Universität Potsdam

Institut für Informatik

Lehrstuhl Maschinelles Lernen

Intelligent Data Analysis / Machine Learning

Tobias Scheffer, Shuwen Deng, David Reich, Sasha Roewer

April 10, 2024

Organization

- English and German lecture videos in Moodle:
  - Watch them in the privacy of your personal shelter.
  - There is no live lecture, go watch the video!
  - Write down any question that you have!
  - Come up with 3 good questions in any case.
- Q&A: ask all your questions!
  - Every Tuesday, 10:15-11:45, live in 02.70.0.10 and via Zoom (the link is on Moodle).
  - Meeting link and all resources are available on Moodle.
  - You **have to** watch the lecture video beforehand!
  - This is you weekly opportunity to ask all your questions—there is no email support.

Organization

- Labs and exercises (mandatory):
  - English exercise G2 (Shuwen Deng, David Reich): Tue, 14:15-15:45 in room 02.70.0.09, starting on 16.04.2024.
  - German exercise G3 (Sasha Roewer): Thursday, 10:15-11:45 in room 02.70.0.08, starting on 11.04.2024.
  - English exercise G1 (Shuwen Deng, David Reich): Thu, 12:15-13:45 in room 02.70.0.10. Starting on 11.04.2024.
- You **have to** complete the homework beforehand.
- You have to mark 70% of the homework in Moodle and present your solutions in the exercise.
- Submitting homework by email is not possible.

Modules

- Bachelor Informatik Computational Science
  - Mandatory (Intelligente Datenanalyse)
- Master Cognitive Systems
  - Mandatory
- Master Data Science
  - Mandatory
- Master Computational Science
  - Only if you did not take this lecture within the Bachelor‘s program
  - Maschinelles Lernen / Maschinelles Lernen II
  - IDA in den Naturwissenschaften

Exams

- For Students of all bachelor programs:
  - Successfully complete labs and exercises
  - Written exam for 1h immediately followed by 15 min oral exam..
- For Students of master programs:
  - Successfully complete labs and exercises
  - Successfully complete semester project
  - 15 minutes presentation of the semester project + 15 minutes oral exam.

Organization

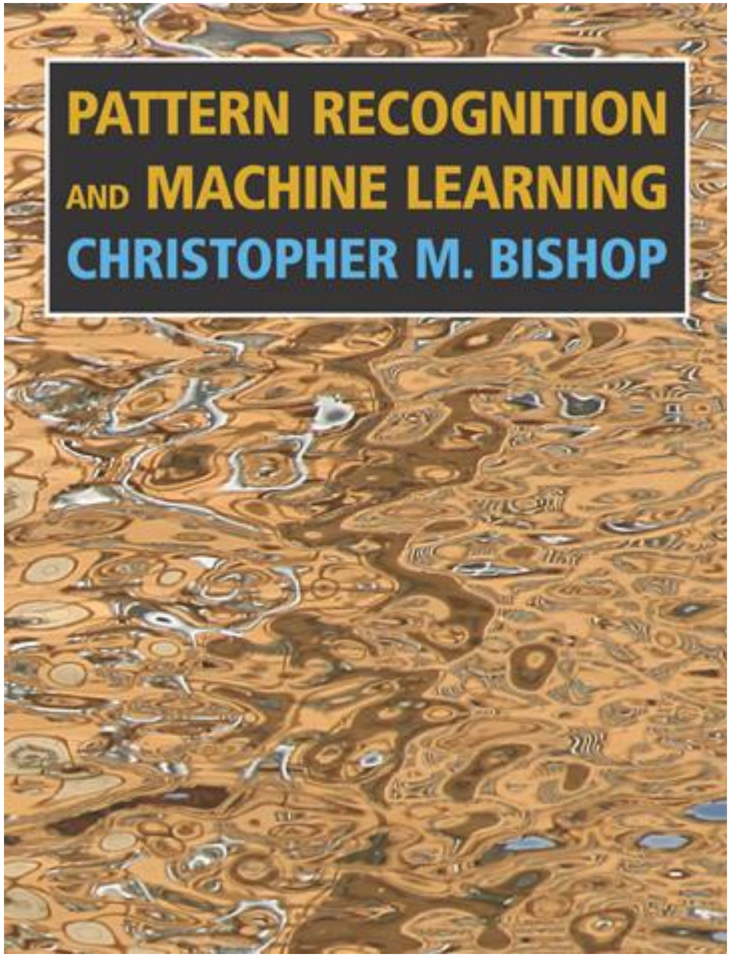
- Moodle page:
  - Slides and lecture videos.
  - Links to video conferences.
  - Introductory mathematics videos and tutorials.
  - Homework to be completes for the next lab.

This Week

- Tutorial on statistics and mathematical foundations is online
  - Recommended for MSc Cognitive Systems.
- Two lectures, “Introduction to Python” and “Models, Data, Learning Algorithms” are online.
  - Skip “Introduction to Python” if you are familiar with Python, numpy, pandas, seaborn.
  - Q&A for Python on 16.04.2023
  - Q&A for Models, Data, … on 16.04.2023
  - Thu 18. – Tue 23.04.: first exercise is due.
- Labs take place from tomorrow.

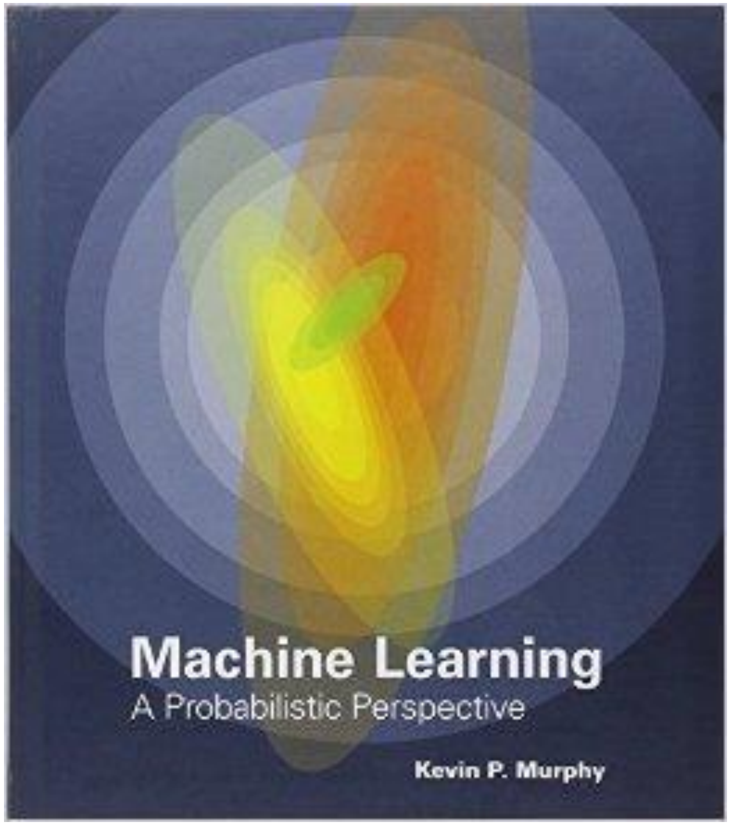
Literature

- Chris Bishop: Pattern Recognition and Machine Learning.
- 30 Copies available in the library
- Can also be found online.



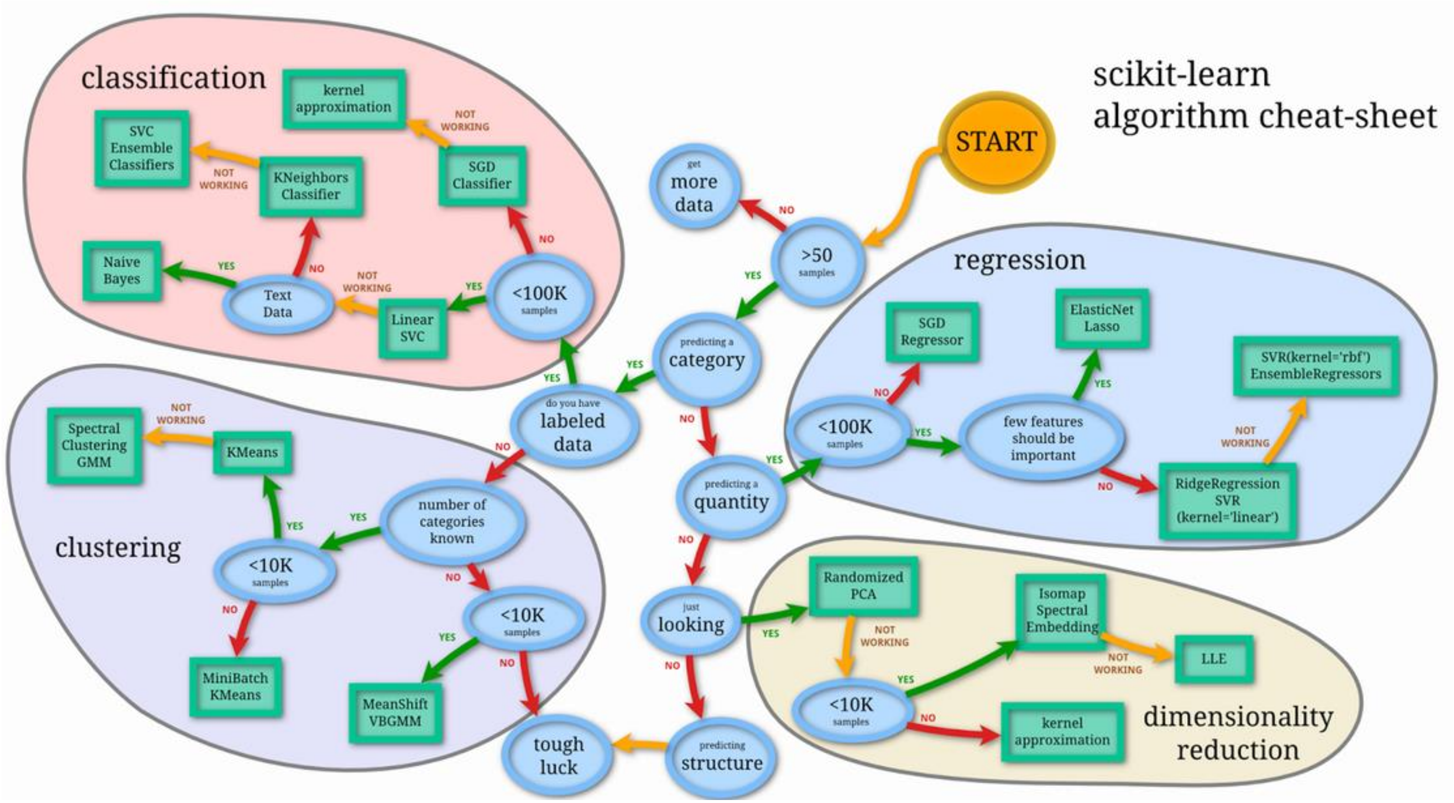
Literature

- Kevin Murphy: Machine Learning: a probabilistic perspective
- Can also be found online.



Intelligent Data Analysis

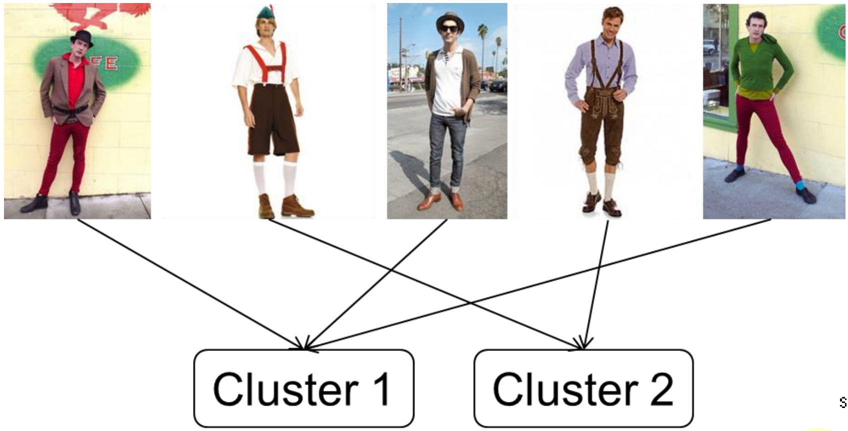
- Introduction to Python.

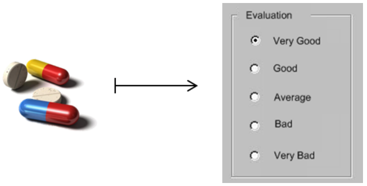


Machine Learning

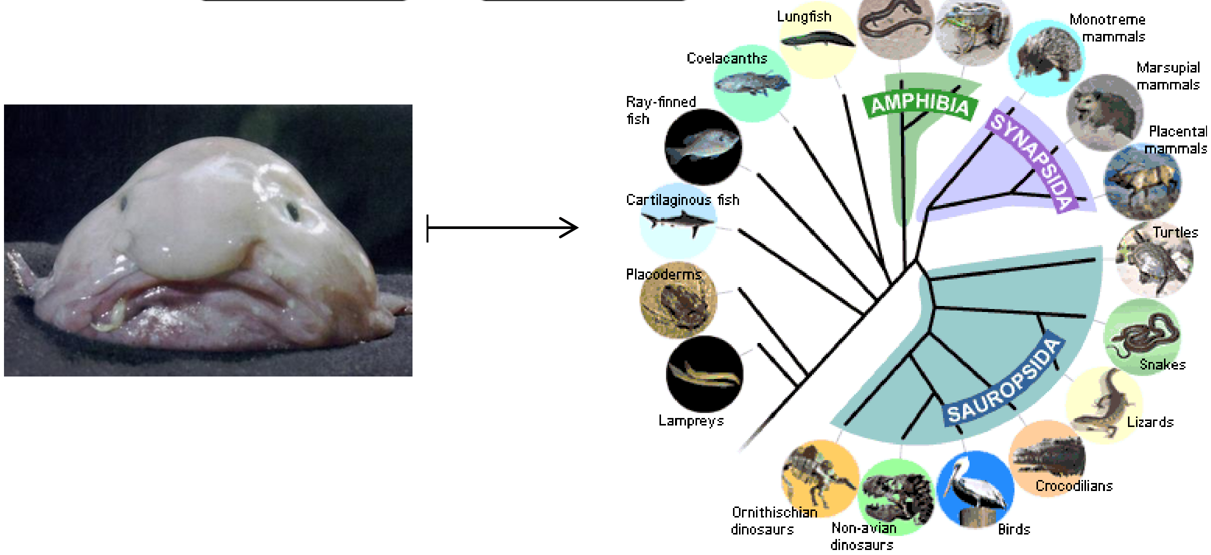
Intelligent Data Analysis

- Problem analysis, basic concepts.




Evaluation

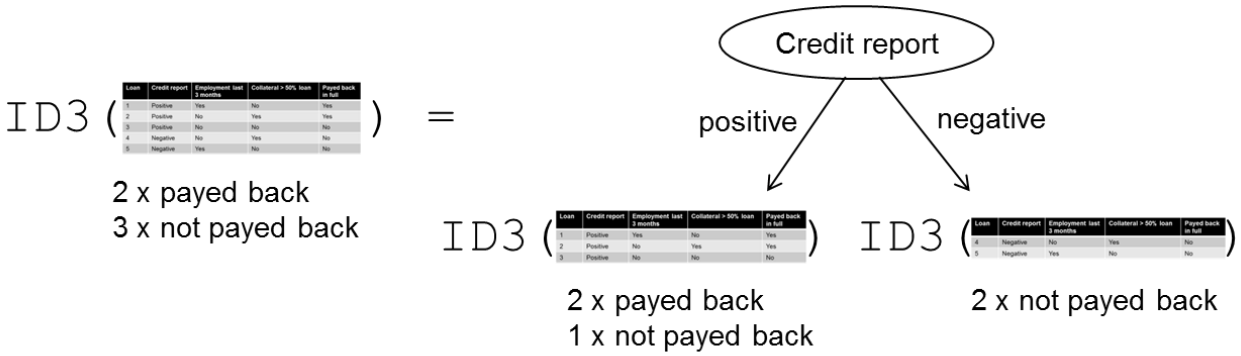
- ● Very Good
- ○ Good
- ○ Average
- ○ Bad
- ○ Very Bad



Intelligent Data Analysis

- Decision trees, random forests.

```
1. ID3(L)
    1. If all data in L have same class y, then return leaf
       node with class y.
    2. Else
        1. Choose attribute $x_j$ that separates L into subsets
           $L_1, \dots, L_k$ with most homogenous class distributions.
        2. Let $L_i = \{(x, y) \in L : x_j = i\}$.
        3. Return test node with attribute $x_j$ and children
           ID3($L_1$,), …, ID3($L_k$).
```



Loan	Credit report	Employment last 3 months	Collateral > 50% loan	Payed back in full
1	Positive	Yes	No	Yes
2	Positive	No	Yes	Yes
3	Positive	No	No	No
4	Negative	No	Yes	No
5	Negative	Yes	No	No

2 x payed back
3 x not payed back

Loan	Credit report	Employment last 3 months	Collateral > 50% loan	Payed back in full
1	Positive	Yes	No	Yes
2	Positive	No	Yes	Yes
3	Positive	No	No	No

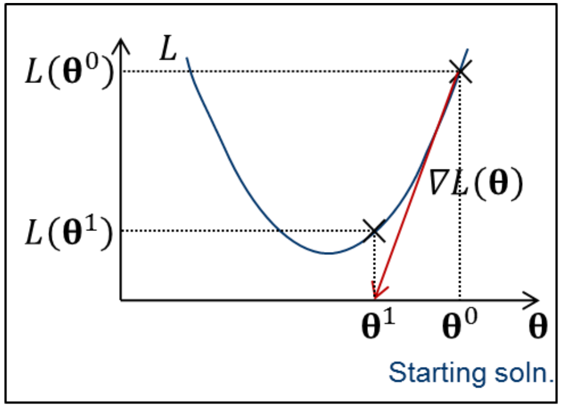
2 x payed back
1 x not payed back

Loan	Credit report	Employment last 3 months	Collateral > 50% loan	Payed back in full
4	Negative	No	Yes	No
5	Negative	Yes	No	No

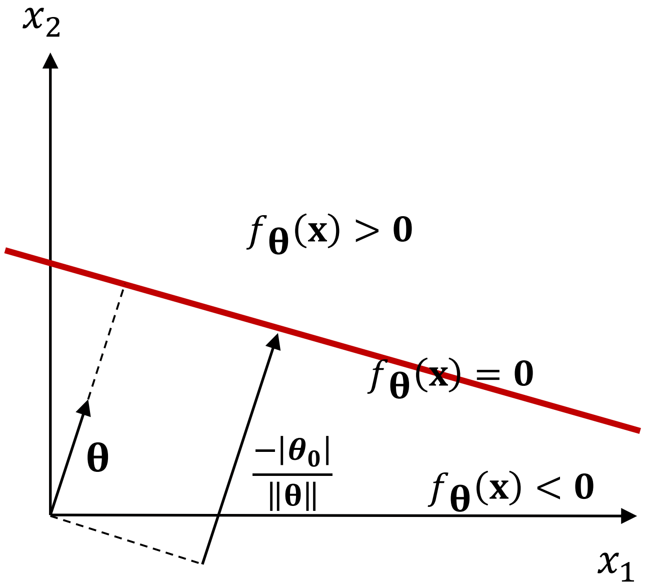
2 x not payed back

Intelligent Data Analysis

- Linear classification and regression models.



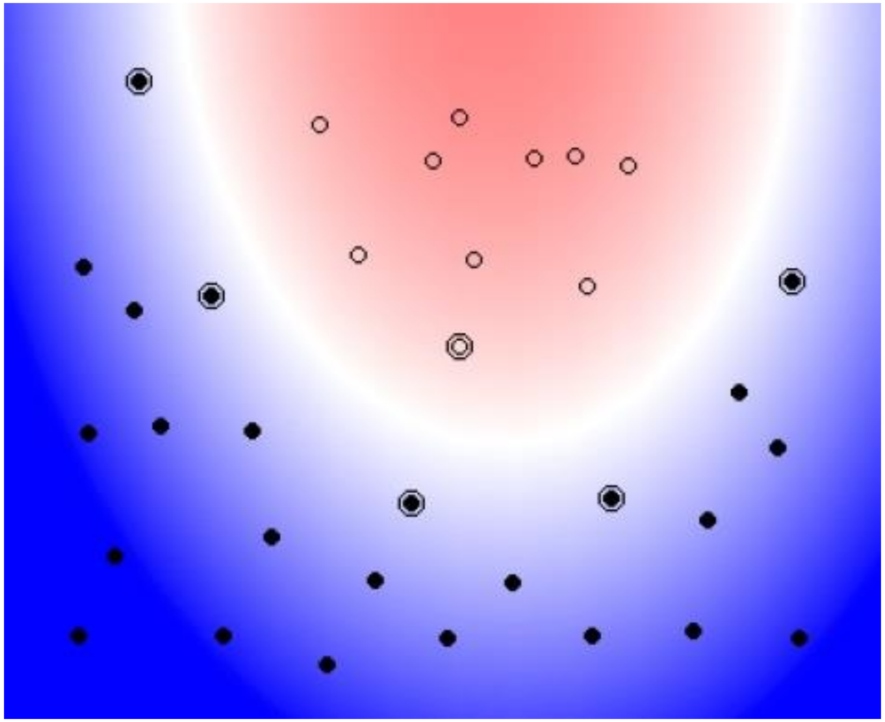
```
RegERM(Data: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$)
    Set $\boldsymbol{\theta}^0 = \mathbf{0}$ and $t = 0$
    DO
        Compute gradient $\nabla L(\boldsymbol{\theta}^t)$
        Compute step size $\alpha^t$
        Set $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \alpha^t \nabla L(\boldsymbol{\theta}^t)$
        Set $t = t + 1$
    WHILE $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t+1}\| > \varepsilon$
RETURN $\boldsymbol{\theta}^t$
```


x_2
$f_{\boldsymbol{\theta}}(\mathbf{x}) > \mathbf{0}$
$f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{0}$
$\boldsymbol{\theta}$
$\frac{-|\boldsymbol{\theta}_0|}{\|\boldsymbol{\theta}\|}$
$f_{\boldsymbol{\theta}}(\mathbf{x}) < \mathbf{0}$
x_1

Machine Learning

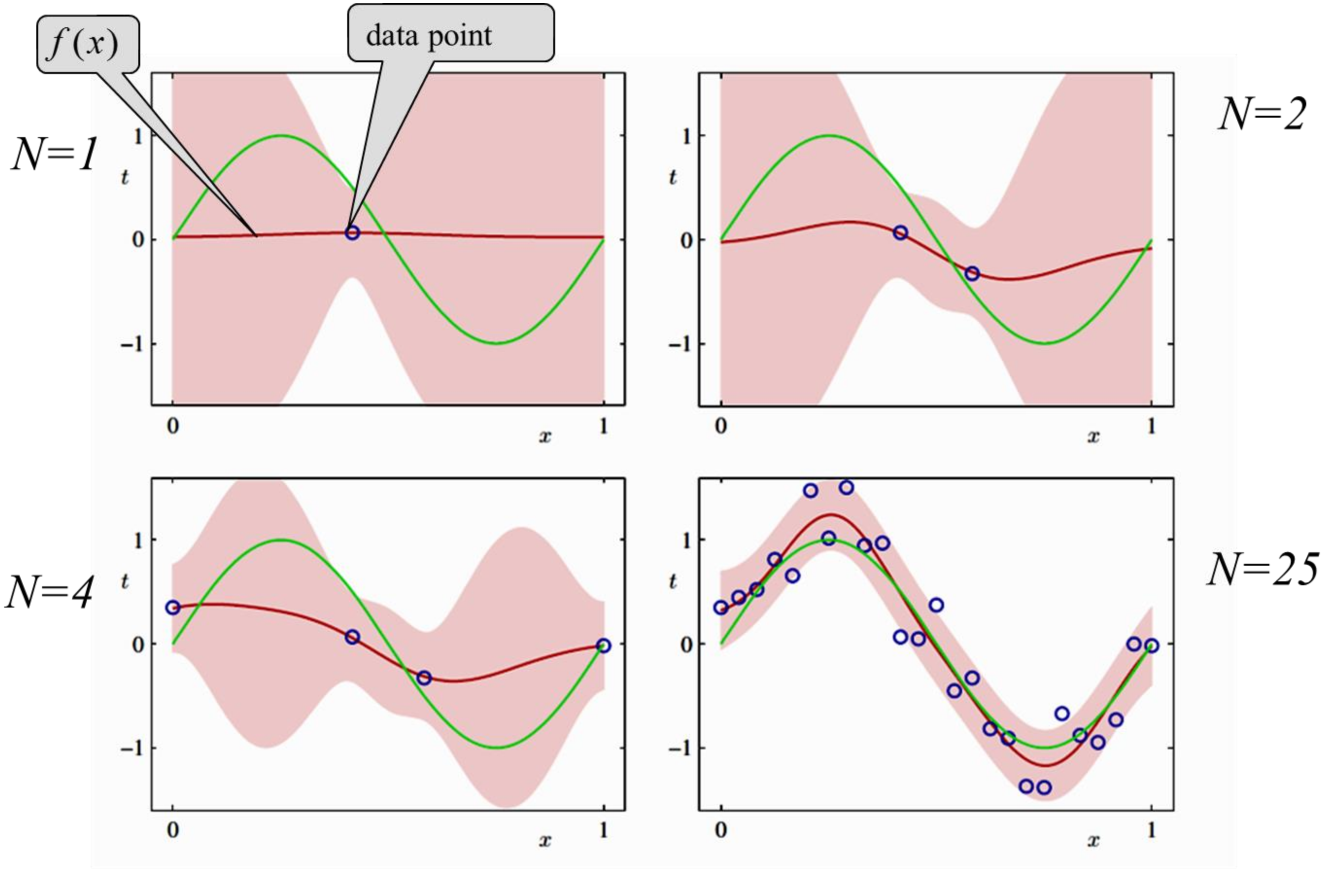
Intelligent Data Analysis

- Kernel methods.



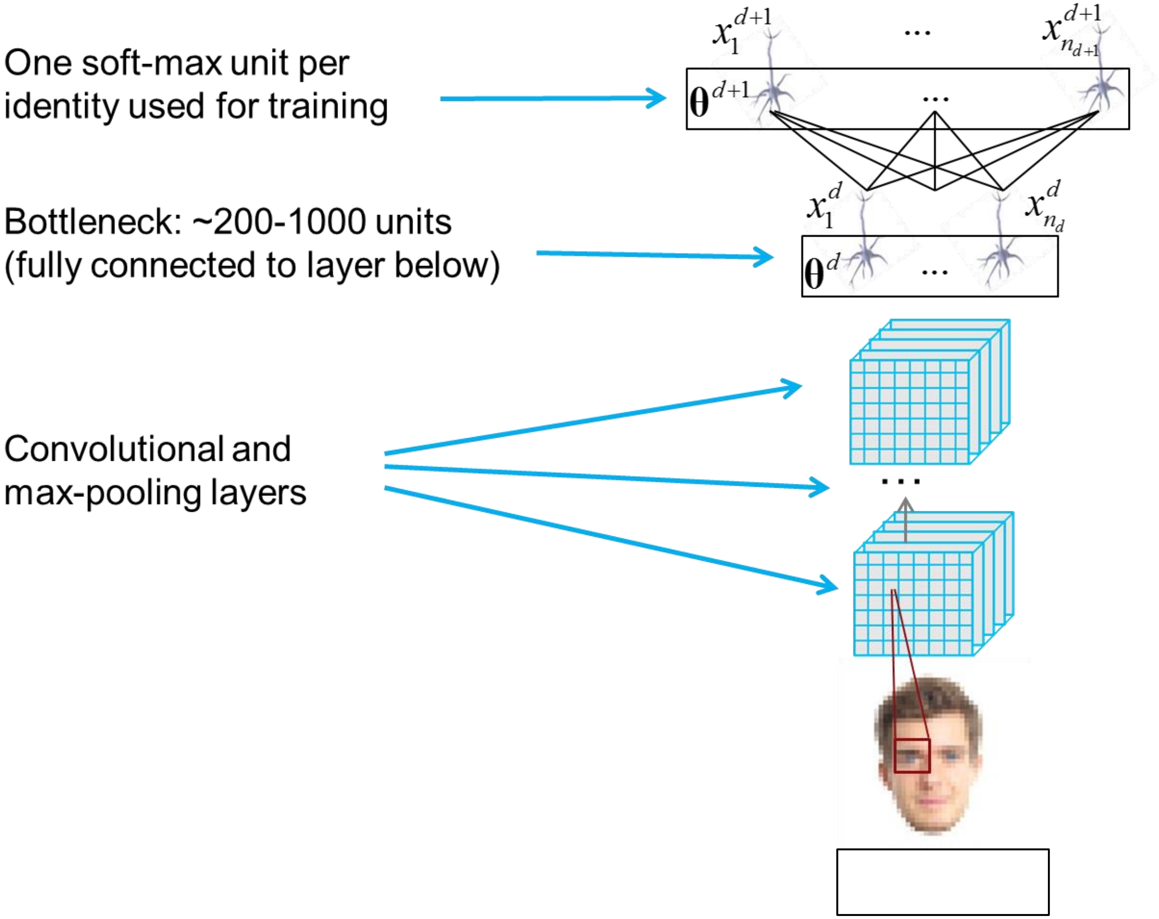
Intelligent Data Analysis

- Bayesian learning.



Intelligent Data Analysis

- Neural Networks.



One soft-max unit per identity used for training

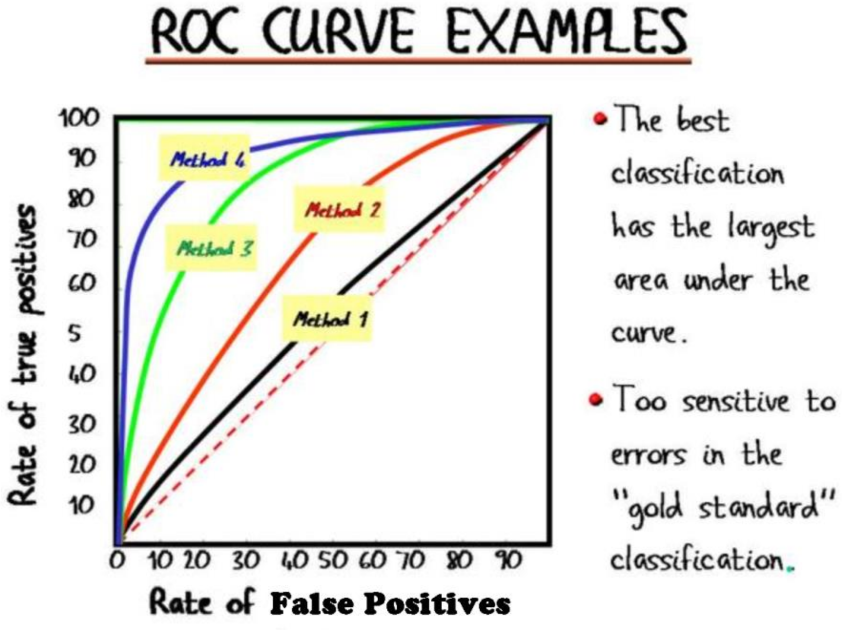
Bottleneck: ~200-1000 units (fully connected to layer below)

Convolutional and max-pooling layers

Machine Learning

Intelligent Data Analysis

- Model evaluation.



- The best classification has the largest area under the curve.
- Too sensitive to errors in the "gold standard" classification.