Course Objectives:
- Introduction of basic machine learning and data mining techniques.
- Get a practical feel through the Weka software environment.
- Will more or less follow textbook
- Weka Software: download Weka3.6 (see syllabus)

Course Grade:
- Three midterms (20% each)
- Final (30%)
- Homework (10%)

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

Examples: Classification
- Example 1: Spam Detection
  - Given: an email message X
  - Problem: Determine whether X is a spam
  - Data: A labeled set of email messages

- Example 2: Hand-Written Character Recognition
  - Given: hand-written character
  - Problem: determine the corresponding letter
  - Data: a set of training examples
Examples: Regression

- Given a customer (annual income, family size, location, job, ...), determine the annual credit card spending based on a sample of customers data.
- Same type as classification except the output is a numerical value.
- General setup: Given a set of training examples \((x, f(x))\), determine an approximation of the function \(f\).

Examples: Clustering

- Group a large collection of Web documents into a small number of groups, each of which contains “similar” documents.
- Cluster a market into distinct clusters, where a cluster represents a group of similar customers to be targeted with a distinct marketing mix.

Machine learning techniques

- *Algorithms for acquiring structural descriptions from examples – Supervised Learning*
- 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 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>No</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Myope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### 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>
</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>
</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>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>No</td>
<td>Normal</td>
<td>Soft</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Reduced</td>
<td>None</td>
</tr>
<tr>
<td>Young</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</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>Myope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</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>Pre-presbyopic</td>
<td>Hypermetrope</td>
<td>Yes</td>
<td>Normal</td>
<td>None</td>
</tr>
<tr>
<td>Presbyopic</td>
<td>Myope</td>
<td>No</td>
<td>Reduced</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>Myope</td>
<td>Yes</td>
<td>Normal</td>
<td>Hard</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>
**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 = hypermetropic 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 = hypermetropic and astigmatic = yes then recommendation = none
- If age = presbyopic and spectacle prescription = hypermetropic and astigmatic = yes then recommendation = none

**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>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>1.9</td>
<td>Iris setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris setosa</td>
</tr>
<tr>
<td>7.0</td>
<td>3.2</td>
<td>4.7</td>
<td>1.4</td>
<td>Iris versicolor</td>
</tr>
<tr>
<td>6.4</td>
<td>3.2</td>
<td>4.7</td>
<td>1.5</td>
<td>Iris versicolor</td>
</tr>
<tr>
<td>6.3</td>
<td>3.3</td>
<td>6.0</td>
<td>2.5</td>
<td>Iris virginica</td>
</tr>
<tr>
<td>5.8</td>
<td>2.7</td>
<td>5.1</td>
<td>1.9</td>
<td>Iris virginica</td>
</tr>
</tbody>
</table>

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

**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>4</td>
<td>5</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>4.3%</td>
</tr>
<tr>
<td>Wage increase second year</td>
<td>Percentage</td>
<td>5%</td>
<td>5%</td>
<td>4.4%</td>
<td>4.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Wage increase third year</td>
<td>Percentage</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Cost of living adjustment</td>
<td>(none,tcf,tc)</td>
<td>tsf</td>
<td>tsf</td>
<td>tcf</td>
<td>4%</td>
<td>4%</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>40</td>
</tr>
<tr>
<td>Pension</td>
<td>(none,ret-all, emp)</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Standby pay</td>
<td>Percentage</td>
<td>5%</td>
<td>5%</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td>Shift-work supplement</td>
<td>Percentage</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Education allowance</td>
<td>(yes,no)</td>
<td>11</td>
<td>15</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Statutory holidays</td>
<td>(Number of days)</td>
<td>avg</td>
<td>gen</td>
<td>gen</td>
<td>avg</td>
<td>avg</td>
</tr>
<tr>
<td>Vacation</td>
<td>(below-avg,avg,gen)</td>
<td>no</td>
<td>?</td>
<td>?</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Long-term disability</td>
<td>(yes,no)</td>
<td>11</td>
<td>15</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Dental plan contribution</td>
<td>(none,half,full)</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>half</td>
<td>full</td>
</tr>
<tr>
<td>Bereavement assistance</td>
<td>(yes,no)</td>
<td>11</td>
<td>15</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Health plan contribution</td>
<td>(none,half,full)</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>half</td>
<td>full</td>
</tr>
<tr>
<td>Acceptability of contract</td>
<td>(good,bad)</td>
<td>bad</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
</tbody>
</table>
**Decision trees for the labor data**

- **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

**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

**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)
- Expert time consuming and highly trained person

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

- 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

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
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)

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

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)

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
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?