UNIVERSITÉ DE NEUCHÂTEL

MASTER IN COMPUTER SCIENCE

Course 32032: Machine Learning and Data Mining

Chapter 1

Introduction to Data Mining

Prof. Marcelo Pasin

UNINE / University of Neuchatel
HES-SO / University of Applied Sciences and Arts Western Switzerland

Fall 2020

Contents

Chapter 1. What's it all about?

Contents

Information

- Example 1: *in vitro* fertilization
  - Given: embryos described by 60 features
  - Problem: selection of embryos that will survive
  - Data: historical records of embryos and outcome

- Example 2: cow culling
  - Given: cows described by 700 features
  - Problem: selection of cows that should be culled
  - Data: historical records and farmers' decisions

Transform data in information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, …
- This data is a potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: patterns underlying the data
- We are concerned with machine learning techniques for automatically finding patterns in data
- Patterns that are found may be represented as *structural descriptions* or as black-box models

Structural descriptions

Example: if-then rules [Table 1.1]

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
…	…	…	…	…

If tear production rate = reduced
    then recommendation = none

Otherwise, if age = young and astigmatic = no
    then recommendation = soft

Machine learning

- Definition of "learning"
  - To get knowledge of by study, experience, or being taught
  - To become aware by information or from observation
  - To commit to memory
  - To be informed of, ascertain; to receive instruction

Difficult to measure

Trivial for computers

- Operational definition
  - Things learn when they change their behaviour in a way that makes them perform better in the future
- Does learning imply intention?

Data mining

- Finding patterns in data that provide insight or enable fast and accurate decision making
- Strong, accurate patterns are needed to make decisions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing
- Machine learning techniques identify patterns in data and provide many tools for data mining
- Of primary interest are machine learning techniques that provide structural descriptions

Contents

The weather problem

Conditions for playing a certain game [Table 1.2]

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	Normal	False	Yes
…	…	…	…	…

If outlook = sunny and humidity = high	then play = no
If outlook = rainy and windy = true	then play = no
If outlook = overcast	then play = yes
If humidity = normal	then play = yes
If none of the above	then play = yes

Classification x association

Classification rule: predicts value of a given attribute
(the classification of an example)

If outlook = sunny and humidity = high
then play = no

Association rule: predicts value of arbitrary attribute (or combination)

If temperature = cool then humidity = normal

If humidity = normal and windy = false
then play = yes

If outlook = sunny and play = no
then humidity = high

If windy = false and play = no
then outlook = sunny and humidity = high

Mixed-attribute problem

Numeric attributes (use inequalities)

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
…	…	…	…	…

If outlook = sunny and humidity > 83	then play = no
If outlook = rainy and windy = true	then play = no
If outlook = overcast	then play = yes
If humidity < 85	then play = yes
If none of the above	then play = yes

Complete rule set for contact lens data

[Table 1.1]

If tear production rate = reduced then recommendation = none

If age = young and astigmatic = no and tear production rate = normal
then recommendation = soft

If age = pre-presbyopic and astigmatic = no and tear production rate = normal
then recommendation = soft

If age = presbyopic and spectacle prescription = myope and astigmatic = no
then recommendation = none

If spectacle prescription = hypermetrope and astigmatic = no and tear production rate = normal
then recommendation = soft

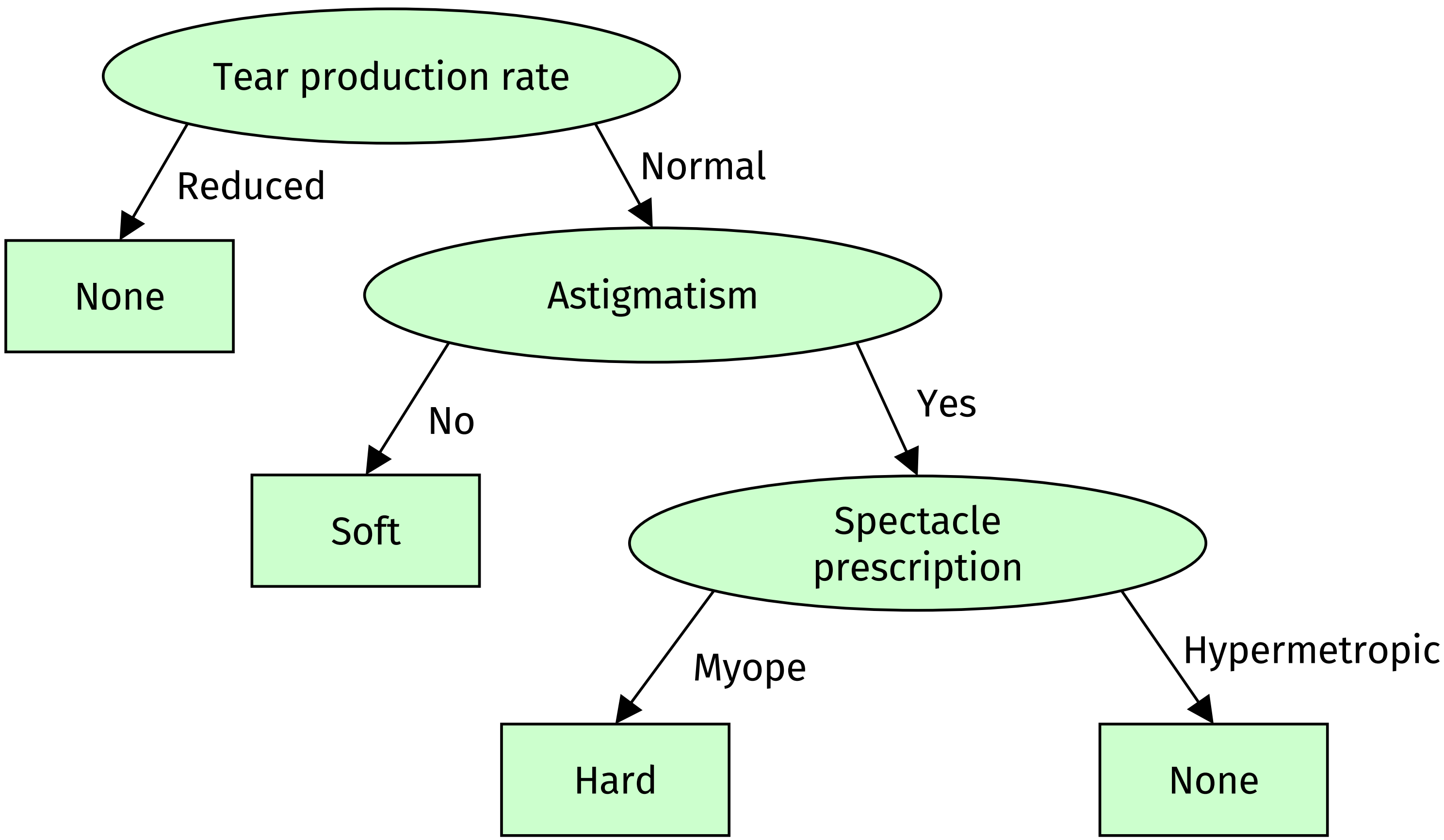
If spectacle prescription = myope and astigmatic = yes and tear production rate = normal
then recommendation = hard

If age young and astigmatic = yes and tear production rate = normal
then recommendation = hard

If age = pre-presbyopic and spectacle prescription = hypermetrope and astigmatic = yes
then recommendation = none

If age = presbyopic and spectacle prescription = hypermetrope and astigmatic = yes
then recommendation = none

Decision tree for contact lens data



Iris flower classification

Seminal work from Fisher, mid-1930s [Table 1.4]

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Iris flower rules

If petal-length < 2.45 then Iris-setosa
If sepal-width < 2.10 then Iris-versicolor
If sepal-width < 2.45 and petal-length < 4.55 then Iris-versicolor
If sepal-width < 2.95 and petal-width < 1.35 then Iris-versicolor
If petal-length ≥ 2.45 and petal-length < 4.45 then Iris-versicolor
If sepal-length ≥ 5.85 and petal-length < 4.75 then Iris-versicolor
If sepal-width < 2.55 and petal-length < 4.95 and petal-width > 1.55 then Iris-versicolor
If petal-length ≥ 2.45 and petal-length < 4.95 and petal-width < 1.55 then Iris-versicolor
If sepal-length ≥ 6.55 and petal-length < 5.05 then Iris-versicolor
If sepal-width < 2.75 and petal-width < 1.65 and sepal-length < 6.05 then Iris-versicolor
If sepal-length ≥ 5.85 and sepal-length < 5.95 and petal-length < 4.85 then Iris-versicolor
If petal-length ≥ 5.15 then Iris-virginica
If petal-width ≥ 1.85 then Iris-virginica
If petal-width ≥ 1.75 and sepal-width < 3.05 then Iris-virginica
If petal-length ≥ 4.95 and petal-width < 1.55 then Iris-virginica

Cumbersome rules, need something more compact

Predicting CPU performance

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

Linear regression

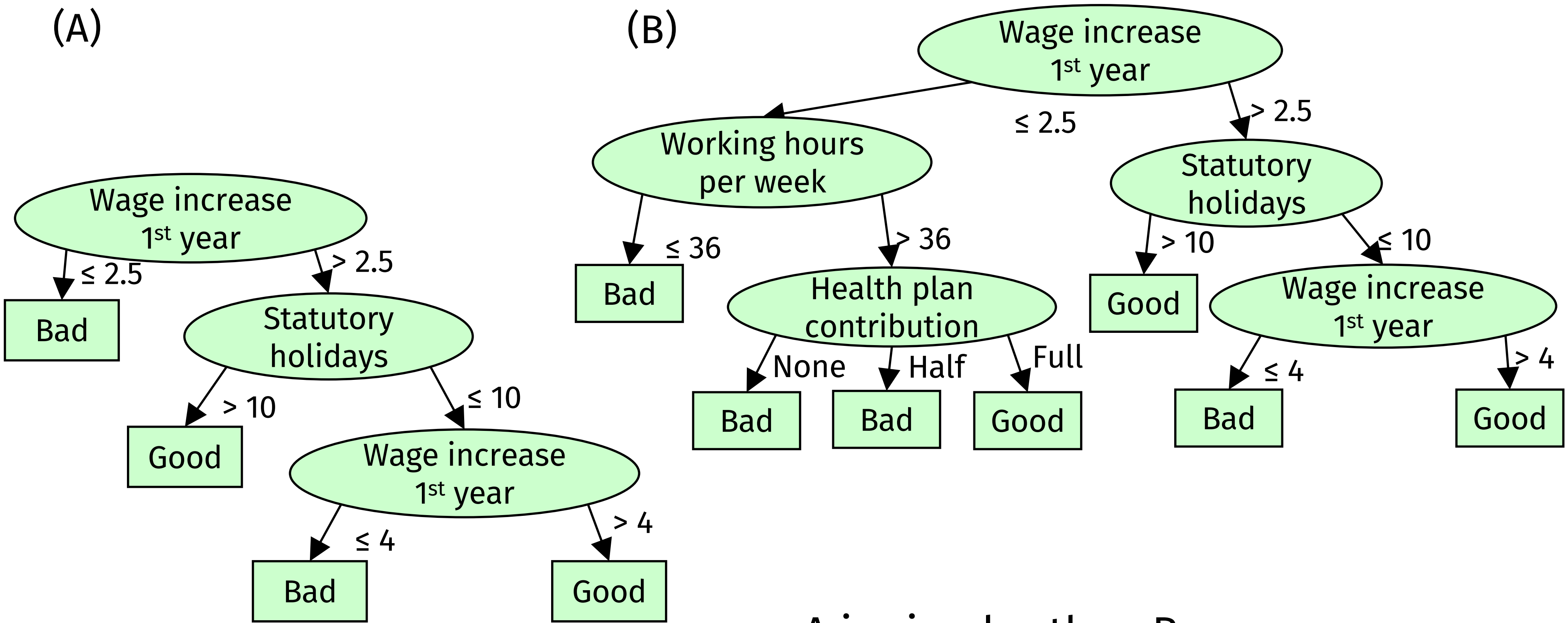
PRP = - 55.9 + 0.0489 MYCT + 0.0153 MMIN
+ 0.0056 MMAX + 0.6410 CACH - 0.2700 CHMIN
+ 1.480 CHMAX

Data from labour negotiations

Attribute	Type	1	2	3	...	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third year	Percentage	?	?	?		?
Cost of living adjustment	{none,tcf,tc}	none	tcf	?		none
Working hours per week	(Number of hours)	28	35	38		40
Pension	{none,ret-allw, empl-cntr}	none	?	?		?
Standby pay	Percentage	?	13%	?		?
Shift-work supplement	Percentage	?	5%	4%		4
Education allowance	{yes,no}	yes	?	?		?
Statutory holidays	(Number of days)	11	15	12		12
Vacation	{below-avg,avg,gen}	avg	gen	gen		avg
Long-term disability assistance	{yes,no}	no	?	?		yes
Dental plan contribution	{none,half,full}	none	?	full		full
Bereavement assistance	{yes,no}	no	?	?		yes
Health plan contribution	{none,half,full}	none	?	full		half
Acceptability of contract	{good,bad}	bad	good	good		good

- Realistic dataset, probably impossible to get exact classification

Decision trees for labour data



- A is simpler than B
- Overfitting

Soybean diseases

	Attribute	Number of values	Sample value
Environment	Time of occurrence	7	July
	Precipitation	3	Above normal
…			
Seed	Condition	2	Normal
	Mold growth	2	Absent
…			
Fruit	Condition of fruit pods	4	Normal
	Fruit spots	5	?
Leaf	Condition	2	Abnormal
	Leaf spot size	3	?
…			
Stem	Condition	2	Abnormal
	Stem lodging	2	Yes
…			
Root	Condition	3	Normal
Diagnosis		19	Diaporthe stem canker

Soybean classification success story

- Questionary with ~680 diseased plants
  - 35 attributes (small sets of values)
- Labelled with the diagnosis of an expert
  - 19 disease categories altogether
- Selected 300 training examples
  - "far apart" in example space
- Performed better that expert (97% x 72%)

Role of domain knowledge

If leaf condition is normal
and stem condition is abnormal
and stem cankers is below soil line
and canker lesion color is brown
then
diagnosis is rhizoctonia root rot

If leaf malformation is absent
and stem condition is abnormal
and stem cankers is below soil line
and canker lesion color is brown
then
diagnosis is rhizoctonia root rot

"leaf condition is normal" implies
"leaf malformation is absent"!


15
min

Contents

Fielded applications

- Previous examples: toy examples
- Real learning systems used to gain knowledge
  - Web mining
  - Loan judgement
  - Screening images
  - Power demand forecast
  - Machine diagnosis
  - Marketing and sales
- Comprehensible decision structure:
  key feature of success

Web mining

- World Wide Web is a huge application area
  - Find out best pages for a given subject
  - Find out best pages for a given user
  - Offer proper advertisements for any given user
- There is a huge commercial interest making money by mining the Web

Machine learning for Web mining

- Page Rank
  - Page prestige measured by the pages pointing to it
- Use human judgement of web pages
  - Apply learnt patterns to new pages and infer judgement
- Mine the user queries and behaviour as well
  - Terms often searched are more important
  - Items clicked are more important
- Social networks
  - Allow for associating users in groups

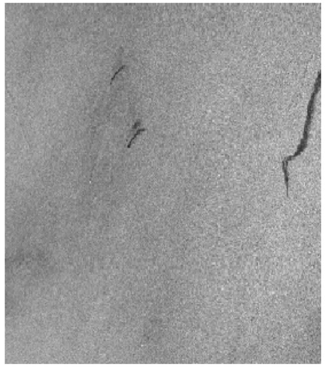
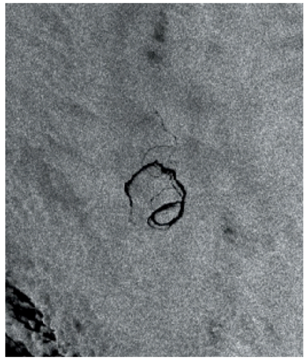
Process loan applications

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

Machine learning to Process loan applications

- 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, etc
- Learned rules
  - Correct on 70% of cases
  - Human experts only 50%
  - Could be used to explain decisions to customers

Screening images


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

Machine learning for Screening images

- Extract dark regions from normalized image
- 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
  - Regions from same image form a batch
  - Requirement: adjustable false-alarm rate
- Use as a filter (user support)

Power load forecasting

- Electricity supply companies need forecast of future demand for power
  - Forecasts of min/max load for each hour allow for significant savings
- Given: manually constructed load model that assumes “normal” climatic conditions
- Problem: adjust for weather conditions
- Static model consist of
  - Base load for the year
  - Load periodicity over the year
  - Effect of holidays

Machine learning for Power load forecasting

- Prediction corrected using “most similar” days
- Attributes
  - Temperature, humidity, wind speed
  - Cloud cover readings
  - Difference between actual load and predicted load
- Average difference among three “most similar” days added to static model
- Linear regression coefficients form attribute weights in similarity function
- Same performance as trained human but **faster**

Machine fault diagnosis

- Diagnosis: classical domain of expert systems
- Given
  - Fourier analysis of vibrations
  - Measured at various points of a device's mounting
- Question: Which fault is present?
  - Used for preventative maintenance of electromechanical motors and generators
  - Information very noisy
  - Before: diagnosis by expert/hand-crafted rules

Machine learning for Machine fault diagnosis

- Available: 600 faults with expert's diagnosis
  - ~300 unsatisfactory, rest used for training
  - Attributes augmented by intermediate concepts that embodied causal domain knowledge
- Learned rules outperformed hand-crafted ones
  - Expert not satisfied with initial rules because they did not relate to his domain knowledge
  - Further background knowledge resulted in more complex rules that were satisfactory

Marketing and sales

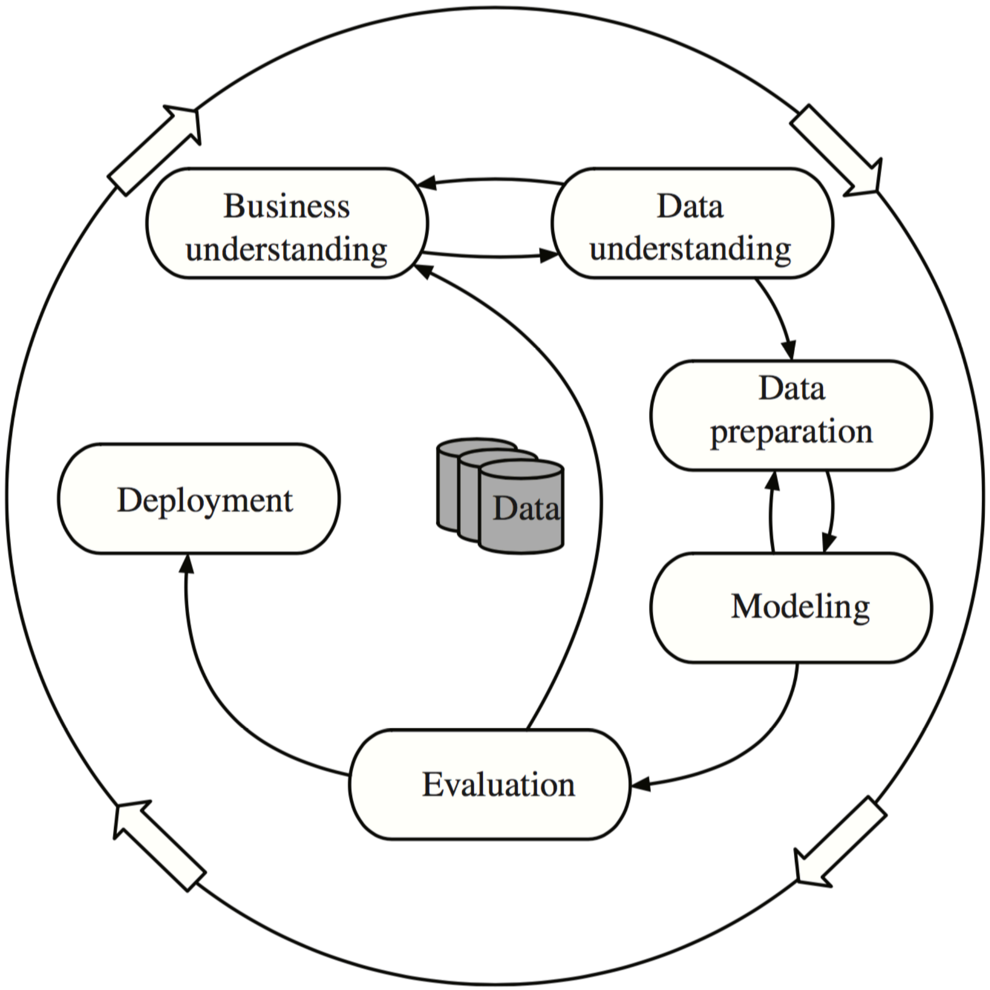
- Companies precisely record massive amounts of marketing and sales data
- Applications
  - Customer loyalty
    - Identifying customers that are likely to defect
    - Detect changes in their behaviour (e.g. banks/phone companies)
- Special offers
  - Identifying profitable customers
    - Ex.: reliable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

Machine learning for Marketing and sales

- Market basket analysis
- Association techniques
  - Find groups of items that tend to occur together in a transaction
  - Used to analyse checkout data
- Historical analysis of purchasing patterns
  - Identifying prospective customers
  - Focusing promotional mailouts (targeted campaigns are cheaper than mass-market)
  - Example: Thursdays, customers often purchase diapers and beer together (young parents stock up for a weekend)
  - Planning store layouts, limiting discounts to one of a set, coupons for a matching product

Contents

Tha data mining process



Machine learning and statistics

- Historical difference (grossly oversimplified):
  - Statistics: testing hypotheses
  - Machine learning: finding the right hypothesis
- Huge overlap
  - Decision trees (C4.5 and CART)
  - Nearest-neighbour methods
- Today: perspectives have converged
  - Most machine learning algorithms employ statistical techniques

Contents

Generalization as search

- Inductive learning
- Find a concept description that fits the data
- Example: rule sets as description language
  - Enormous, but finite, search space
- Simple solution:
  - Enumerate the concept space
  - Eliminate descriptions that do not fit examples
  - Surviving descriptions contain target concept

Enumerating the concept space

- Search space for weather problem
  - 4 x 4 x 3 x 3 x 2 = 288 possible combinations
  - With 14 rules: 2.7 x 10^{34} possible rule sets
- Other practical problems
  - More than one description may survive
  - No description may survive
    - Language is unable to describe target concept
    - Data may contain noise
- Another view of generalization as search
  - Hill-climbing in description space according to pre-specified matching criterion
  - Many practical algorithms use heuristic search that cannot guarantee to find the optimum solution

Bias

- Important decisions in learning systems:
  - Concept description language
  - Order in which the space is searched
  - Way that overfitting to the particular training data is avoided
- These form the "bias" of the search:
  - Language bias
  - Search bias
  - Overfitting-avoidance bias

Language bias

- Important question
  - Is language universal
    or does it restrict what can be learned?
- Universal language can express arbitrary subsets of examples
- If language includes logical *or* ("disjunction"), it is universal
  - Example: rule sets
- Domain knowledge can be used to exclude some concept descriptions *a priori* from the search

Search bias

- Search heuristic
  - "Greedy" search: performing the best single step
  - "Beam search": keeping several alternatives
- Direction of search
  - *General-to-specific*
    - Ex.: specializing a rule by adding conditions
  - *Specific-to-general*
    - Ex.: generalizing an individual instance into a rule

Overfitting avoidance bias

- It can be seen as a form of search bias
- Modified evaluation criterion
  - Ex.: balancing simplicity and number of errors
- Modified search strategy
  - E.g., pruning (simplifying a description)
    - Pre-pruning: stops at a simple description before search proceeds to an overly complex one
    - Post-pruning: generates a complex description first and simplifies it afterwards

Contents

Data mining and ethics

- Ethical issues arise in practical applications
- Anonymizing data is difficult
  - 85% of Americans can be identified from just zip code, birth date and sex
- Data mining often used to discriminate
  - Ex.: loan applications using some information (e.g., sex, religion, race) is unethical
- Ethical situation depends on application
  - Ex.: same information ok in medical application
- Attributes may contain problematic information
  - Ex.: area code may correlate with race

Ethics wider issues

- Important questions
  - Who is permitted access to the data?
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?

- Caveats must be attached to results
- Purely statistical arguments are never sufficient
- Are resources put to good use?