F20/21DL Data Mining and Machine Learning

Diana Bental
Ekaterina Komendantskaya
(with material from David Corne and slides from http://www.cs.waikato.ac.nz/ml/weka/book.html)

- Diana Bental
  - Office hour in EM1.05 Monday 10.15-11.15am
  - d.s.bental@hw.ac.uk Extension 3367
- Ekaterina Komendantskaya
  - ek19@hw.ac.uk Extension 8283
- Lectures
  - Thursday 13.15 - 14.15 EM 1.83
  - Friday 12.15 – 13.15 EM 3.36
- Labs – from Week 2
  - 4th Year Thursday 10.15 – 11.15 EM G.45/EMG.46
  - 5th Year / MSc Thursday 15.15– 16.15 EM G.45/EMG.46

Mechanics

- We're using Vision
  - Course Information, Contacts
  - Learning Materials – lecture slides, links and websites
  - Assessment - coursework specifications and submission
- We're using F21DL on Vision. If you don't have F21DL 2018-2019 (Data Mining and Machine Learning) on your list of courses then email me and I will add you.
  - d.s.bental@hw.ac.uk
- It's a bit of a work in progress so please let us know if there are issues

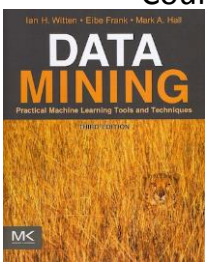
Module assessment

- 50% exam
  - Electronic exam
- Three main items of coursework
- CW 1: 15% CW 2: 15% CW 3: 20%
  - Group coursework
  - Divide yourselves into groups of 4 and sign up on Vision
  - CW1 Released next week
- Extra bit added to each c/w for MSc students
- Weeks 6-10 Weekly Class Exercises on Vision– no marks but you must complete them!

Coursework Software

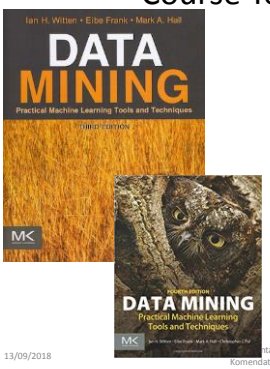
- Using WEKA 3
- Available on Linux
- http://www.cs.waikato.ac.nz/ml/weka/
- Online tutorials and MOOCS

Course Textbook



- **Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition** Ian H. Witten, Eibe Frank and Mark A Hall, Elsevier 2011
- Sections on WEKA 3

Course Textbook



- **Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition** Ian H. Witten, Eibe Frank and Mark A Hall, Elsevier 2011
- Sections on WEKA 3
- New 4th Edition
  - "Deep learning"

- Part 1
- Introduction & Terminology
- Input Preparation
- Output - Knowledge Representations
- Algorithms and Basic Statistics
- Evaluation
- Coursework 1

- Part 2
- 4 learning approaches
  - Probability: Bayesian Networks
  - Unsupervised Learning: Clustering
  - Supervised Linear Learning: Decision Trees
  - Supervised Non-Linear Learning: Neural Networks
- Courseworks 2 & 3

This lecture

- What can we do with data?
- Where does data come from?
- What should we do with data?

Data Mining and Machine Learning– Why are they important?

- Data are being generated in enormous quantities
- Data are being collected over long periods of time
- Data are being kept for long periods of time
- Computing power is formidable and cheap
- A variety of Data Mining and Machine Learning software is available
- All of these data contain `hidden knowledge' – facts, rules, patterns, that can be usefully exploited ….. if we can find them

What can we do with data?

- Answer simple questions like:
- How many female clients do we have?
- How much paint did we sell in 2007?
- Which is the most profitable branch of our supermarket?
- Which postcodes suffered the most dropped calls in July?

We can, but …

More interesting things that can be done with data

- Answer difficult and valuable questions like:
  - How can we predict ovarian cancer early enough to treat it successfully?
  - How can I make significant profit on the stock market next month?
  - Two different authors claim to have written this story – how can we resolve the dispute?
  - How can we get our customers to spend more money in the store?
  - Is this loan applicant a good credit risk?
  - Is this sonar image a mine or a rock?
  - What other websites will this customer be interested in?

Data Mining - Definition & Goal

- Definition
  - Data Mining is the exploration and analysis of (often) large quantities of data in order to discover meaningful patterns and rules
- Goal
  - To permit some other goal to be achieved or performance to be improved through a better understanding of the data

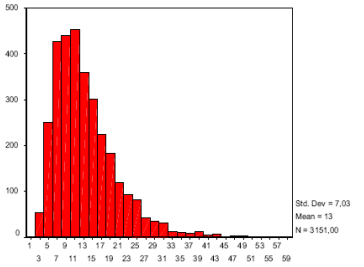
Some examples of large databases

- Shopping basket data: much commercial DM is done with this. In one store, 18,000 baskets per month
- Tesco has >500 stores. Per year, 100,000,000 baskets ?
- The Internet
  - ~ 4.7 billion (searchable) pages
  - 305 billion printed pages
- Lots of datasets: UCI Machine Learning repository

- How can we begin to understand and exploit such datasets? Especially the big ones?

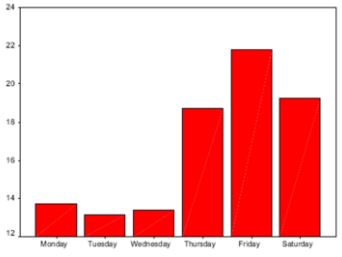
Like this …

Figure 1: Average number of distinct items purchased per visit



and this…

Figure 4: Distribution of visits per day of the week



Data Mining & Machine Learning

- Data Mining is the process of discovering patterns and inferring associations in raw data
- … a collection of techniques intended to analyse small or large amounts of data
- … can employ a range of techniques, either individually or in combination with each other
- Is Machine Learning the same?

Philosophy – Can machines "learn"?

- **Dictionary definitions - *knowledge***
  - "To get knowledge of by study, experience, or being taught"
  - "To become aware, by information or from observation"
    - How would we know if a computer had learnt?
  - "To commit to memory"
  - "To be informed of, ascertain; to receive instruction"
    - Trivial for computers
- **Operational definition - *performance***
  - Things learn when they change their behaviour in a way that makes them perform better in the future.
- **Do *intention* and *purpose* matter? Or *reflection*?**

Data Mining and Machine Learning

- Given some data, we want to extract information that is
  - Implicit
  - Previously unknown
  - Potentially useful (and non-trivial)
- We need programs to extract patterns and regularities from the data
- Strong patterns lead to good predictions
- But
  - Most patterns are not interesting
  - Patterns may be inexact (or spurious)
  - Data may be garbled or missing

Some applications in the field

- The result of learning—or the learning method itself—is deployed in practical applications….

1. Processing loan applications (American Express)

- Given:
  - questionnaires with financial and personal information
- Question:
  - should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
  - Solution: reject all borderline cases?
  - No! Borderline cases are most active customers

Enter machine learning….

- 1000 training examples of borderline cases
- 20 attributes:
  - age
  - years with current employer
  - years at current address
  - years with the bank
  - other credit cards possessed,…
- Learned rules: correct on 70% of cases
- human experts only 50%
- Rules could be used to explain decisions to customers

2. Screening images

- Given: radar satellite images of coastal waters
  - Problem: detect oil slicks in those images
  - Oil slicks appear as dark regions with changing size and shape
- Not easy: lookalike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel

Enter machine learning…

- Extract the dark regions from normalized images
- Attributes:
  - size of region
  - shape, area
  - intensity
  - sharpness and jaggedness of boundaries
  - proximity of other regions
  - info about background
- Constraints:
  - Few training examples—oil slicks are rare!
  - Unbalanced data: most dark regions aren't slicks
  - Requirement: adjustable false-alarm rate

3. Marketing and sales

- Companies precisely record massive amounts of marketing and sales data
- Applications:
  - Customer loyalty:
    identifying customers who are likely to defect by detecting changes in their behavior
    (e.g. banks/phone companies)
  - Special offers:
    identifying profitable customers
    (e.g. reliable owners of credit cards who need extra money during the holiday season)

More marketing and sales

- Market basket analysis
- *Association learning* techniques
  - Find groups of items that tend to occur together in a transaction
  - Analyze checkout data
- Historical analysis of purchasing patterns
- Identify prospective customers
- Focus promotional mailings
  - targeted campaigns are cheaper than mass-marketed ones

…. and many more

- The result of learning—or the learning method itself—is deployed in practical applications
  - Electricity supply forecasting
  - Diagnosis of machine faults
  - Separating crude oil and natural gas
  - Reducing banding in rotogravure printing
  - Finding appropriate technicians for telephone faults
  - Scientific applications: biology, astronomy, chemistry
  - Monitoring intensive care patients
  - Web mining
  - Recommender systems (Automatic selection of TV programs, online sales, films, music)
  - Etc.
- Usually care about *performance* (prediction) but *structure* matters too

(Some) Ethics of Data Mining

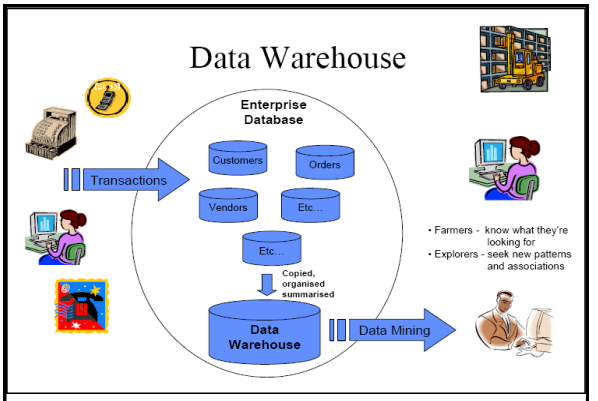
- Ethical issues arise in practical applications
- Personal Information
  - anonymizing data is difficult
  - 85% of Americans can be identified from just zip code, birth date and sex
- Discrimination
  - e.g. loan applications: using some information (e.g. sex, religion, race) is unethical
  - Ethical situation depends on the application
  - e.g. a medical applications may need the same information
- Attributes may contain problematic information
  - E.g. *area code* may correlate with *race*

Ethics of Data Mining

- Important questions:
  - Who is allowed access to the data?
    - Social norms – a library will not tell you who has a book
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?
- Learning from existing data
  - Are we just reinforcing old behaviour patterns?
- Caveats must be attached to results
- Purely statistical arguments are never sufficient!
- Are results put to good use?
  - Make it easier and quicker for shoppers, or more profitable for the shop?

Garbage in, garbage out….

- Historic example:
  - 1980s, "expert system" for a local police force
    - Recognise the "modus operandi" of burglars
    - Identify the criminals
    - Analyse existing criminal records
  - But:
    - Important attributes were missing
      - » *Broken* glass - common
      - » *Cut* glass - rare
      - » Records only stored
        - *Glass – broken / cut*
    - Class values were inaccurate
      - "*Taken into consideration*"



Data Warehousing

- Hopefully – a source of "clean" data
- "A subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process"
  - W. H. Inmon, "What is a Data Warehouse?" Prism Tech Topic, Vol. 1, No. 1, 1995 -- a very influential definition.
- "A copy of transaction data, specifically structured for query and analysis"
  - Ralph Kimball, from his 2000 book, "The Data Warehouse Toolkit"

Take away

- What can we do with data?
  - Many interesting and useful things
- Real data is often not in the form we want it
- What should we do with data?
  - Learn interesting and useful things
  - Respect privacy and ethics
  - Look critically at the "truth" of what we have learnt
  - Use good quality data

Friday

- Basics
- Examples (you can investigate in Weka)