So far

- Data mining as “knowledge discovery”
- Frequent itemsets
- Descriptive analysis
- Clustering
- Seriation
- DWH/OLAP/BI
“Machine learning”
- Terminology, foundations, general framework.

Supervised machine learning
- Basic ideas, algorithms & toy examples.

Statistical challenges
- P-values, significance, consistency, stability

State of the art techniques
- SVM, kernel methods, graphical models, latent variable models, boosting, bagging, LASSO, on-line learning, deep learning, reinforcement learning, …
A Dear Child has Many Names

Data mining, Data analysis, Statistical analysis, Pattern discovery, Statistical learning, Machine learning, Predictive analytics, Business intelligence, Data-driven statistics, Inductive reasoning, Pattern analysis, Knowledge discovery from databases, Analytical processing, …
.. is mainly about methods for **modeling data** and **mining patterns**.

- **To gain knowledge**
  - Bioinformatics, LHC physics, Web analytics, ...

- **To infer intelligent behaviour from data**
  - Spam filtering, Automated recommendations, OCR, robotics, fraud detection, ...

- **To automatically organize data**
  - Data summarization, compression, noise reduction, ...
Typical approaches

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.21</td>
</tr>
<tr>
<td>0.49</td>
<td>0.42</td>
</tr>
<tr>
<td>0.67</td>
<td>0.34</td>
</tr>
<tr>
<td>1.02</td>
<td>0.55</td>
</tr>
<tr>
<td>1.53</td>
<td>0.71</td>
</tr>
<tr>
<td>0.31</td>
<td>0.33</td>
</tr>
<tr>
<td>0.35</td>
<td>0.70</td>
</tr>
<tr>
<td>0.21</td>
<td>1.20</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Typical approaches

- Clustering ("Unsupervised learning")
Typical approaches

- Regression, classification ("Supervised learning")
Typical approaches

- Outlier detection
Typical approaches

- Frequent pattern mining

AATAACGGCCCGATGAGGAAACGAACGGTCGCACT
AAAGATGAGACATGTCGCCGAAAGGTGCATAAGTTAT
GGACGAAAAACTTTCTTCTCGCCCTTTGATGTGCCCC
AGCGC GGATGAGGATCAGCCCCCGCATTTAGTTCA
ATATGCGAGCTTTCGCGCTCGGAAAGGGCAATAAA
GCAGCGGCCCCGATGAGGGGTGTACTAGATTGGA
TGGGTGGTTTCAAGATCTCGGCTTTACCCCCCTTTATCA
ACCCTGCTACAGACTCGTTGAGAATGCTACGGGATC
Typical approaches

- “Specific” pattern mining
Machine learning: How?

The approach depends strongly on **application**

The general principle is the same, though:

1. Define a **set of patterns** of interest
2. Define a **measure of goodness** for the patterns
3. Find the **best pattern** in the data
Machine learning: How?

- The approach depends strongly on application

The general principle is the same, though:

1. Define a set of patterns of interest
2. Define a measure of goodness for the patterns
3. Find the best pattern in the data

Hence, heavy use of statistics and optimization. (In other words, heavy maths).
## Supervised learning

<table>
<thead>
<tr>
<th>Observation</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer of 2003 was <strong>cold</strong></td>
<td>Winter of 2003 was <strong>warm</strong></td>
</tr>
<tr>
<td>Summer of 2004 was <strong>cold</strong></td>
<td>Winter of 2004 was <strong>cold</strong></td>
</tr>
<tr>
<td>Summer of 2005 was <strong>cold</strong></td>
<td>Winter of 2005 was <strong>cold</strong></td>
</tr>
<tr>
<td>Summer of 2006 was <strong>hot</strong></td>
<td>Winter of 2006 was <strong>warm</strong></td>
</tr>
<tr>
<td>Summer of 2007 was <strong>cold</strong></td>
<td>Winter of 2007 was <strong>cold</strong></td>
</tr>
<tr>
<td>Summer of 2008 was <strong>warm</strong></td>
<td>Winter of 2008 was <strong>warm</strong></td>
</tr>
<tr>
<td>Summer of 2009 was <strong>warm</strong></td>
<td><strong>Winter of 2009 will be ?</strong></td>
</tr>
</tbody>
</table>
# Supervised learning

<table>
<thead>
<tr>
<th>Observation</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study=hard,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=slack,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=hard,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=slack,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=slack,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=slack,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=hard,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>Study=slack,</td>
<td>Professor=😊</td>
</tr>
<tr>
<td>?</td>
<td></td>
</tr>
</tbody>
</table>

5.11.2009
## Supervised learning

<table>
<thead>
<tr>
<th>Day</th>
<th>Observation</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>I was not using magnetic bracelet™</td>
<td>In the evening I had a headache</td>
</tr>
<tr>
<td>Tue</td>
<td>I was using magnetic bracelet™</td>
<td>In the evening I had less headache</td>
</tr>
<tr>
<td>Wed</td>
<td>I was using magnetic bracelet™</td>
<td>In the evening no headache!</td>
</tr>
<tr>
<td>Thu</td>
<td>I was using magnetic bracelet™</td>
<td>The headache is gone!!</td>
</tr>
<tr>
<td>Fri</td>
<td>I was not using magnetic bracelet™</td>
<td>No headache!!</td>
</tr>
</tbody>
</table>
Supervised learning

<table>
<thead>
<tr>
<th>Day</th>
<th>Observation</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>I was <strong>not</strong> using magnetic bracelet™</td>
<td>In the evening I had a headache</td>
</tr>
<tr>
<td>Tue</td>
<td>I was using magnetic bracelet™</td>
<td>In the evening I had less headache</td>
</tr>
<tr>
<td>Wed</td>
<td>I was using magnetic bracelet™</td>
<td>In the evening no headache!</td>
</tr>
<tr>
<td>Thu</td>
<td>I was using magnetic bracelet™</td>
<td>The headache is gone!!</td>
</tr>
<tr>
<td>Fri</td>
<td>I was <strong>not</strong> using magnetic bracelet™</td>
<td>No headache!!</td>
</tr>
</tbody>
</table>

Magnetic bracelet™ cures headache
Supervised learning
Supervised learning
Supervised learning
Supervised learning
Supervised learning
Supervised learning

- Formally,

Let $\mathcal{X}$ and $\mathcal{Y}$ be some sets and let there be a dataset of training samples:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$$

Find a function $f_D : \mathcal{X} \rightarrow \mathcal{Y}$ generalizing the functional relationship present in the data.
Regression

- $\mathcal{X} = \mathbb{R}$, \( \mathcal{Y} = \mathbb{R} \).
- \( D = \{(0.50, 0.26), (0.43, 0.08), (0.26, 0.00), \ldots\} \)
- \( f_D(x) = x^2 \).
Classification

- \( \mathcal{X} = \mathbb{R}^2 \), \( \mathcal{Y} = \{ \text{blue, red} \} \).
- \( D = \{ ((1.3, 0.8), \text{red}), ((2.5, 2.3), \text{blue}), \ldots \} \).
- \( f_D(x_1, x_2) = \text{if } x_1 + x_2 > 3 \text{ then blue else red.} \)
The “Dumb User” Perspective

Weka, RapidMiner, MSSSAS, Clementine, SPSS, R, …
The “Dumb User” Perspective

Validation

\[ f(x) = \sum_i \omega_i x_i \]
Classification demo: Iris dataset

- 150 measurements, 4 attributes, 3 classes
Classification demo: Iris dataset

```
Tree View

- petalwidth
  - <= 0.6
    - Iris-setosa (50.0)
  - > 0.6
    - petalwidth
      - <= 1.7
        - petallength
          - <= 4.9
            - Iris-versicolor (48.0/1.0)
          - > 4.9
            - Iris-virginica (46.0/1.0)
      - > 1.7
        - Iris-virginica (3.0)
    - petalwidth
      - <= 1.5
        - Iris-virginica (3.0)
      - > 1.5
        - Iris-versicolor (3.0/1.0)
```
Validation

\[
\begin{array}{ccc}
  a & b & c \\
  50 & 0 & 0 & a = \text{Iris-setosa} \\
  0 & 49 & 1 & b = \text{Iris-versicolor} \\
  0 & 2 & 48 & c = \text{Iris-virginica} \\
\end{array}
\]

Correctly Classified Instances 147 98%
Incorrectly Classified Instances 3 2%
Kappa statistic 0.97
Mean absolute error 0.0233
Root mean squared error 0.108
Relative absolute error 5.2482 %
Root relative squared error 22.9089 %
Total Number of Instances 150
Validation

\[ a \ b \ c \quad \text{classified as} \]

\[
\begin{align*}
50 & \quad 0 & \quad 0 & \quad | \quad a = \text{Iris-setosa} \\
0 & \quad 49 & \quad 1 & \quad | \quad b = \text{Iris-versicolor} \\
0 & \quad 2 & \quad 48 & \quad | \quad c = \text{Iris-virginica}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Class</th>
<th>setosa</th>
<th>versic.</th>
<th>virg.</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>1</td>
<td>0.98</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Precision</td>
<td>1</td>
<td>0.961</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Recall</td>
<td>1</td>
<td>0.98</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>F-Measure</td>
<td>1</td>
<td>0.97</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>ROC Area</td>
<td>1</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Validation

\[
\begin{array}{ccc}
a & b & c & \text{classified as} \\
50 & 0 & 0 & a = \text{Iris-setosa} \\
0 & 49 & 1 & b = \text{Iris-versicolor} \\
0 & 2 & 48 & c = \text{Iris-virginica} \\
\end{array}
\]

versic.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
<td>0.98</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.02</td>
</tr>
<tr>
<td>Precision</td>
<td>0.961</td>
</tr>
<tr>
<td>Recall</td>
<td>0.98</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.97</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Validation

\[
\begin{array}{ccc}
\text{a} & \text{b} & \text{c} \\
50 & 0 & 0 & \text{classified as Iris-setosa} \\
0 & 49 & 1 & \text{Iris-versicolor} \\
0 & 2 & 48 & \text{Iris-virginica} \\
\end{array}
\]

“True positives” “False positives”

\[
\text{TP Rate} = \frac{\text{TP}}{\text{positive examples}} = 0.98
\]

\[
\text{FP Rate} = \frac{\text{FP}}{\text{negative examples}} = 0.02
\]

\[
\text{Precision} = \frac{\text{TP}}{\text{positives}} = 0.961
\]

\[
\text{Recall} = \frac{\text{TP}}{\text{positive examples}} = 0.98
\]

\[
\text{F-Measure} = \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}} = 0.97
\]

\[
\text{ROC Area} \sim \text{Pr}(s(\text{false}) < s(\text{true})) = 0.99
\]
## Classification summary

<table>
<thead>
<tr>
<th>Positives</th>
<th>Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted = Yes</td>
<td>Predicted = No</td>
</tr>
<tr>
<td><strong>Actual = Yes</strong></td>
<td>True positives (TP)</td>
</tr>
<tr>
<td><strong>Actual = No</strong></td>
<td>False positives (FP) (Type I, α-error)</td>
</tr>
</tbody>
</table>
## Classification summary

<table>
<thead>
<tr>
<th></th>
<th>Positives</th>
<th>Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted = Yes</td>
<td>Predicted = No</td>
</tr>
<tr>
<td>Actual = Yes</td>
<td>True positives (TP)</td>
<td>False negatives (FN)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Type II, β-error)</td>
</tr>
<tr>
<td>Actual = No</td>
<td>False positives (FP)</td>
<td>True negatives (FN)</td>
</tr>
<tr>
<td></td>
<td>(Type I, α-error)</td>
<td></td>
</tr>
</tbody>
</table>

- **Recall**
- **Accuracy**
- **Precision**

F-measure = harmonic_mean(Precision, Recall)
“Training” classifiers on data

- Thus, a “good” classifier is the one which has good Accuracy/Precision/Recall.

- Hence, machine learning boils down to finding a function that optimizes these parameters for given data.
“Training” classifiers on data

- Thus, a “good” classifier is the one which has good Accuracy/Precision/Recall.

- Hence, machine learning boils down to **finding a function** that optimizes these parameters for **given data**.

- Yet, there’s a catch
“Training” classifiers on data

- We want our algorithm to perform well on “unseen” data!

  - This makes algorithms and theory way more complicated.

  - This makes validation somewhat more complicated.
Proper validation

- You **may not test your algorithm on the same data that you used to train it!**
Proper validation

- You **may not test your algorithm** on the same data that you used to train it!
Proper validation :: Holdout

Split

Training set

Testing set

Validation
Proper validation

- What are the “sufficient” sizes for the test/training sets and why?
- What if the data is scarce?
  - Cross-validation
  - K-fold cross-validation
  - Leave-one-out cross-validation
  - Bootstrap .632+
Intermediate summary

- **Supervised learning** = predicting f(x) well.
- For classification, “well” = high accuracy/precision/recall on unseen data.
- To achieve that, most training algorithms will try to optimize their accuracy/precision/recall on training data.
- We can then validate how good they are on test data.
Three examples of approaches

- Ad-hoc
  - Decision tree induction
- Probabilistic modeling
  - Naïve Bayes classifier
- Objective function optimization
  - Linear least squares regression
Decision Tree Induction :: ID3

- **Iterative Dichotomizer 3**
  - Simple yet popular decision tree induction algorithm
  - Builds a decision tree top-down, starting at the root.

Ross Quinlan
## ID3

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temp</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Which split is the most informative?
Entropy

- Entropy measures the “informativeness” of a probability distribution.

\[ H(p) = - \sum_{i} p_i \log_2 p_i \]

- A split is informative if it reduces entropy.
Information gain of a split

- Before split:
  \[ p_{\text{no}} = \frac{5}{14}, \quad p_{\text{yes}} = \frac{9}{14}, \quad H(p) = 0.94 \]

- After split on outlook:

\[
\begin{align*}
H &= 0.97 \\
H &= 0 \\
H &= 0.97 \\
&= 0.97 \left( \frac{5}{14} \right) + 0 \left( \frac{4}{14} \right) + 0.97 \left( \frac{5}{14} \right) \\
&= 0.69
\end{align*}
\]

Information gain

\[ = 0.94 - 0.69 = 0.25 \]
ID3

1. Start with a single node
2. Find the attribute with the largest information gain
3. Split the node according to this attribute
4. Repeat recursively on subnodes
C4.5

- C4.5 is an extension of ID3
  - Supports continuous attributes
  - Supports missing values
  - Supports pruning

- There is also a C5.0
  - A commercial version with additional bells & whistles
Decision trees

- **The goods:**
  - Easy & efficient
  - Interpretable and pretty

- **The bads**
  - Rather ad-hoc
  - Can overfit unless properly pruned
  - Not the best model for all classification tasks
Three examples of approaches

- Ad-hoc
  - Decision tree induction
- Probabilistic modeling
  - Naïve Bayes classifier
- Objective function optimization
  - Linear least squares regression
Next: Naïve Bayes Classifier

To be continued…
Questions?

Why are we doing this?
What problem are we solving?
Is this actually useful?
Are we adding value?
Will this change behavior?
Is there an easier way?
What's the opportunity cost?
Is it really worth it?