What’s it all about?

- Data vs information
- Data mining and machine learning
- Structural descriptions
  - Rules: classification and association
  - Decision trees
- Datasets
  - Weather, contact lens, CPU performance, labor negotiation data, soybean classification
- Fielded applications
  - Ranking web pages, loan applications, screening images, load forecasting, machine fault diagnosis, market basket analysis
- Generalization as search
- Data mining and ethics

Data vs. information

- Society produces huge amounts of data
  - Sources: business, science, medicine, economics, geography, environment, sports, ...
- Potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
  - Data: recorded facts
  - Information: patterns underlying the data

Information is crucial

- 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
Data mining

- Extracting
  - implicit,
  - previously unknown,
  - potentially useful
  information from data
- Needed: programs that detect patterns and regularities in the data
- Strong patterns ⇒ good predictions
  - 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

- Algorithms for acquiring structural descriptions from examples
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived
    (may be even more important)
- Methods originate from artificial intelligence, statistics, and research on databases

Structural descriptions

- Example: if-then rules

<table>
<thead>
<tr>
<th>Age</th>
<th>Spectacle prescription</th>
<th>Astigmatism</th>
<th>Tear production rate</th>
<th>Recommended lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>Myope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Myope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Can machines really learn?

- Definitions of “learning” from dictionary:
  - 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

  }  Trivial for computers
  }  Difficult to measure

- Operational definition:
  - Things learn when they change their behavior in a way that makes them perform better in the future.

  }  Does a slipper learn?

- Does learning imply intention?
The weather problem

- Conditions for playing a certain game

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>Mild</td>
<td>Normal</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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 vs. association rules

- 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

Weather data with mixed attributes

- Some attributes have numeric values

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>85</td>
<td>85</td>
<td>False</td>
<td>No</td>
</tr>
<tr>
<td>Sunny</td>
<td>80</td>
<td>90</td>
<td>True</td>
<td>No</td>
</tr>
<tr>
<td>Overcast</td>
<td>83</td>
<td>86</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>Rainy</td>
<td>75</td>
<td>80</td>
<td>False</td>
<td>Yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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
### The contact lenses data

<table>
<thead>
<tr>
<th>Age</th>
<th>Spectacle prescription</th>
<th>Astigmatism</th>
<th>Tear production rate</th>
<th>Recommended lenses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>Myope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Myope</td>
<td>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Young</td>
<td>Myope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Myope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Myope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Myope</td>
<td>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Myope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Pre-presbyopic</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Myope</td>
<td>No</td>
<td>Normal</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Myope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Hypermetrope</td>
<td>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Normal</td>
<td>None</td>
</tr>
</tbody>
</table>

Data Mining: Practical Machine Learning Tools and Techniques (Chapter 1) 13

### A complete and correct rule set

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

### A decision tree for this problem

Data Mining: Practical Machine Learning Tools and Techniques (Chapter 1) 14

### Classifying iris flowers

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>7.0</td>
<td>3.2</td>
<td>4.7</td>
<td>1.4</td>
</tr>
<tr>
<td>52</td>
<td>6.4</td>
<td>3.2</td>
<td>4.5</td>
<td>1.5</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>6.3</td>
<td>3.3</td>
<td>6.0</td>
<td>2.5</td>
</tr>
<tr>
<td>102</td>
<td>5.8</td>
<td>2.7</td>
<td>5.1</td>
<td>1.9</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- If petal length < 2.45 then Iris setosa
- If sepal width < 2.10 then Iris versicolor
...
Predicting CPU performance

- Example: 209 different computer configurations

<table>
<thead>
<tr>
<th>Cycle time (ns)</th>
<th>Main memory (Mb)</th>
<th>Cache (Kb)</th>
<th>Channels</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCT</td>
<td>MMIN</td>
<td>MMAX</td>
<td>CACH</td>
<td>CHMIN</td>
</tr>
<tr>
<td>1</td>
<td>125</td>
<td>256</td>
<td>6000</td>
<td>256</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>8000</td>
<td>32000</td>
<td>32</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>208</td>
<td>480</td>
<td>512</td>
<td>8000</td>
<td>32</td>
</tr>
<tr>
<td>209</td>
<td>480</td>
<td>1000</td>
<td>4000</td>
<td>0</td>
</tr>
</tbody>
</table>

- Linear regression function

$$\text{PRP} = -55.9 + 0.0489 \text{MYCT} + 0.0153 \text{MMIN} + 0.0056 \text{MMax} + 0.6410 \text{CACH} - 0.2700 \text{CHMIN} + 1.480 \text{CHMAX}$$

Data from labor negotiations

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

Decision trees for the labor data

Soybean classification

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Number of values</th>
<th>Sample value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment</td>
<td>Time of occurrence</td>
<td>July</td>
</tr>
<tr>
<td></td>
<td>Precipitation</td>
<td>3</td>
</tr>
<tr>
<td>Seed</td>
<td>Condition</td>
<td>2</td>
</tr>
<tr>
<td>Mold growth</td>
<td>Condition</td>
<td>2</td>
</tr>
<tr>
<td>Fruit</td>
<td>Condition of fruit pods</td>
<td>4</td>
</tr>
<tr>
<td>Fruit spots</td>
<td>Condition</td>
<td>5</td>
</tr>
<tr>
<td>Leaf</td>
<td>Condition</td>
<td>2</td>
</tr>
<tr>
<td>Leaf spot size</td>
<td>Condition</td>
<td>3</td>
</tr>
<tr>
<td>Stem</td>
<td>Condition</td>
<td>2</td>
</tr>
<tr>
<td>Stem lodging</td>
<td>Condition</td>
<td>2</td>
</tr>
<tr>
<td>Root</td>
<td>Condition</td>
<td>3</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Diaporthe stem canker</td>
<td>19</td>
</tr>
</tbody>
</table>
The 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

But in this domain, “leaf condition is normal” implies
“leaf malformation is absent”!

Fielded applications

- The result of learning—or the learning method itself—is deployed in practical applications
  - Processing loan applications
  - Screening images for oil slicks
  - Electricity supply forecasting
  - Diagnosis of machine faults
  - Marketing and sales
  - Separating crude oil and natural gas
  - Reducing banding in rotogravure printing
  - Finding appropriate technicians for telephone faults
  - Scientific applications: biology, astronomy, chemistry
  - Automatic selection of TV programs
  - Monitoring intensive care patients

Processing loan applications
(American Express)

- 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
- 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
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 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

Load forecasting

- Electricity supply companies need forecast of future demand for power
- Forecasts of min/max load for each hour ⇒ 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

Enter machine learning

- Prediction corrected using “most similar” days
- Attributes:
  - temperature
  - humidity
  - wind speed
  - cloud cover readings
  - plus 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
### Diagnosis of machine faults

- **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?
- **Preventative maintenance of electromechanical motors and generators**
- **Information very noisy**
- **So far:** diagnosis by expert/hand-crafted rules

### Enter machine learning

- **Available:** 600 faults with expert’s diagnosis
- **~300 unsatisfactory, rest used for training**
- **Attributes augmented by intermediate concepts that embodied causal domain knowledge**
- **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**
- **Learned rules outperformed hand-crafted ones**

### Marketing and sales I

- **Companies precisely record massive amounts of marketing and sales data**
- **Applications:**
  - **Customer loyalty:** identifying customers that 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 that need extra money during the holiday season)

### Marketing and sales II

- **Market basket analysis**
  - **Association techniques find groups of items that tend to occur together in a transaction** (used to analyze checkout data)
- **Historical analysis of purchasing patterns**
- **Identifying prospective customers**
  - **Focusing promotional mailouts** (targeted campaigns are cheaper than mass-marketed ones)
Machine learning and statistics

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

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
  - \[ 4 \times 4 \times 3 \times 3 \times 2 = 288 \] possible combinations
  - With 14 rules \[ \Rightarrow 2.7 \times 10^{34} \] possible rule sets
- Other practical problems:
  - More than one description may survive
  - No description may survive
    - Language is unable to describe target concept
    - or data contains noise
- Another view of generalization as search: hill-climbing in description space according to pre-specified matching criterion
  - Most practical algorithms use heuristic search that cannot guarantee to find the optimum solution

Statisticians

- Sir Ronald Aylmer Fisher
  - Born: 17 Feb 1890 London, England
  - Died: 29 July 1962 Adelaide, Australia
  - Numerous distinguished contributions to developing the theory and application of statistics for making quantitative a vast field of biology

- Leo Breiman
  - Developed decision trees
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*
    - E.g. specializing a rule by adding conditions
  - *Specific-to-general*
    - E.g. generalizing an individual instance into a rule

Overfitting-avoidance bias

- Can be seen as a form of search bias
- Modified evaluation criterion
  - E.g. 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
Data mining and ethics I

- 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
  - E.g. loan applications: using some information (e.g. sex, religion, race) is unethical
- Ethical situation depends on application
  - E.g. same information ok in medical application
- Attributes may contain problematic information
  - E.g. area code may correlate with race

Data mining and ethics II

- 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?