

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)

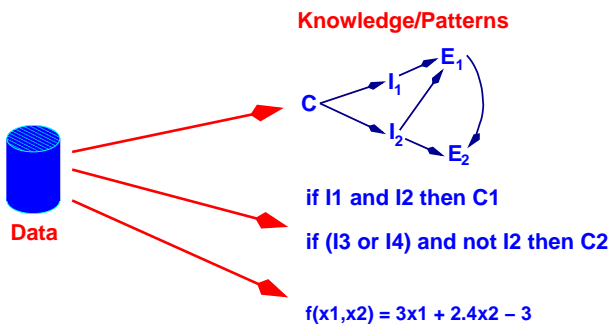
Course Information:

<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.

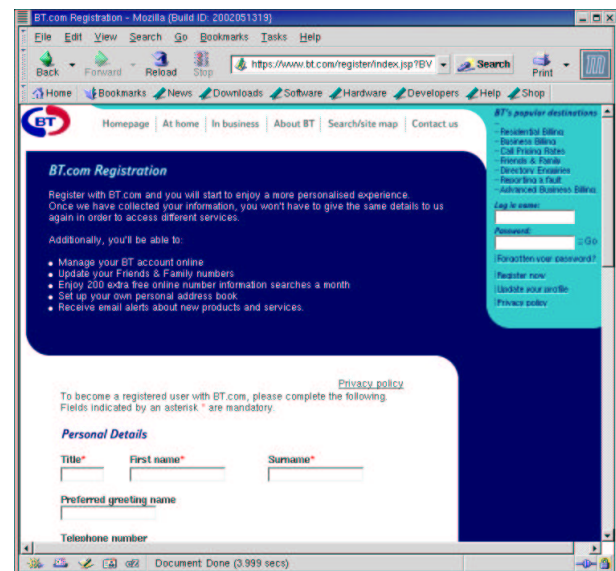
Data mining: what is it?



Process data, taking into account:

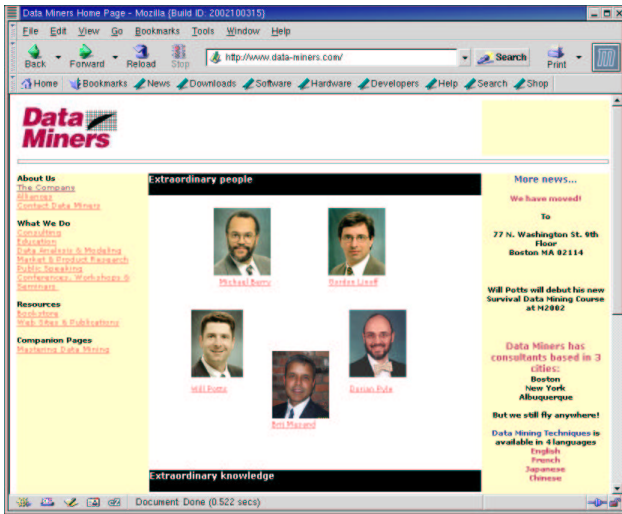
- **assumptions** about data (meaning, relevance, purpose)
- **knowledge** about domain from which data were obtained
- target **representation** (rules, decision trees, polynomial, etc.) – often called **models**

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



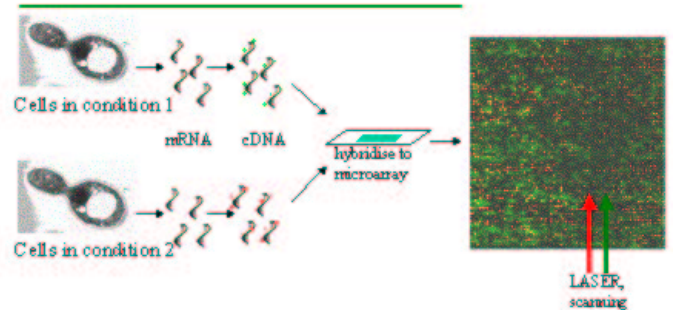
Software houses:

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

Data mining is business – hardware

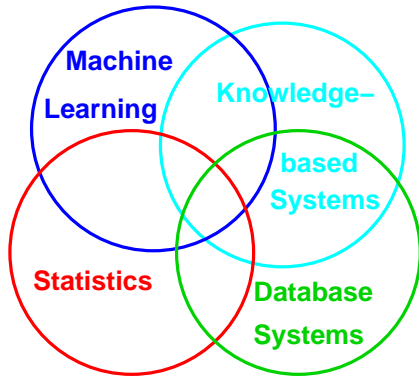


In a failing economy – Bioinformatics



- **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 negotiations
% in Canadian industry
% Creators: Collective Bargaining 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
3,4.5,4.5,5,5,?,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 $D = \langle x_1, \dots, x_N \rangle$, instance x_i

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

– Classification

$$f(x_i) = c \in \{T, \perp\}$$

with $x_{i,j} \in \{T, \perp\}$, and f classifier

– Prediction/regression

$$f(x_i) = c \in \mathbb{R}$$

with $x_i \in \mathbb{R}^p$, and f predictor

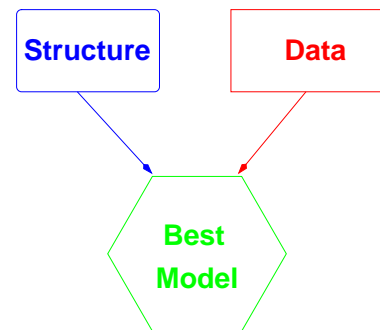
- **Unsupervised learning:**

– Clustering

$$f(x_i) = k \in \{1, \dots, m\}$$

with f clustering function, $x_i \in \mathbb{R}^p$ and k encoder

Learning and search

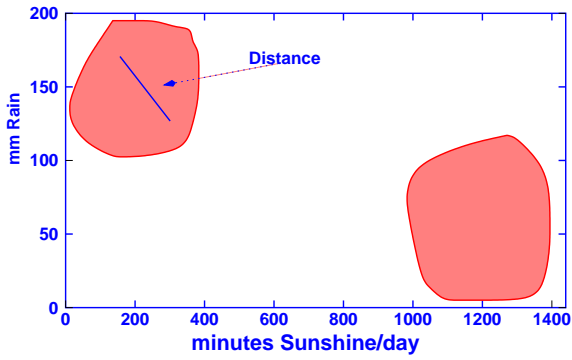


- **Supervised learning:**

- 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)



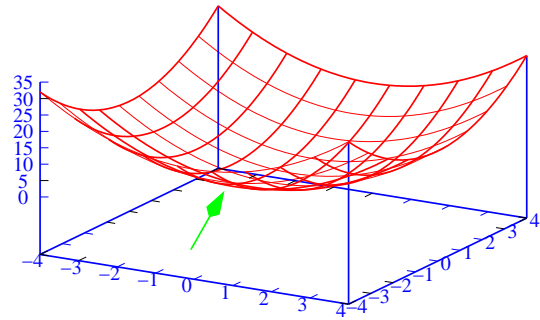
- **Unsupervised learning:**

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

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

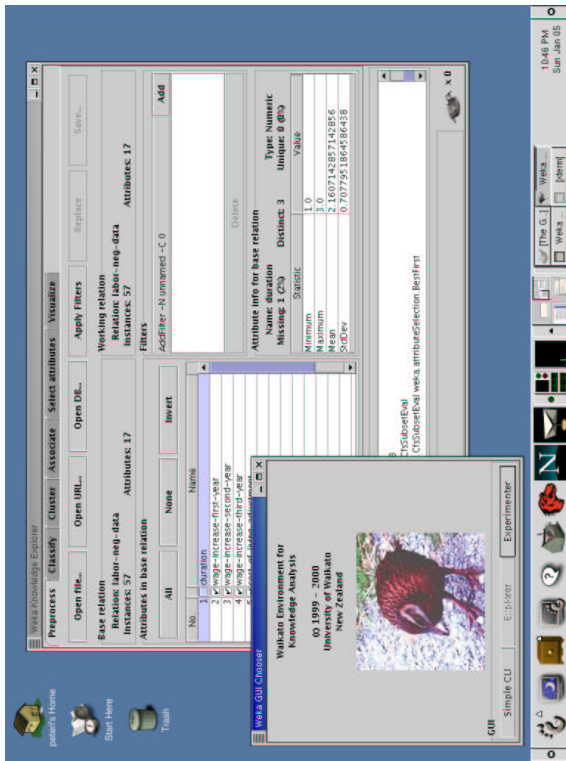
with $d \in \mathbb{R}$

Learning and search (continued)



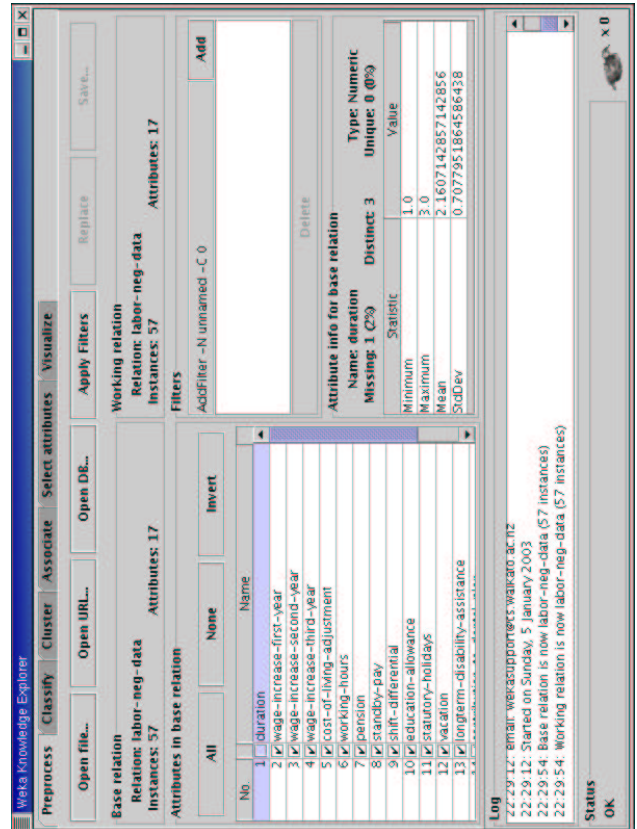
- Developing (including learning) a model can be viewed as searching through a search space of possible models
- Search may be very expensive computationally
- Special search principles (e.g. heuristics) may be required

WEKA – Waikato Environment for Knowledge Analysis



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

WEKA – Preprocessing



WEKA – Classification by decision tree

The screenshot shows the WEKA Knowledge Explorer interface with the 'Classifier' tab selected. The classifier is 'J48 -C 0.25 -M 2'. The 'Test options' section has 'Use training set' selected. The 'Classifier output' pane displays the following decision tree structure:

```

=== Classifier model (full training set) ===
J48 pruned tree
-----
wage-increase-first-year <= 2.5: bad (15.27/2.27)
wage-increase-first-year > 2.5
| statutory-holidays <= 10: bad (10.77/4.77)
| statutory-holidays > 10: good (30.96/1.0)
Number of Leaves :    3
Size of the tree :    5
Time taken to build model: 0.14 seconds

=== Evaluation on training set ===
=== Summary ===
    
```

The 'Result list' shows a single entry: '22:33:27 - J48.J48'. The 'Log' pane shows the execution timeline, and the 'Status' is 'OK'.

WEKA – Visualisation

The screenshot shows the WEKA Knowledge Explorer interface with the 'Visualize' tab selected. The 'X' axis is 'wage-increase-first-year (Num)' and the 'Y' axis is 'wage-increase-second-year (Num)'. The 'Colour' is set to 'class (Nom)'. The 'Plot: labor-neg-data' shows a scatter plot with points colored by class ('bad' in blue, 'good' in red). The 'Class colour' legend at the bottom indicates 'bad' in blue and 'good' in red. The 'Result list' shows '22:33:27 - J48.J48' and the 'Status' is 'OK'.

WEKA – Classification by Naive Bayes

The screenshot shows the WEKA Knowledge Explorer interface with the 'Classifier' tab selected. The classifier is 'NaiveBayes'. The 'Test options' section has 'Use training set' selected. The 'Classifier output' pane displays the following Naive Bayes model parameters:

```

Class alive: Prior probability = 0.51
AGE: Normal Distribution, Mean = 7.1714 StandardDev = 2.342 WeightSum = 70
GENERAL-HEALTH-STATUS: Discrete Estimator, Counts = 1 1 67 (Total = 69)
BULKY-DISEASE: Discrete Estimator, Counts = 63 9 (Total = 72)
HISTOLOGICAL-CLASSIFICATION: Discrete Estimator, Counts = 44 27 (Total = 71)
CLINICAL-STAGE: Discrete Estimator, Counts = 51 9 5 2 8 (Total = 75)
CLINICAL-PRESENTATION: Discrete Estimator, Counts = 51 11 1 11 (Total = 74)
CT-RT-SCHEDULE: Discrete Estimator, Counts = 1 59 4 10 (Total = 74)
SURGERY: Discrete Estimator, Counts = 59 11 3 (Total = 73)
Class death: Prior probability = 0.49
AGE: Normal Distribution, Mean = 8.3582 StandardDev = 2.2038 WeightSum = 67
GENERAL-HEALTH-STATUS: Discrete Estimator, Counts = 1 17 51 (Total = 69)
BULKY-DISEASE: Discrete Estimator, Counts = 38 31 (Total = 69)
HISTOLOGICAL-CLASSIFICATION: Discrete Estimator, Counts = 17 52 (Total = 69)
CLINICAL-STAGE: Discrete Estimator, Counts = 33 11 11 3 14 (Total = 72)
CLINICAL-PRESENTATION: Discrete Estimator, Counts = 40 12 4 15 (Total = 72)
CT-RT-SCHEDULE: Discrete Estimator, Counts = 4 42 9 13 (Total = 68)
SURGERY: Discrete Estimator, Counts = 52 11 5 (Total = 68)
    
```

The 'Result list' shows two entries: '22:33:27 - J48.J48' and '23:10:11 - NaiveBayes'. The 'Log' pane shows the execution timeline, and the 'Status' is 'OK'.

WEKA – Attribute selection

The screenshot shows the WEKA Knowledge Explorer interface with the 'Attribute Evaluator' tab selected. The 'Attribute Evaluator' is 'CfsSubsetEval'. The 'Search Method' is 'BestFirst -D 1 -N 5'. The 'Attribute Selection Mode' has 'Use full training set' selected. The 'Attribute selection output' pane displays the following results:

```

Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 112
Merit of best subset found: 0.363

Attribute Subset Evaluator (supervised, Class (nominal): 17 class):
CFS Subset Evaluator

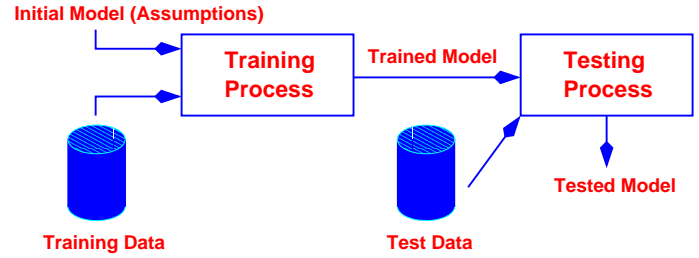
Selected attributes: 2,3,11,13 : 4
wage-increase-first-year
wage-increase-second-year
statutory-holidays
longterm-disability-assistance
    
```

The 'Result list' shows an entry: '22:38:34 - BestFirst + CfsSubsetEval'. The 'Log' pane shows the execution timeline, and the 'Status' is 'OK'.

R: statistical data analysis



Data mining & ML cycle

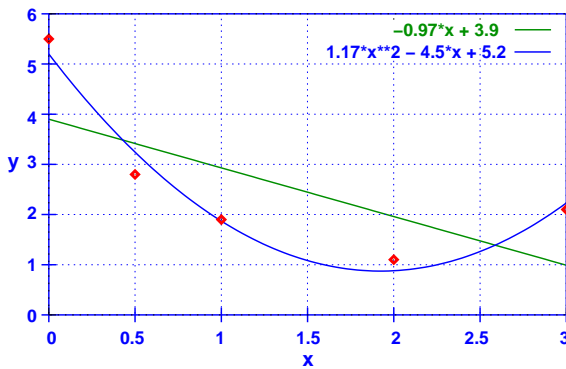


Datasets:

- 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:



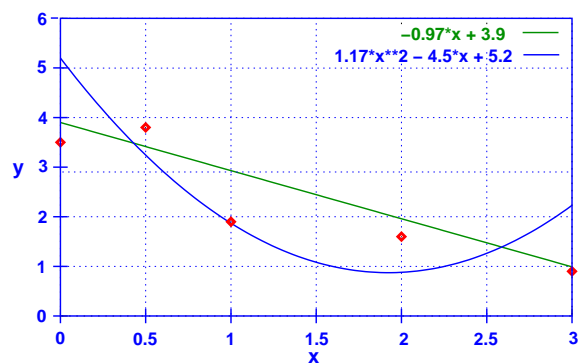
- Fitted (least squares) functions:

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

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

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:

$$f(x) = a_1x + a_0$$

$$g(x) = a_2x^2 + a_1x + a_0$$

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:

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

- **Mathematical expectation** of $g(x)$ given probability distribution P

- Discrete case:

$$E(g(X)) = \sum_X g(X)P(X)$$

- Continuous case:

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

- Example: discrete mean:

$$E(X) = \sum_X XP(X)$$

Properties

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

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

- $E(c) = c$, with c constant
Proof (for continuous case):

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

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**:

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

with expectation E over training data T

- **Bias**:

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

model assumption effects

- **Variance**:

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

effects of variation in data

Bias-variance decomposition

- Mean squared error:

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

- Bias (note that $E(c) = c$):

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

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

- Variance:

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

Bias-variance decomposition

- Mean squared error:

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

- Bias:

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

- Variance:

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

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

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