# Machine Learning
# Chapter 1. Introduction

1

# 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
  - 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

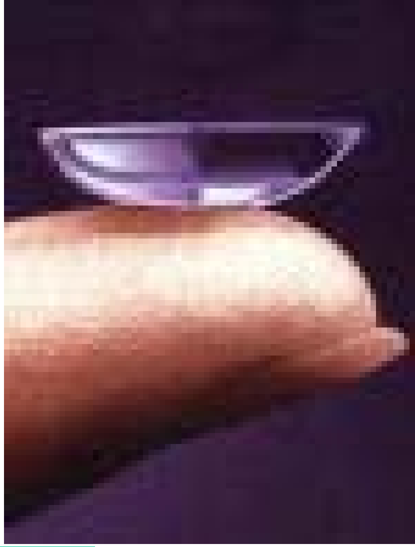
  information from data
- Needed: programs that detect patterns and regularities in the data
- Strong patterns $\Rightarrow$ good predictions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing

# 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



```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

| 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 |
| … | … | … | … | … |

# Can machines really learn?

- Definitions of “learning” from dictionary:

  **To get knowledge of by study, experience, or being taught**

  Difficult to measure

  **To become aware by information or from observation**

  **To commit to memory**

  **To be informed of, ascertain; to receive instruction**

  Trivial for computers

- 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

| 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
```

# Ross Quinlan

- Machine learning researcher from 1970's
- University of Sydney, Australia

1986 "Induction of decision trees" *ML Journal*

1993 *C4.5: Programs for machine learning*. Morgan Kaufmann

199? Started

RULEQUEST RESEARCH
*data mining tools*

# 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

| 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
```

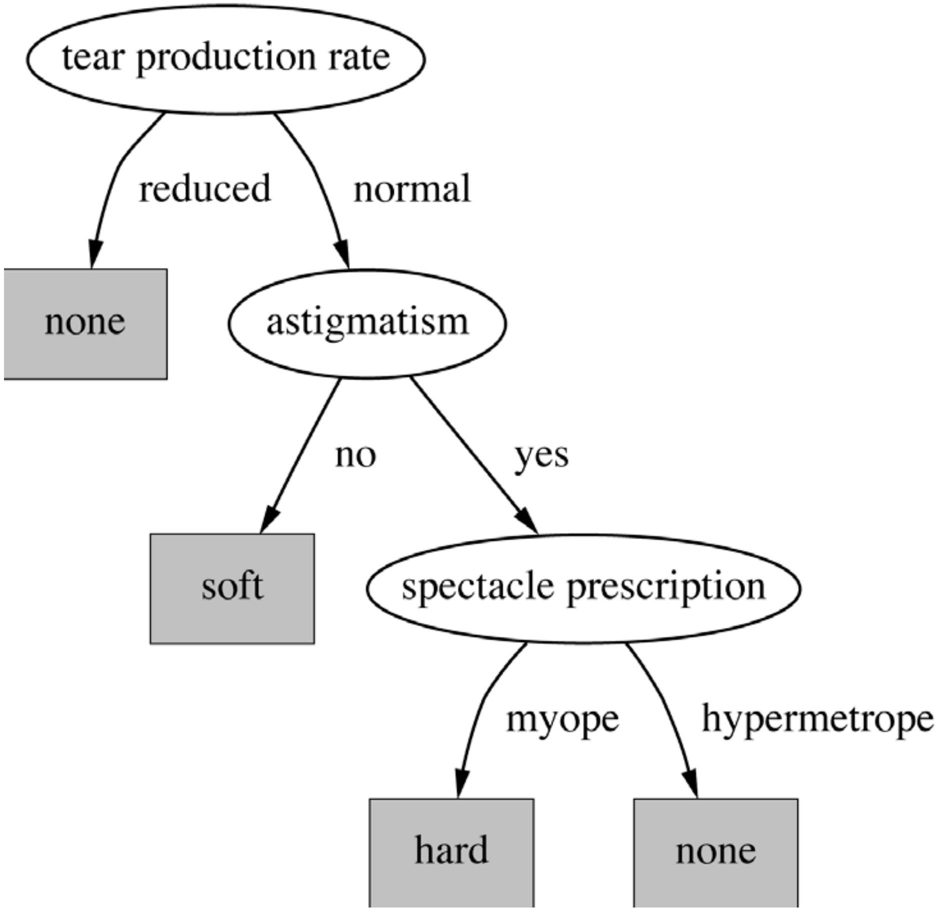
# The contact lenses data

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

# 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 = 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
```

# A decision tree for this problem



# Classifying iris flowers

| | 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 |
| … | | | | | |

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

# Predicting CPU performance

- Example: 209 different computer configurations

| | 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 |

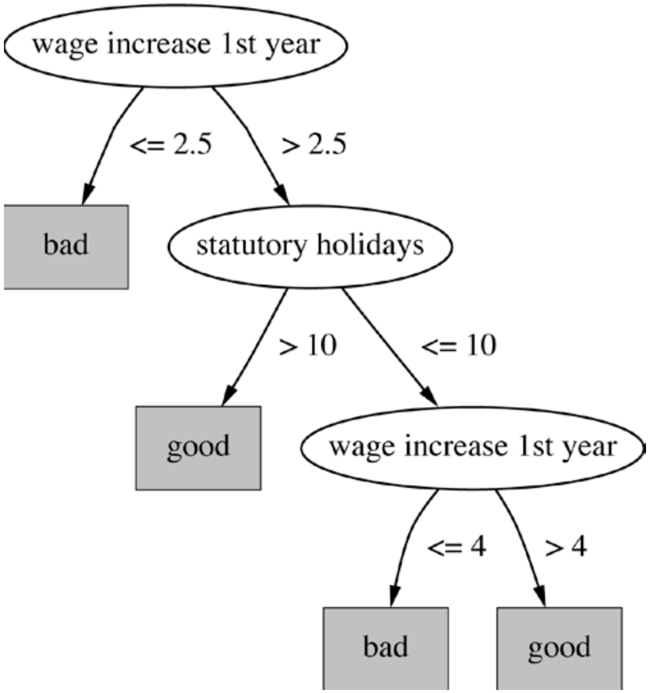
- Linear regression function

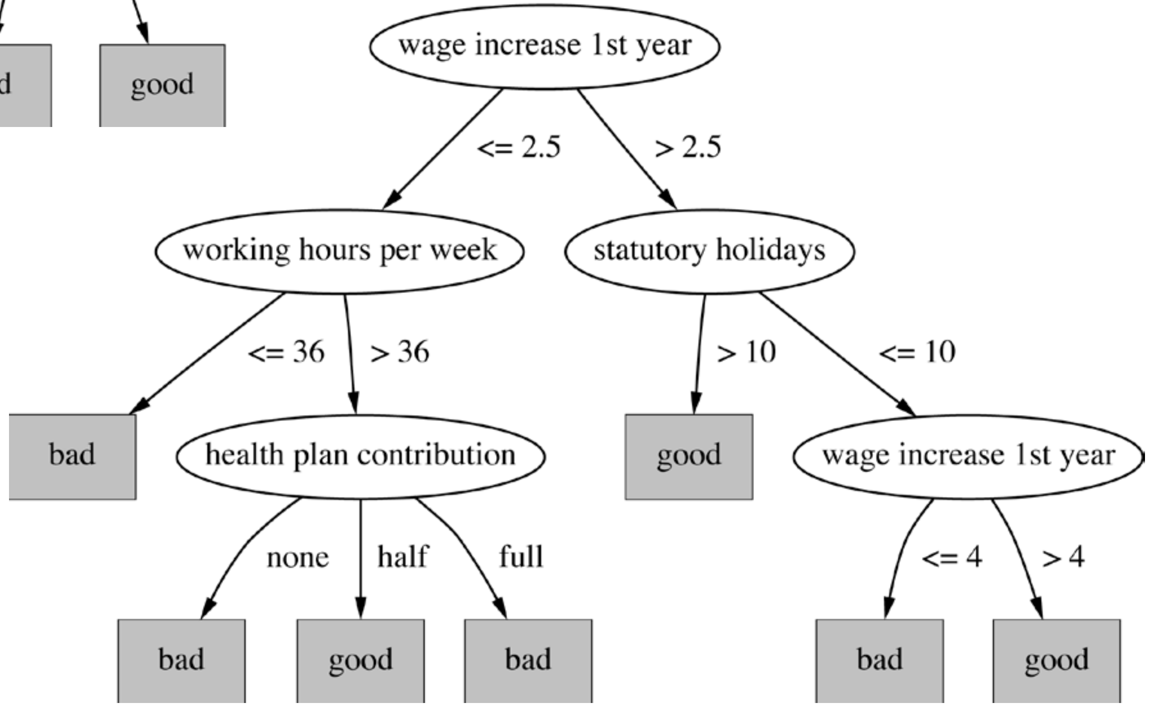
```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
      + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

# Data from labor 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 |

# Decision trees for the labor data



# Soybean classification

| | 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 | ? |
| *Leaves* | Condition | 2 | Abnormal |
| | Leaf spot size | 3 | ? |
| *…* | | | |
| *Stem* | Condition | 2 | Abnormal |
| | Stem lodging | 2 | Yes |
| *…* | | | |
| *Roots* | Condition | 3 | Normal |
| *Diagnosis* | | 19 | Diaporthe stem canker |

# 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
  - Reducing banding in rotogravure printing
  - Autoclave layout for aircraft parts
  - Automatic classification of sky objects
  - Automated completion of repetitive forms
  - Text retrieval

# Processing loan application

(American Express)

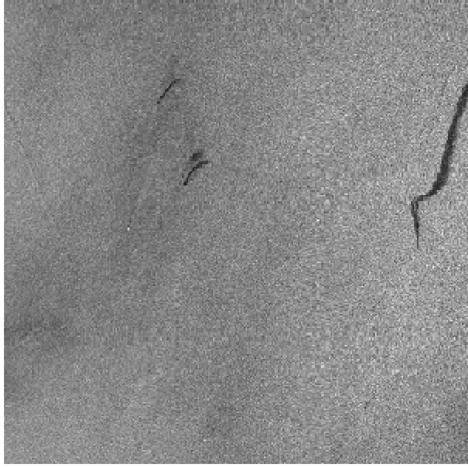
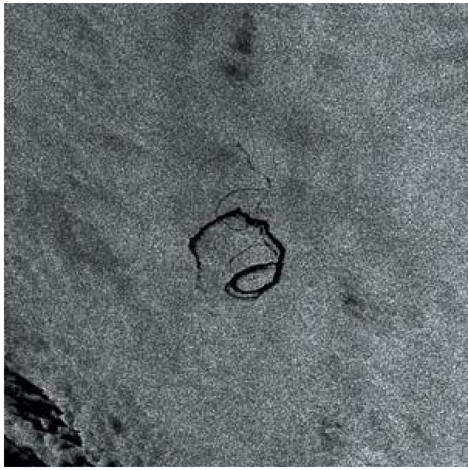
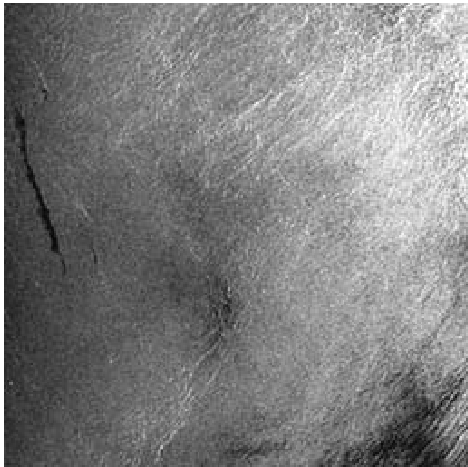
- Given: questionnaire with financial and personal information
- Question: should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?
  - No! Borderline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

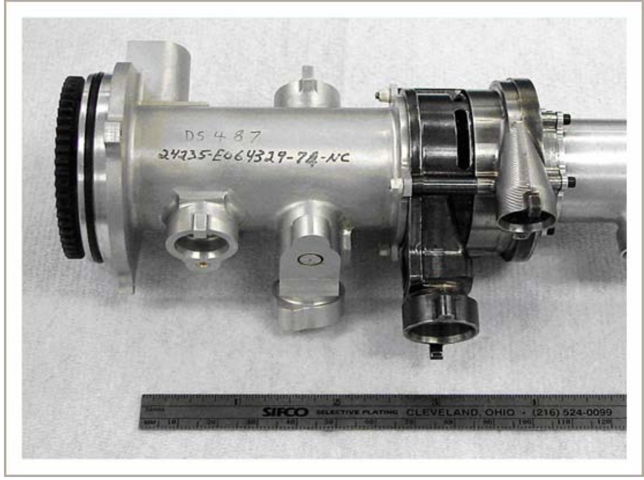
# Load forecasting

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

# 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
- *1984 Classification and Regression Trees*. Wadsworth.

# 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 $\Rightarrow$ $2.7 \times 10^{34}$ possible rule sets
- Solution: candidate-elimination algorithm
- Other practical problems:
  - More than one description may survive
  - No description may survive
    - Language is unable to describe target concept
    - *or* data contains noise

# The version space

- Space of consistent concept descriptions
- Completely determined by two sets
  - *L*: most specific descriptions that cover all positive examples and no negative ones
  - *G*: most general descriptions that do not cover any negative examples and all positive ones
- Only *L* and *G* need be maintained and updated
- But: still computationally very expensive
- And: does not solve other practical problems

# Version space example

- Given: red or green cows or chicken

| | |
|---|---|
| *L*={} | *G*={<*, *>} |
| <green,cow>: positive | |
| *L*={<green, cow>} | *G*={<*, *>} |
| <red,chicken>: negative | |
| *L*={<green, cow>} | *G*={<green,*>,<*,cow>} |
| <green, chicken>: positive | |
| *L*={<green, *>} | *G*={<green, *>} |

# Candidate-elimination algorithm

```
Initialize L and G
For each example e:
  If e is positive:
    Delete all elements from G that do not cover e
    For each element r in L that does not cover e:
      Replace r by all of its most specific generalizations
        that 1. cover e and
             2. are more specific than some element in G
    Remove elements from L that
      are more general than some other element in L
  If e is negative:
    Delete all elements from L that cover e
    For each element r in G that covers e:
      Replace r by all of its most general specializations
        that 1. do not cover e and
             2. are more general than some element in L
    Remove elements from G that
      are more specific than some other element in G
```

# 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
- 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?