Data Mining

Lecturer:

- Peter Lucas

Assessment:

- Written exam at the end of part II
- Practical assessment

'Compulsory' study material:

- Transparencies
- Handouts (mostly on the Web)

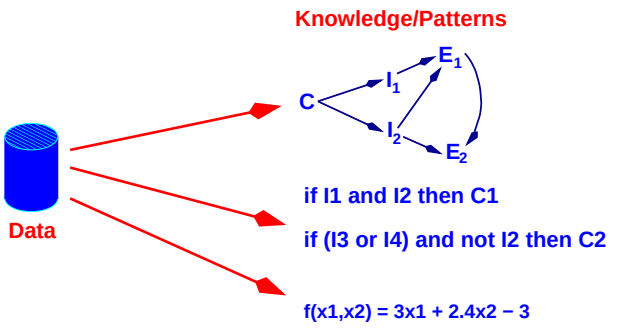
Course Information:

`http://www.cs.kun.nl/~peterl/teaching/DM`

Background literature

- I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, 2000.
- M. Berthold and D.J. Hand, *Intelligent Data Analysis: An Introduction*, Springer, Berlin, 1999.
- T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, 2001.
- D. Hand, H. Mannila and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.
- T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

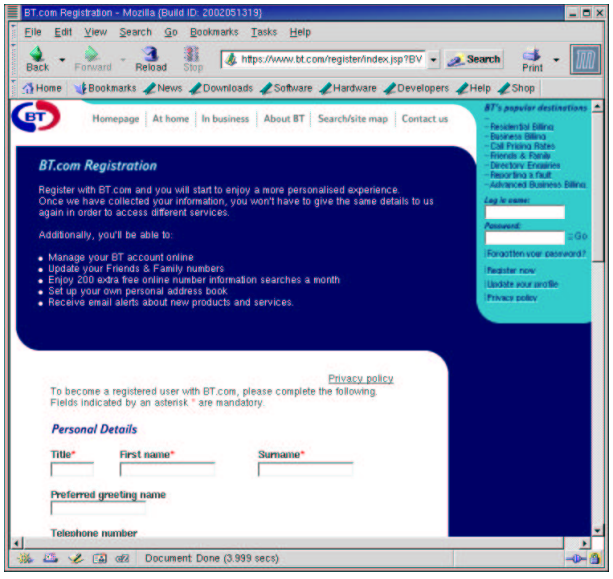
Data mining: what is it?



Process data, taking into account:

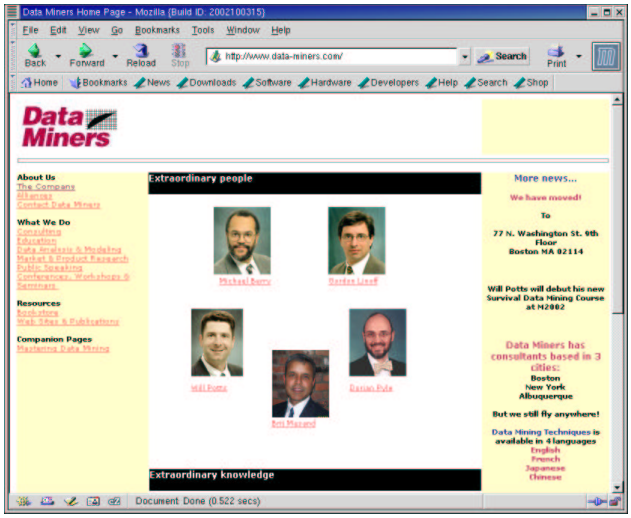
- assumptions about data (meaning, relevance, purpose)
- knowledge about domain from which data were obtained
- target representation (rules, decision trees, polynomial, etc.) – often called **models**

Electronic customer support



- Many companies are now collecting electronic information about their customers
- This information can be explored

Data mining is business – consultancy



Consultants help companies:

- setting up data-mining environments
- training people

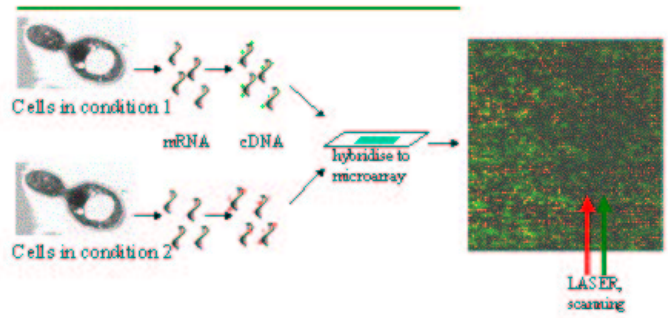
Data mining is business – software

Software houses:

- develop data-mining tools
- train people in using these tools

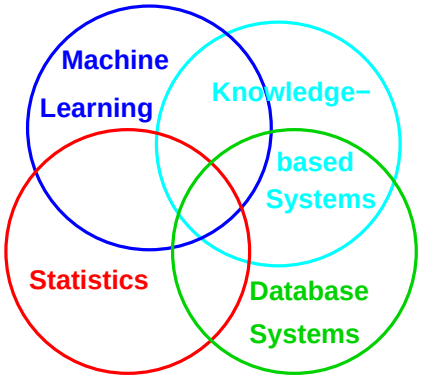
Data mining is business – hardware

In a failing economy – Bioinformatics



- **Microarray:** expression of genetic feature
- Analysis: data-mining, machine learning
- Purpose: characterisation of cells, e.g. cancer cells

Data mining – relationships



Data-mining draws upon various fields:

- Statistics – model construction and evaluation
- Machine learning
- Knowledge-based systems – representation
- Database systems – data extraction

Datasets – ARFF: Attribute Relation File Format

```
% Title: Final settlements in labor negotitions
% in Canadian industry
% Creators: Collective Barganing Review, montly publication,
% Labour Canada, Industrial Relations Information Service,
% Ottawa, Ontario, K1A 0J2, Canada, (819) 997-3117

@relation labor-neg-data
@attribute duration real
@attribute wage-increase-first-year real
@attribute wage-increase-second-year real
@attribute wage-increase-third-year real
@attribute cost-of-living-adjustment {none,tcf,tc}
@attribute working-hours real
@attribute pension {none,ret_allw,empl_contr}
                    ...
@attribute contribution-to-health-plan {none,half,full}
@attribute class {bad,good}

@data
1,5,?,?,?,40,?,?,2,?,11,average,?,?,yes,?,good
3,3.7,4,5,tc,?,?,?,?,yes,?,?,?,?,yes,?,good
3,4.5,4.5,5,?,40,?,?,?,?,12,average,?,half,yes,half,good
2,2,2.5,?,?,35,?,?,6,yes,12,average,?,?,?,?,good
3,6.9,4.8,2.3,?,40,?,?,3,?,12,below_average,?,?,?,?,good
2,3,7,?,?,38,?,12,25,yes,11,below_average,yes,half,yes,?,good
2,7,5.3,?,?,?,?,?,?,?,11,?,yes,full,?,?,good
3,2,3,?,tcf,?,empl_contr,?,?,yes,?,?,yes,half,yes,?,good
3,3.5,4,4.5,tcf,35,?,?,?,?,13,generous,?,?,yes,full,good
```

Problem types

Given a dataset DS $= (A, D)$, with attributes A and multiset $D = \langle \mathbf{x}_1, \ldots, \mathbf{x}_N \rangle$, instance $\mathbf{x}_i$

- **Preprocessing:** DS $\rightarrow$ DS$'$
- **Attribute selection:** $A \rightarrow A'$, with $A' \subseteq A$
- **Supervised learning:**
  - Classification

    $$f(\mathbf{x}_i) = c \in \{\top, \bot\}$$

    with $x_{i,j} \in \{\top, \bot\}$, and f *classifier*
  - Prediction/regression

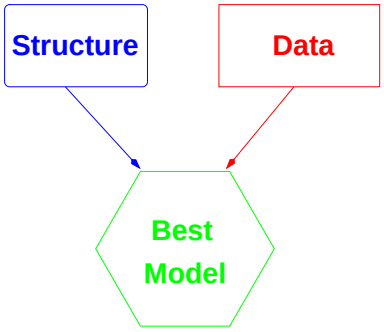
    $$f(\mathbf{x}_i) = c \in \mathbb{R}$$

    with $\mathbf{x}_i \in \mathbb{R}^p$, and f *predictor*
- **Unsupervised learning:**
  - Clustering

    $$f(\mathbf{x}_i) = k \in \{1, \ldots, m\}$$

    with f *clustering function*, $\mathbf{x}_i \in \mathbb{R}^p$ and k *encoder*

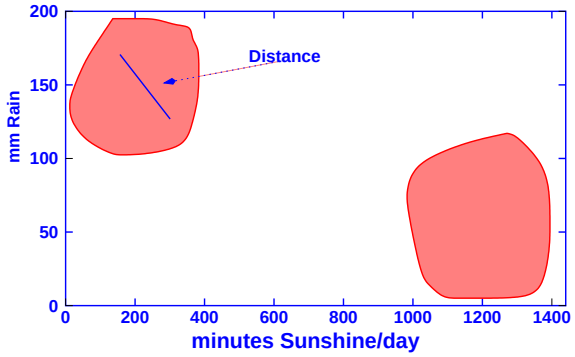
Learning and search



- **Supervised learning:**
  - Output (class) variable known and *indicated* for every instance
  - Aim is to learn a model that predicts the output (class)

Day	Average Temp.	Rain (mm)	Pressure (mb)
1	3	0.7	1011
2	2.1	0	1024
⋮	⋮	⋮	⋮

Learning and search (continued)

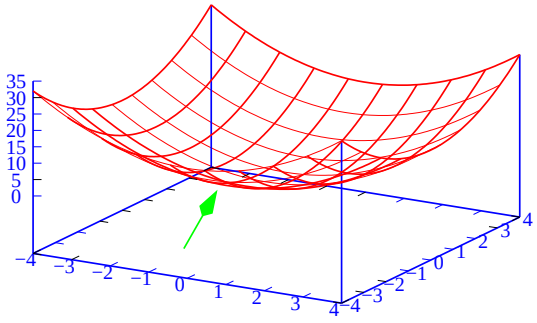


- **Unsupervised learning:**
  - No class variable indicated
  - Finding 'similar' (clusters) cases using e.g. similarity or distance measures:

$$||\mathbf{x} - \mathbf{y}|| = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2} < d$$

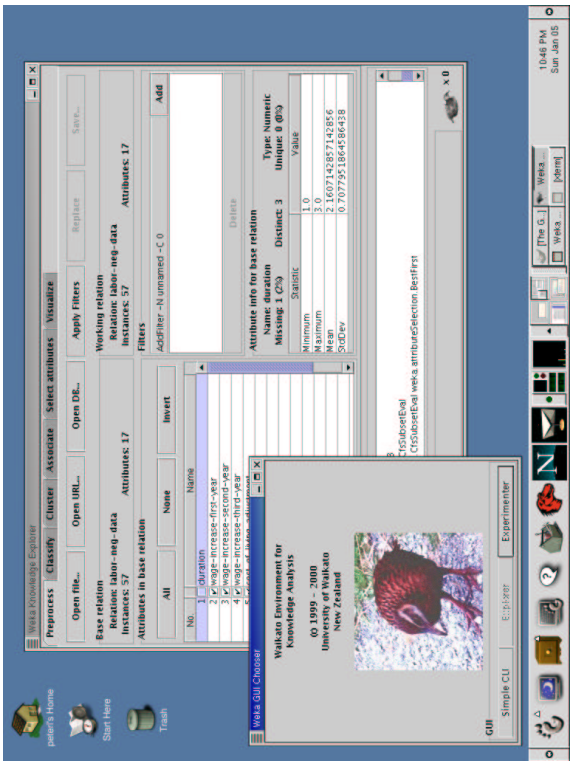
with $d \in \mathbb{R}$

Learning and search (continued)



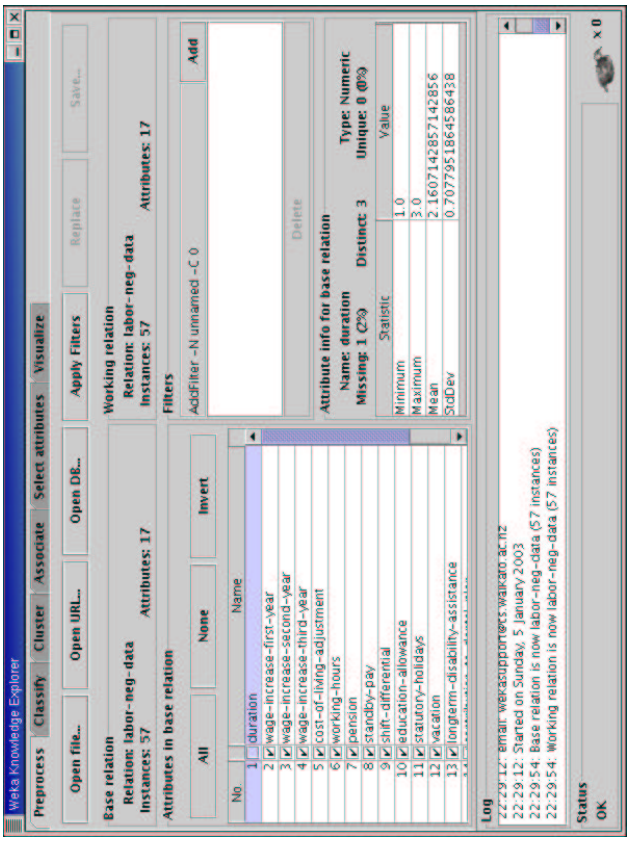
- Developing (including learning) a model can be viewed as searching through a search space of possible models
- Search may be very expensive computationally
- Special search principles (e.g. heuristics) may be required

WEKA – Waikato Environment for Knowledge Analysis

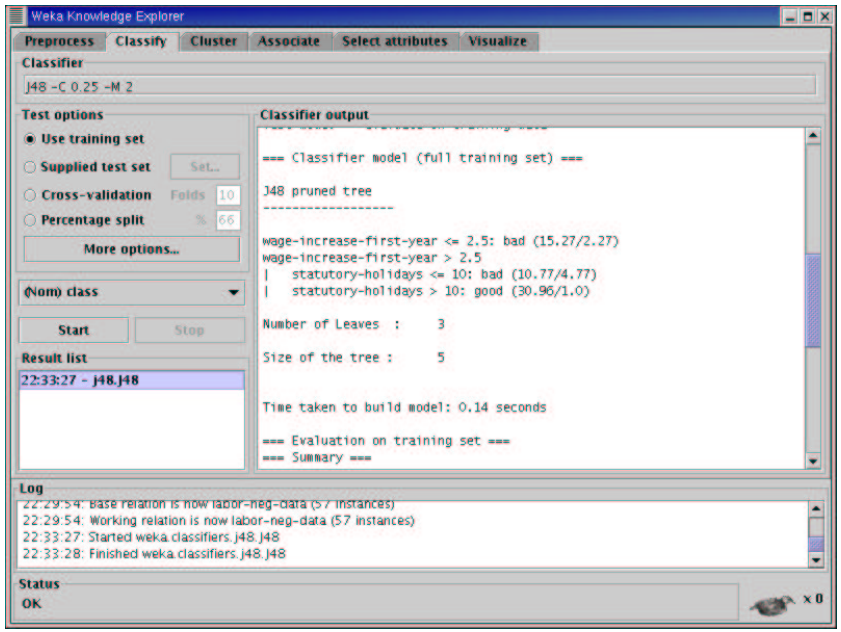


http://www.cs.waikato.ac.nz/ml/weka

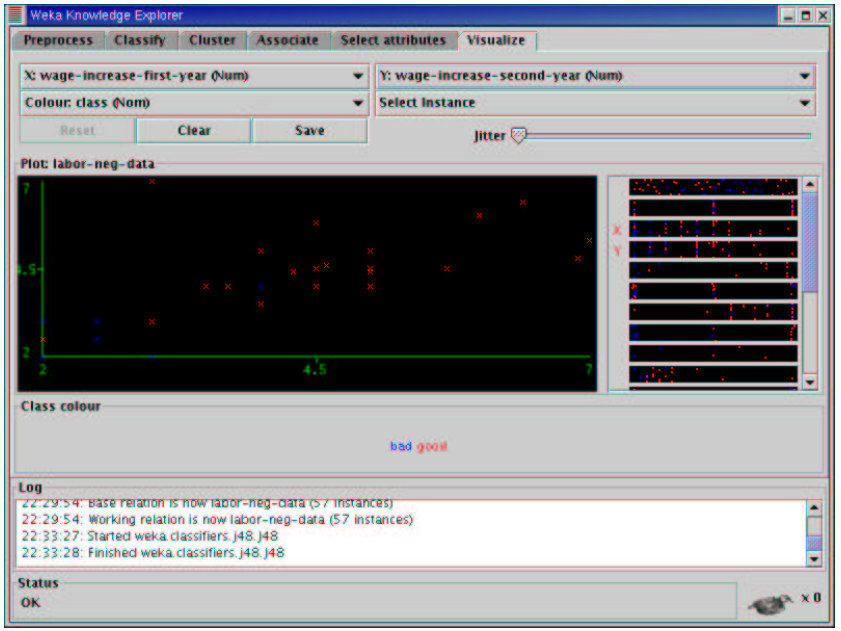
WEKA – Preprocessing



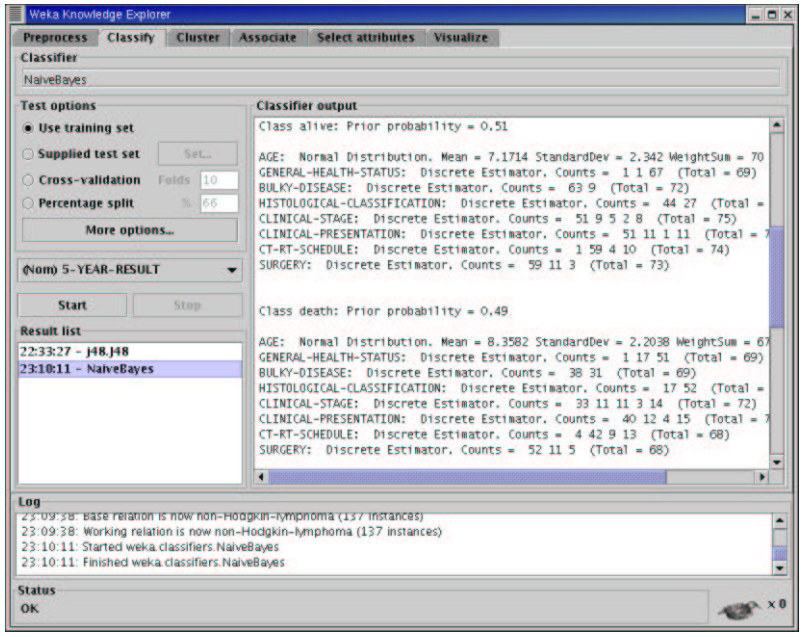
WEKA – Classification by decision tree



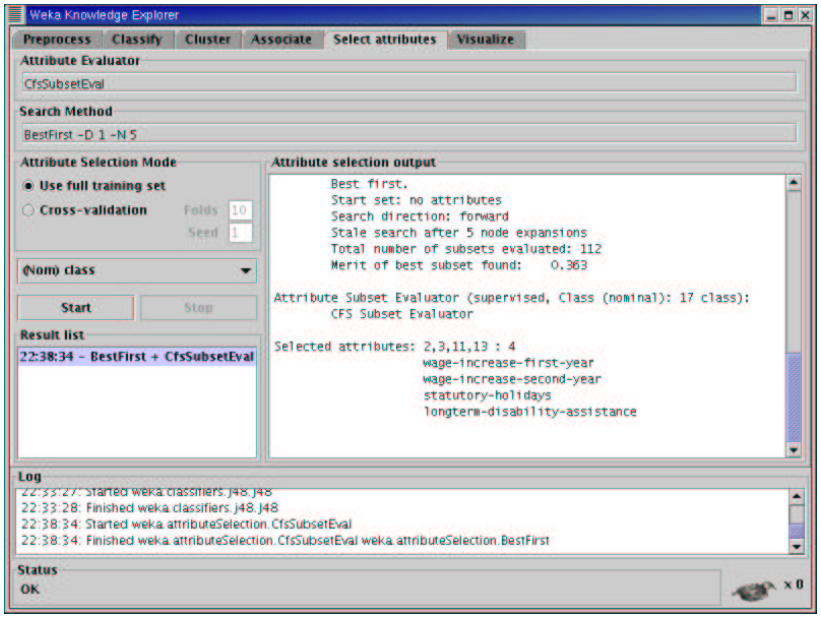
WEKA – Visualisation



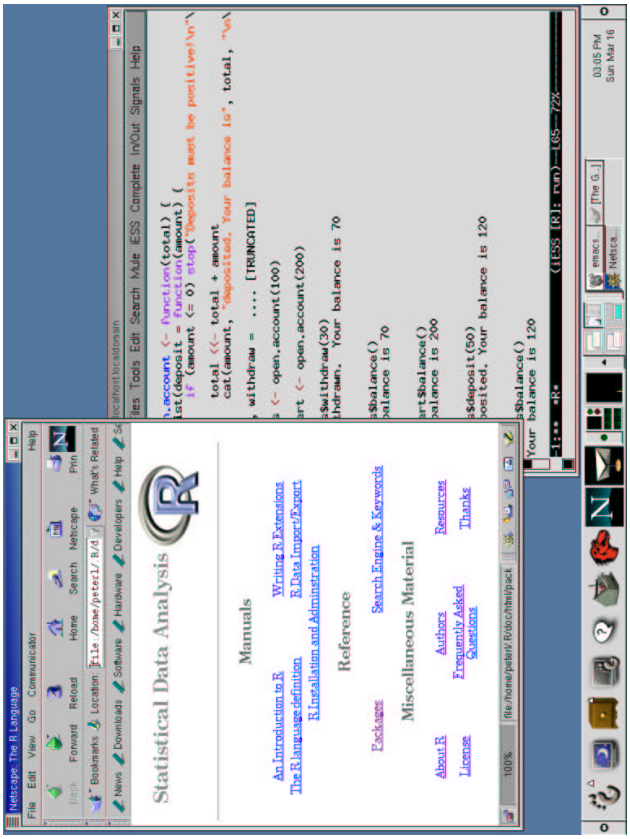
WEKA – Classification by Naive Bayes



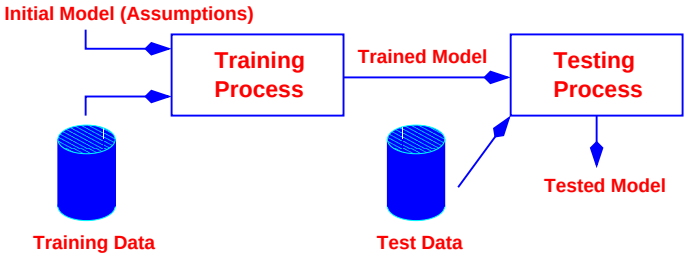
WEKA – Attribute selection



R: statistical data analysis



Data mining & ML cycle

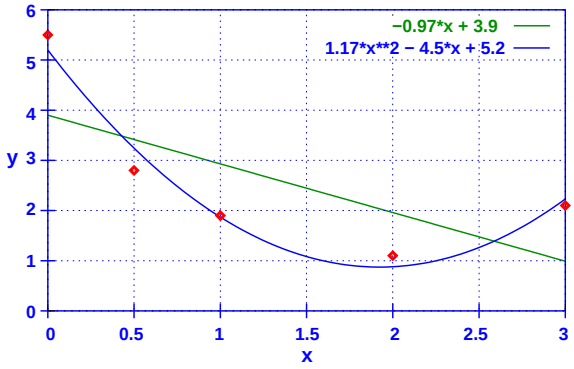


Datasets:

- training data: used for model building
- test data: used for model evaluation
- preferably disjoint datasets

What constitutes a good model? – training

- Suppose a process is governed by the (unknown) function $f(x) = -1x + 4$
- Training data:



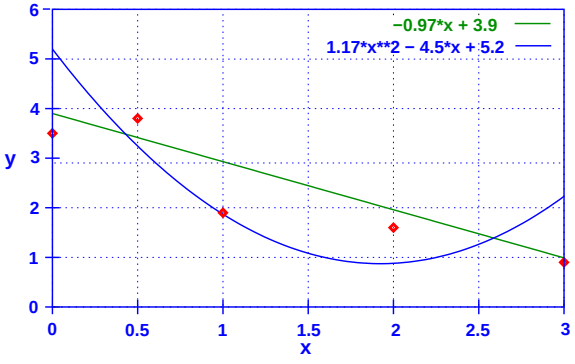
- Fitted (least squares) functions:

$f(x) = -0.97x + 3.9$

$g(x) = 1.17x^2 - 4.5x + 5.2$

What constitutes a good model? – testing

- Suppose a process is governed by the (unknown) function $f(x) = -1x + 4$
- Testing data:



- Fitted (least squares) functions:

$f(x) = -0.97x + 3.9$

$g(x) = 1.17x^2 - 4.5x + 5.2$

Flexibility of model

- Compare:

$$f(x) = a_1 x + a_0$$

$$g(x) = a_2 x^2 + a_1 x + a_0$$

then $f(x) = g(x)$, $\forall x \in \mathbb{R}$, if $a_2 = 0$ (function f special case of g)

- More parameters $\Rightarrow$ more flexibility
- Danger that model overfits training data
- Bias-variance decomposition: analytic description of sources of errors:
  - model assumptions
  - adaptation to data

Basic tools

- X: random variable (discrete or continuous)
- Probability distribution:
  - Discrete: $P(X)$
  - Continuous: $f(x)$ probability density function:

$$P(X \leq x) = \int_{-\infty}^{x} f(x)\mathrm{d}x$$

- Mathematical expectation of $g(x)$ given probability distribution P
  - Discrete case:

$$\mathrm{E}(g(X)) = \sum_X g(X)P(X)$$

  - Continuous case:

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x$$

- Example: discrete mean:

$$\mathrm{E}(X) = \sum_X XP(X)$$

Properties

- $\mathrm{E}(X)$ expresses that the values observed for X are governed by a stochastic, uncertain process
- $\mathrm{E}(ag(X) + bh(X)) = a\mathrm{E}(g(X)) + b\mathrm{E}(h(X))$
  Proof (for continuous case):

$$\begin{aligned}
&\mathrm{E}(a\,g(X) + b\,h(X)) = \\
&= \int_{-\infty}^{\infty} [ag(x) + bh(x)]\, f(x)\mathrm{d}x \\
&= a\int_{-\infty}^{\infty} g(x)f(x)\mathrm{d}x + b\int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \\
&= a\,\mathrm{E}(g(X)) + b\,\mathrm{E}(h(X))
\end{aligned}$$

- $\mathrm{E}(c) = c$, with c constant
  Proof (for continuous case):

$$\begin{aligned}
\mathrm{E}(c) &= \int_{-\infty}^{\infty} cf(x)\mathrm{d}x \\
&= c\int_{-\infty}^{\infty} f(x)\mathrm{d}x \\
&= c \cdot 1
\end{aligned}$$

Bias-variance decomposition

- T: training dataset
- $Y = f(X)$ is predictor of the process
- $\hat{y} = \hat{f}_T(\mathbf{x})$: prediction of y based on training data T
- Mean squared error:

$$M_T(\mathbf{x}) = \mathrm{E}\left(\left[f(\mathbf{x}) - \hat{f}_T(\mathbf{x})\right]^2\right)$$

with expectation E over training data T

- Bias:

$$B_T(\mathbf{x}) = \mathrm{E}(f(\mathbf{x}) - \hat{f}_T(\mathbf{x}))$$

model assumption effects

- Variance:

$$V_T(\mathbf{x}) = \mathrm{E}\left(\left[\hat{f}_T(\mathbf{x}) - \mathrm{E}(\hat{f}_T(\mathbf{x}))\right]^2\right)$$

effects of variation in data

Bias-variance decomposition

- Mean squared error:

$$\begin{aligned} M_T(\mathbf{x}) &= \mathrm{E}([f(\mathbf{x}) - \hat{f}_T(\mathbf{x})]^2) \\ &= \mathrm{E}([f(\mathbf{x})]^2 - 2f(\mathbf{x})\hat{f}_T(\mathbf{x}) + [\hat{f}_T(\mathbf{x})]^2) \\ &= [f(\mathbf{x})]^2 - 2f(\mathbf{x})\mathrm{E}(\hat{f}_T(\mathbf{x})) + \\ &\quad \mathrm{E}([f_T(\mathbf{x})]^2) \\ &= [B_T(\mathbf{x})]^2 + V_T(\mathbf{x}) \end{aligned}$$

- Bias (note that $\mathrm{E}(c) = c$):

$$\begin{aligned} B_T(\mathbf{x}) &= \mathrm{E}(f(\mathbf{x}) - \hat{f}_T(\mathbf{x})) \\ &= \mathrm{E}(f(\mathbf{x})) - \mathrm{E}(\hat{f}_T(\mathbf{x})) \\ &= f(\mathbf{x}) - \mathrm{E}(\hat{f}_T(\mathbf{x})) \end{aligned}$$

$$\begin{aligned} \Rightarrow [B_T(\mathbf{x})]^2 &= [\mathrm{E}(f(\mathbf{x}) - \hat{f}_T(\mathbf{x}))]^2 \\ &= [f(\mathbf{x})]^2 - 2f(\mathbf{x})\mathrm{E}(\hat{f}_T(\mathbf{x})) + \\ &\quad [\mathrm{E}(\hat{f}_T(\mathbf{x}))]^2 \end{aligned}$$

- Variance:

$$\begin{aligned} V_T(\mathbf{x}) &= \mathrm{E}([\hat{f}_T(\mathbf{x}) - \mathrm{E}(\hat{f}_T(\mathbf{x}))]^2) \\ &= \mathrm{E}([\hat{f}_T(\mathbf{x})]^2) - \mathrm{E}(\hat{f}_T(\mathbf{x}))\mathrm{E}(\hat{f}_T(\mathbf{x})) \end{aligned}$$

Bias-variance decomposition

- Mean squared error:

$$\begin{aligned} M_T(\mathbf{x}) &= \mathrm{E}([f(\mathbf{x}) - \hat{f}_T(\mathbf{x})]^2) \\ &= \mathrm{E}([f(\mathbf{x})]^2 - 2f(\mathbf{x})\hat{f}_T(\mathbf{x}) + [\hat{f}_T(\mathbf{x})]^2) \\ &= [f(\mathbf{x})]^2 - 2f(\mathbf{x})\mathrm{E}(\hat{f}_T(\mathbf{x})) + \\ &\quad \mathrm{E}([f_T(\mathbf{x})]^2) \\ &= [B_T(\mathbf{x})]^2 + V_T(\mathbf{x}) \end{aligned}$$

- Bias:
  $[B_T(\mathbf{x})]^2 = [f(\mathbf{x})]^2 - 2f(\mathbf{x})\mathrm{E}(\hat{f}_T(\mathbf{x})) + [\mathrm{E}(\hat{f}_T(\mathbf{x}))]^2$
- Variance:

$$\begin{aligned} V_T(\mathbf{x}) &= \mathrm{E}([\hat{f}_T(\mathbf{x}) - \mathrm{E}(\hat{f}_T(\mathbf{x}))]^2) \\ &= \mathrm{E}([\hat{f}_T(\mathbf{x})]^2) - 2\mathrm{E}(\hat{f}_T(\mathbf{x}))\mathrm{E}(\hat{f}_T(\mathbf{x})) + \\ &\quad [\mathrm{E}(\hat{f}_T(\mathbf{x}))]^2 \\ &= \mathrm{E}([\hat{f}_T(\mathbf{x})]^2) - \mathrm{E}(\hat{f}_T(\mathbf{x}))\mathrm{E}(\hat{f}_T(\mathbf{x})) \end{aligned}$$

  Note that $\mathrm{E}(\mathrm{E}(\hat{f}_T(\mathbf{x}))) = \mathrm{E}(\hat{f}_T(\mathbf{x})) = c$

Course Outline

Theory:

- Learning classification rules (supervised)
- Bayesian networks (from simple to complex) (partially supervised)
- Clustering (unsupervised)

Practice:

- Data-mining software: WEKA
- BayesBuilder
- Practical assessment

Tutorials:

- Exercises