attend your classes because they are beneficial for you, not because this is mandated!

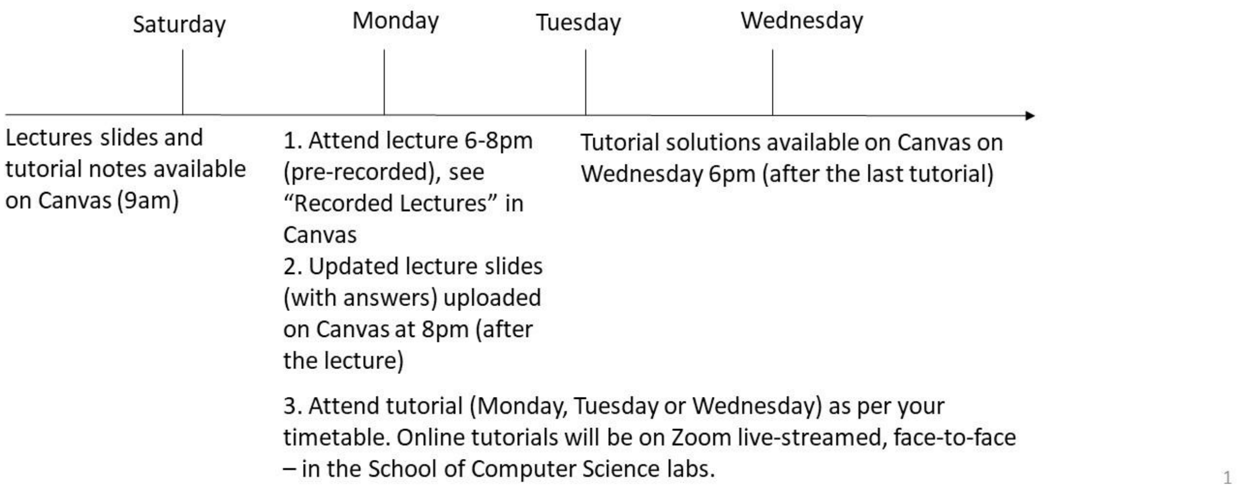
Course website

- 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 (Piazza)
  - assignment submission system (PASTA)

- Important document on Canvas: unit-outline-detailed.pdf – contains the most important information about this course

Course materials

- Lecture slides and tutorial notes will be available in advance on Saturday morning at 9am
- The lecture slides initially may not include the answers to questions and exercises that we will do at the lectures; the complete version with the answers will be uploaded after the lecture
- Tutorial solutions will be available on Wednesday evening after the last tutorial, which finishes at 6pm



Lectures slides and tutorial notes available on Canvas (9am)

1. Attend lecture 6-8pm (pre-recorded), see “Recorded Lectures” in Canvas
2. Updated lecture slides (with answers) uploaded on Canvas at 8pm (after the lecture)

Tutorial solutions available on Canvas on Wednesday 6pm (after the last tutorial)

3. Attend tutorial (Monday, Tuesday or Wednesday) as per your timetable. Online tutorials will be on Zoom live-streamed, face-to-face – in the School of Computer Science labs.

Discussion board Piazza

- We will use Piazza, it is linked to Canvas
  - Activate your Piazza account – I sent you an invitation email on Wednesday
  - If you have joined the course after Wednesday, you can enroll in Piazza yourself, from the Canvas site -> go to Piazza; if you have problems, email me
  - You must have access to Piazza, it will be the main communication channel for this course
  - When you click on Piazza in Canvas, you should be able to see the posts for this course, otherwise you are either not enrolled in Piazza or there is another problem
- Posting questions on Piazza
  - Post your question on Piazza 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)
- If you are shy, you can ask your question anonymously!

Automated marking system PASTA

- We will use PASTA for Assignment 1
- PASTA is an automated marking system – it will automatically test your code against test cases

- It will be linked to Canvas in due time
- It requires VPN access

- More information will be provided later

Lecture slides and recordings

- The lectures will be recorded and available on Canvas
- For Irena’s part – you may not need the recordings:
  - My lecture slides are very detailed, with many examples
  - I put everything important on the slides, including updating the slides after the lecture to add the solutions/answers whenever applicable
  - My slides are self-content and intended to help you revise and catch-up quickly

Assessment overview

- Three components:
  1. Assignment 1 – 15% (week 7)
  2. Assignment 2 – 25% (week 12)
  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; individual
    - Computer program only
    - Submitted via PASTA
    - Code auto-marked by PASTA for correctness

  - Assignment 2 (25%) - due Friday week 12; individual or in pairs (no more than 2 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 3 days late
  - Late penalty of 5% per day will apply
  - Assignments submitted more than 3 days late will not be accepted

- Assignment 1 - multiple submissions allowed
  - You can submit your code in PASTA as many times as you want (unlimited) before the deadline
  - PASTA will show you how many tests you have passes, which determines your mark
- 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. This means a minimum of 24 marks.
  - The exam will be online take-home, 2 hours duration
  - 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 complete OLEO1306 Beginner Programming for Data Analysis
  - It is a 0 credit points OLE course – it is free, can be enrolled and discontinued at any time during the semester
  - See the document on Canvas on how to enroll in OLE courses
- 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 Colaboratory – Google's Jupyter notebook environment

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

- In summary, there are 3 documents on Canvas related to the practical part of this course:
  - How to install Jupyter Notebook
  - How to enroll in OLEO1306 Beginner Programming for Data Analysis
  - 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 for both 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 (Assignment 2 report 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

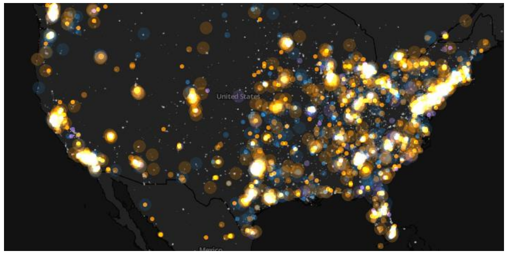
- Data explosion – society produces and stores huge amounts of data
  - Due to automated data collection tools and sensors, mature database technology, cheaper and more powerful computers
  - Sources: business, science, medicine, economics, environment, web, etc.
- Examples:
  - purchase data – supermarket, 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

amazon.com

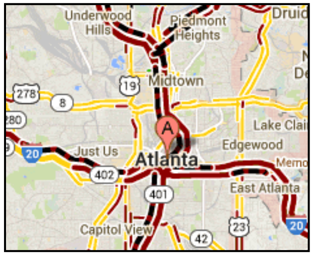
Google

facebook

E-Commerce



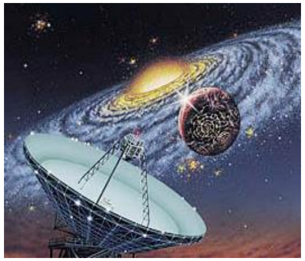
Social Networking: Twitter



Traffic

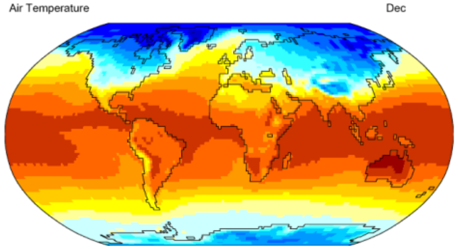
More examples

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

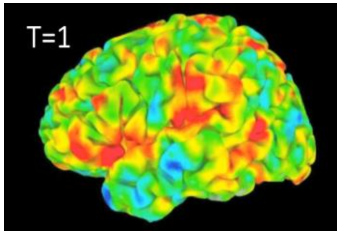


Sky survey data

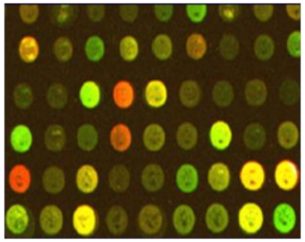
Sensor networks



Temperature



fMRI brain data



Gene expression data

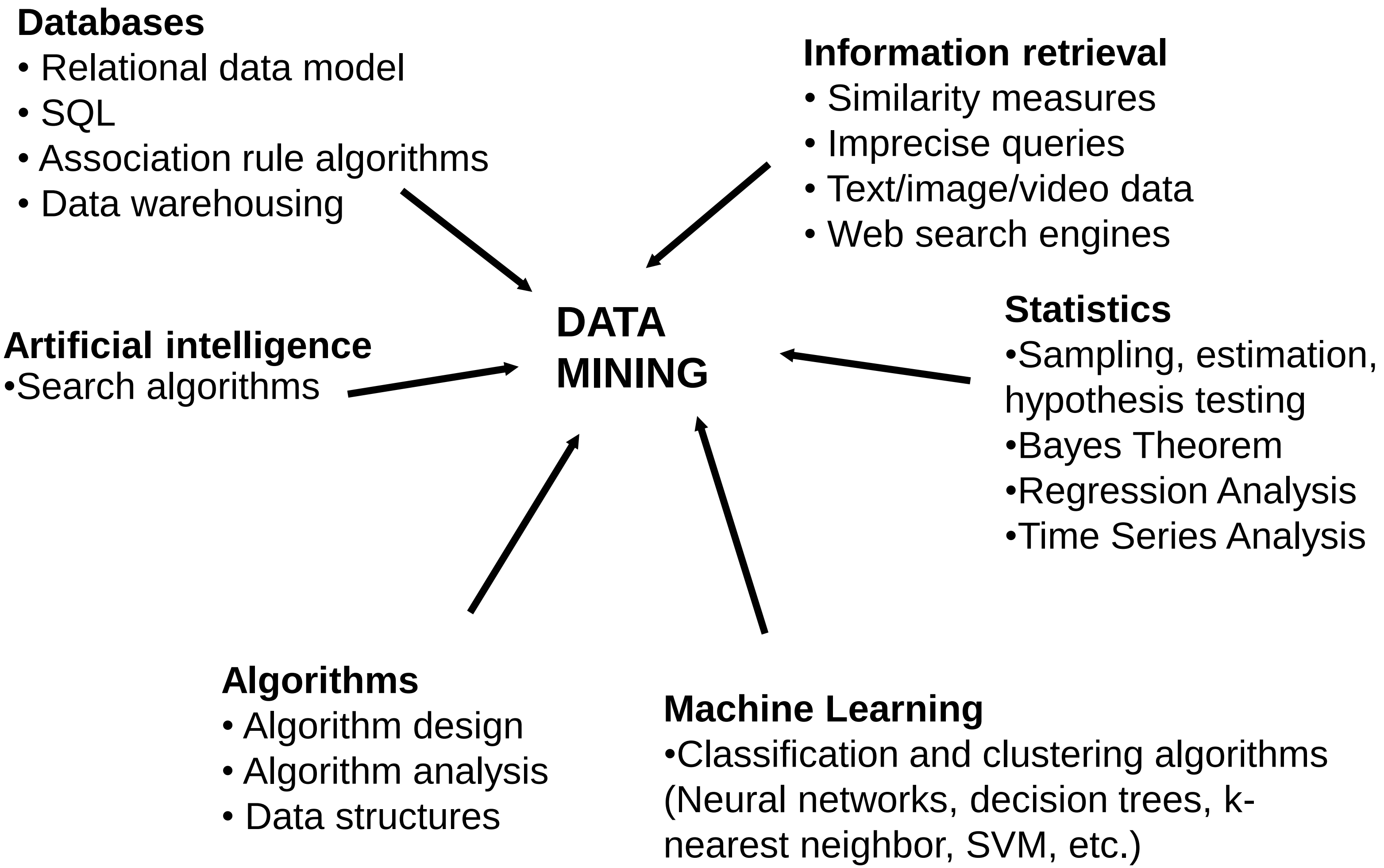
  - medical records and scans
  - biological data (high-throughput) – cytometry, gene expression

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.)

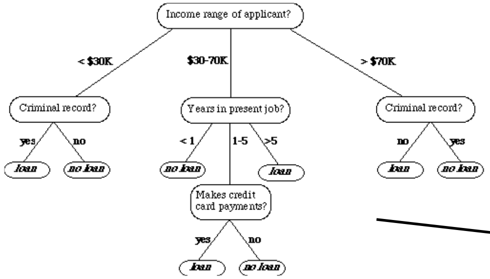
ML and DM tasks

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

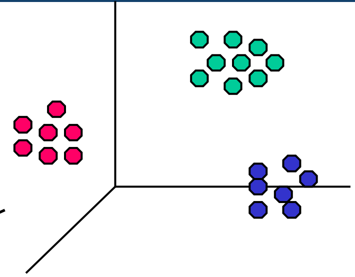
Main tasks - diagram



Supervised learning
(classification and regression)

Data

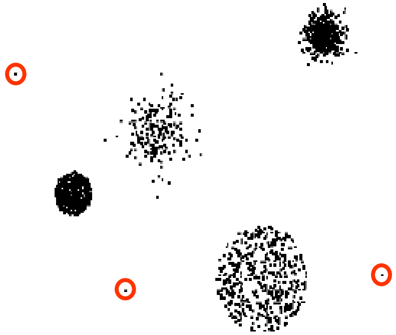
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



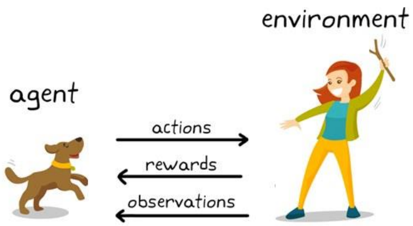
Unsupervised learning
(clustering)

Association rule mining

Anomaly detection



Reinforcement learning



Topics covered

Week	Date	Topic	Lecturer
1	1 March	Administrative matters and course overview. Introduction to machine learning and data mining. Data: cleaning, pre-processing and similarity measures.	Irena
2	8 March	Nearest neighbour. Rule-based algorithms.	Irena
3	15 March	Linear regression. Logistic regression. Overfitting and regularization.	Irena
4	22 March	Naïve Bayes. Evaluating machine learning methods. Assignment 1 out (Monday)	Irena
5	29 March	Decision trees. Ensembles.	Irena
		Mid-semester break	
6	12 April	Support vector machines. Kernels. Dimensionality reduction.	Irena
7	19 April	Neural networks - perceptrons and backpropagation algorithm. Assignment 1 due (Friday)	Chang
8	26 April	Deep neural networks: convolutional and recurrent.	Chang
9	3 May	Clustering I:Partitional, model-based and hierarchical. Assignment 2 out (Monday)	Chang
10	10 May	Clustering II: Density-based and grid-based. Evaluating clustering results.	Chang
11	17 May	Markov models – HMM, MEMM, CRF.	Chang
12	24 May	Reinforcement learning. Assignment 2 due (Friday)	Chang
13	31 May	Guest lecture. Revision.	Irena

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

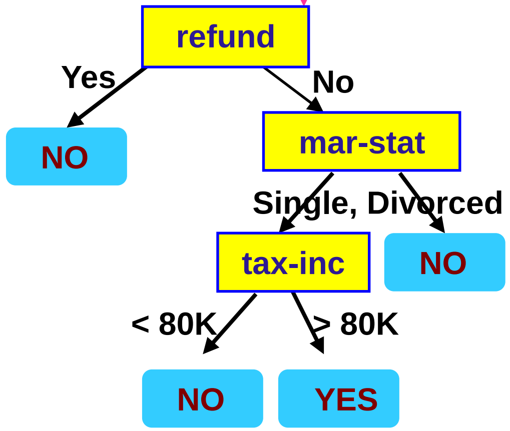
target class

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

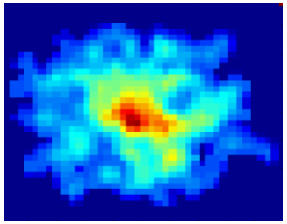
More classification examples (1)

- 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)

More classification examples (2)

- Ex. 3: Direct marketing – find a set of customers likely to buy a product
  - Given a user, is she/he likely to buy a product, e.g. a new mobile phone?
  - Data about
    - the user – phone usage, demographic and lifestyle
    - previous similar products – what are the characteristics of the customers who decided to buy and who didn’t – extract features
  - 2 classes {buy, don’t buy} – build a classifier

- Ex. 4: Sky survey cataloging
  - Task: Predict the class (star or galaxy) of sky objects
  - Data: images from an observatory
  - Dataset: 72 million stars, 20 million galaxies
  - From Fayyad et. al. *Advances in Knowledge Discovery and Data Mining*

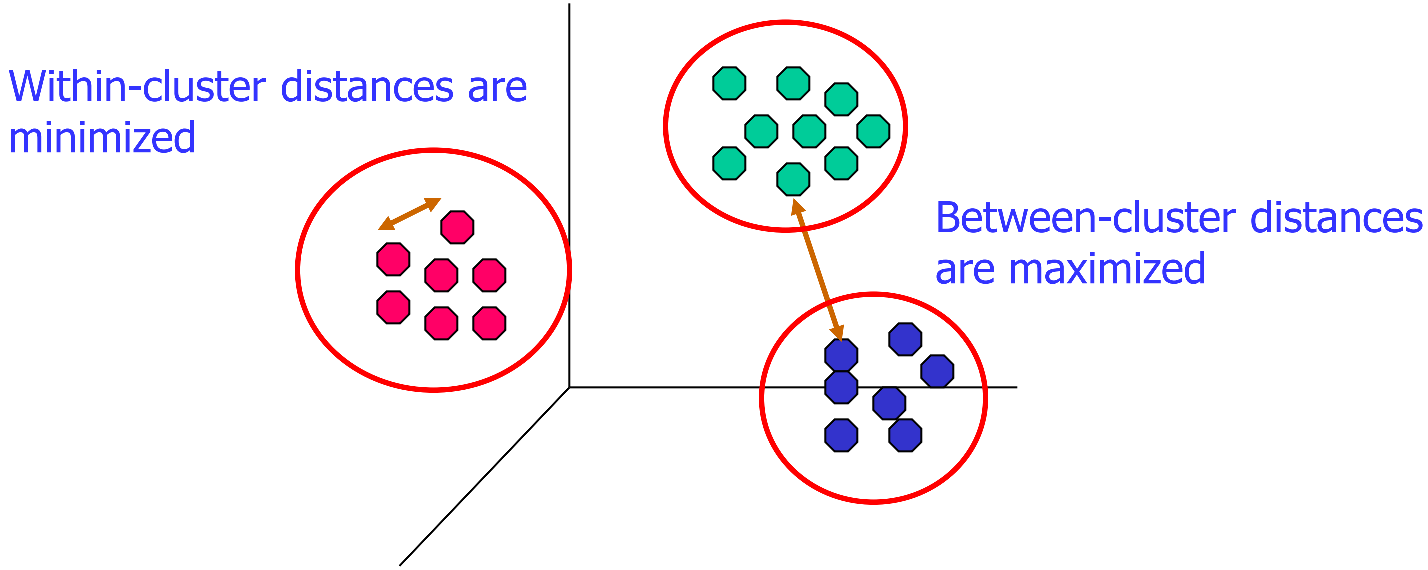


Regression - examples

- 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

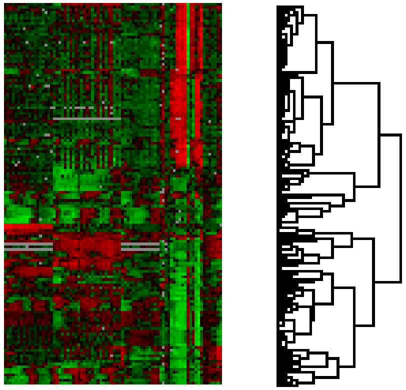
Unsupervised learning

- 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


Within-cluster distances are minimized
Between-cluster distances are maximized

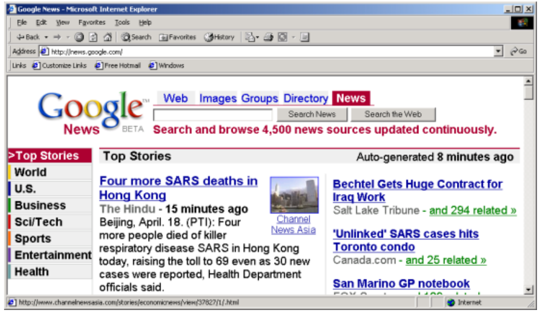
Clustering - examples

- 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


Google News - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://news.google.com/
Go
Links Customize Links Free Hotmail Windows
Google News BETA
Web Images Groups Directory News
Search News
Search the Web
>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Channel News Asia
http://www.channelnewsasia.com/stories/economicnews/view/37827/1/.html
Internet

Search and browse 4,500 news sources updated continuously.

Top Stories Auto-generated 8 minutes ago

Four more SARS deaths in Hong Kong
The Hindu - 15 minutes ago
Beijing, April. 18. (PTI): Four more people died of killer respiratory disease SARS in Hong Kong today, raising the toll to 69 even as 30 new cases were reported, Health Department officials said.

Bechtel Gets Huge Contract for Iraq Work
Salt Lake Tribune - and 294 related »

'Unlinked' SARS cases hits Toronto condo
Canada.com - and 25 related »

San Marino GP notebook

- 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

Association rule mining

- Find combinations of items that occur together
- Also called *market-basket analysis*
- Assumes transaction data

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Potato Chips, Milk**
4	**Beer, Bread, Potato Chips, Milk**
5	**Coke, Potato Chips, Milk**

Rules discovered:

{Milk} --> {Coke}

{Potato Chips, Milk} --> {Beer}

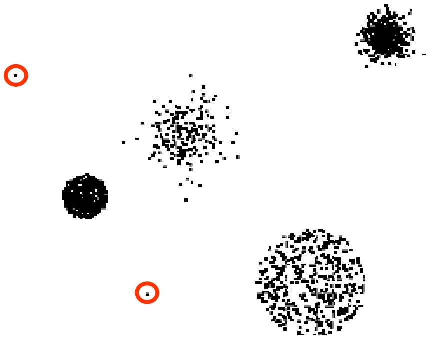
- Sequential version of association rule mining: find *frequent sequences* in data

Association rule mining - examples

- Business and marketing
  - Rules are used for sales promotion, shelf management and inventory management
  - E.g. sales promotion: services purchased together by telecommunication customers (e.g. broad band Internet, call forwarding, etc.) help determine how to bundle these services together to maximize revenue
- Telecommunication alarm diagnosis
  - Find combination of alarms that occur together frequently in the same time period
- Insurance
  - Unusual combinations of insurance claims can be a sign of a fraud
- Medical informatics
  - Find combination of patient symptoms and test results associated with certain diseases
  - Medical histories can give indications of complications based on combinations of treatments

Outlier detection

- Outliers are examples that are significantly different than the others (i.e. are far away from the others)
- In statistics an outlier is typically defined as an example that differs more than 3 standard deviations from the mean of all examples
- Detecting outliers is important for 2 reasons:
  - Outliers are noise and should be removed before data analysis
  - Outliers are the goal of our DM analysis – to detect unusual behavior, e.g. credit card fraud detection or intrusion detection



Outliers should be considered carefully

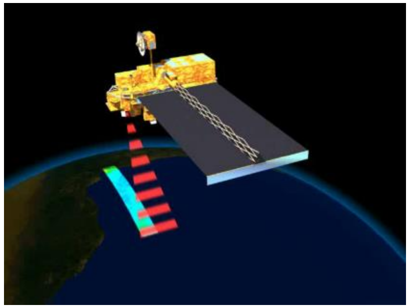
- An example where outliers were mistakenly removed

- Detecting the ozone hole in 1985:
  - Data collected by the British Antarctic Survey showed 10% drop in the ozone concentration for Antarctica
  - However, data collected by a satellite (Nimbus 7) did not show this drop

  - Why?
  - The satellite correctly recorded the ozone concentration but the values were so low that they were treated as outliers by the computer program and discarded!

Outlier detection - examples

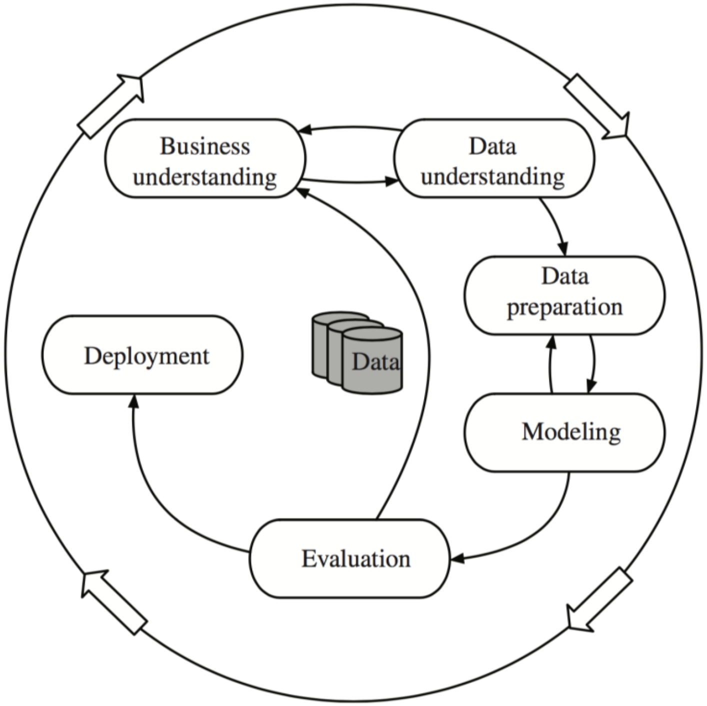
- The goal is to detect significant deviations from normal behavior
  - Fraud detection – e.g. deviation from typical behavior in credit card usage
  - Intrusion detection – monitoring computers and networks for unusual behavior
  - Hurricanes, floods, heat waves and fires prediction – atypical events with significant effect on humans
  - Health care – unusual symptoms or test results may indicate potential health problems and should be investigated
  - Detecting changes in the global forest cover



The DM process

1) Business understanding

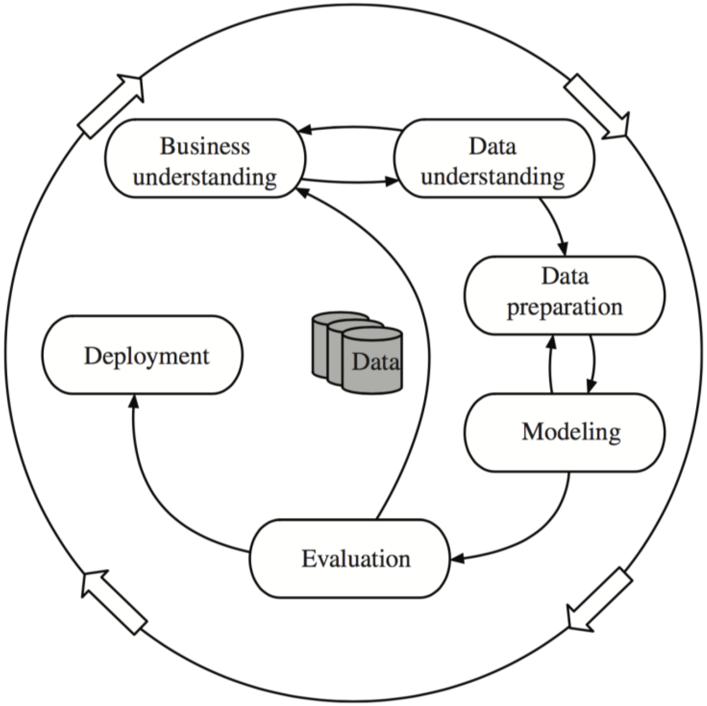
- Investigating the business objectives and requirements
- Deciding whether DM can be applied to meet them
- Determining what kind of data can be collected to build a deployable model



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

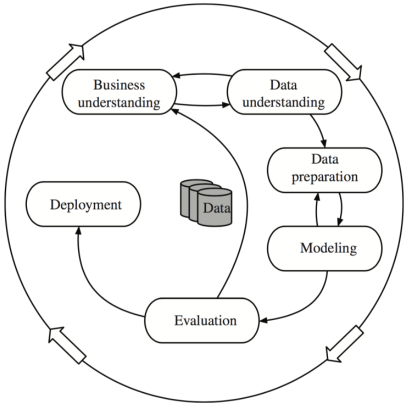
The DM process (4)

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


Business understanding
Data understanding
Data preparation
Deployment
Data
Modeling
Evaluation