

Machine Learning for Determining the Credit Risk of Customers

Peter Geibel

TU Berlin

Faculty IV (Electrical Engineering and Computer Science)

Methods of Artificial Intelligence Research Group

Prof. Fritz Wysotzki

talk at SJTU, Aetna Management School, October 9

Overview

- Decision Support for Credit Risk Management
- German Credit Data
- Scoring Systems
- CAL5
- DIPOL

Basic Problem

- A private customer comes to a bank and applies for a loan.
- Should the credit be given to him?
- **Decision support** for the bank employee is needed:
 - he enters **description of customer and credit** into system:
 - income, age etc.
 - credit amount etc.
 - system **judges the credit risk** of the customer:
 - **classification**: good or bad customer
 - **OR**: rank, e.g. 1–5 (very bad – very good)
 - **OR**: prediction of gain/loss for bank
 - + **explanation**

Basic Problem

Not discussed in this talk:

- software engineering questions (embedding into larger system, connection to some database, programming etc.)
- collection and input of the data
- usage of the system by the bank employee

Basic Idea

Use the **past experience** of the bank (or of other banks) to automatically construct a decision support program.

machine learning terminology:

- past experience = **training set**
- automatically construct = **learn, induce**
- decision support program = **hypothesis, classifier**

find function $f : X_1 \times \dots \times X_n \rightarrow \{good, bad\}$ from I/O examples.

X_i is domain of **feature/attribute/variable** x_i .

German Credit Data Set

STATLOG project, Number of Instances: 1000

Number of Attributes: 20 (7 numerical, 13 categorical)

Attribute x_1 : (qualitative)

Status of existing checking account

A11 : ... < 0 DM

A12 : 0 <= ... < 200 DM

A13 : ... >= 200 DM /

salary assignments for at least 1 year

A14 : no checking account

i.e. $X_1 = \{1, \dots, 4\}$

German Credit Data Set

Attribute 2: (numerical)

Duration in month, *i.e.* $X_2 = real$

Attribute 3: (qualitative)

Credit history

A30 : no credits taken/
all credits paid back duly

A31 : all credits at this bank paid back duly

A32 : all credits at this bank paid back duly

A33 : existing credits paid back duly till now

A34 : delay in paying off in the past

A35 : critical account/
other credits existing (not at this bank)

A36 : other credits existing (not at this bank)

German Credit Data Set

Attribute 4: (qualitative)

Purpose

A40 : car (new)

A41 : car (used)

A42 : furniture/equipment

A43 : radio/television

A44 : domestic appliances

A45 : repairs

A46 : education

A47 : (vacation - does not exist?)

A48 : retraining

A49 : business

A410 : others

German Credit Data Set

Attribute 5: (numerical)

Credit amount

Attribute 6: (qualitative)

Savings account/bonds

A61 : ... < 100 DM

A62 : 100 <= ... < 500 DM

A63 : 500 <= ... < 1000 DM

A64 : .. >= 1000 DM

A65 : unknown/ no savings account

German Credit Data Set

Attribute 7: (qualitative)

Present employment since

A71 : unemployed

A72 : ... < 1 year

A73 : 1 <= ... < 4 years

A74 : 4 <= ... < 7 years

A75 : .. >= 7 years

Attribute 8: (numerical)

Installment rate in percentage of disposable income

German Credit Data Set

Attribute 9: (qualitative)

Personal status and sex

A91 : male : divorced/separated

A92 : female : divorced/separated/married

A93 : male : single

A94 : male : married/widowed

A95 : female : single

Attribute 10: (qualitative)

Other debtors / guarantors

A101 : none

A102 : co-applicant

A103 : guarantor

German Credit Data Set

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Property

A121 : real estate

A122 : if not A121 : building society savings
agreement/life insurance

A123 : if not A121/A122 : car or other

A124 : unknown / no property

Attribute 13: (numerical)
Age in years

German Credit Data Set

Attribute 14: (qualitative)

Other installment plans

A141 : bank

A142 : stores

A143 : none

Attribute 15: (qualitative)

Housing

A151 : rent

A152 : own

A153 : for free

German Credit Data Set

Attribute 16: (numerical)

Number of existing credits at this bank

Attribute 17: (qualitative)

Job

A171 : unemployed/ unskilled - non-resident

A172 : unskilled - resident

A173 : skilled employee / official

A174 : management/ self-employed/
highly qualified employee/ officer

German Credit Data Set

Attribute 18: (numerical)

Number of people being liable to provide maintenance for

Attribute 19: (qualitative)

Telephone

A191 : none

A192 : yes, registered under the customers name

Attribute 20: (qualitative)

foreign worker

A201 : yes

A202 : no

Data Set

A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1
A12 48 A32 A43 5951 A61 A73 2 A92 A101 2 A121 22 A143 A152 1 A173 1 A191 A201 2
A14 12 A34 A46 2096 A61 A74 2 A93 A101 3 A121 49 A143 A152 1 A172 2 A191 A201 1
A11 42 A32 A42 7882 A61 A74 2 A93 A103 4 A122 45 A143 A153 1 A173 2 A191 A201 1
A11 24 A33 A40 4870 A61 A73 3 A93 A101 4 A124 53 A143 A153 2 A173 2 A191 A201 2
A14 36 A32 A46 9055 A65 A73 2 A93 A101 4 A124 35 A143 A153 1 A172 2 A192 A201 1
A14 24 A32 A42 2835 A63 A75 3 A93 A101 4 A122 53 A143 A152 1 A173 1 A191 A201 1
A12 36 A32 A41 6948 A61 A73 2 A93 A101 2 A123 35 A143 A151 1 A174 1 A192 A201 1
A14 12 A32 A43 3059 A64 A74 2 A91 A101 4 A121 61 A143 A152 1 A172 1 A191 A201 1
A12 30 A34 A40 5234 A61 A71 4 A94 A101 2 A123 28 A143 A152 2 A174 1 A191 A201 2
A12 12 A32 A40 1295 A61 A72 3 A92 A101 1 A123 25 A143 A151 1 A173 1 A191 A201 2

1: good customer

2: bad customer

Scoring Systems

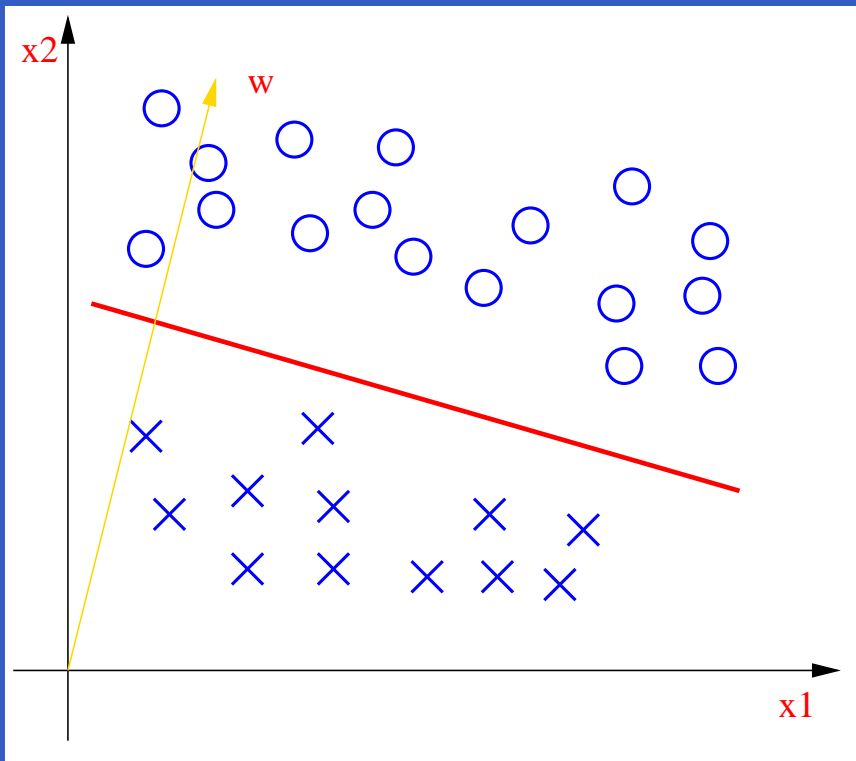
Given: attributes (variables) x_1, \dots, x_n and weights w_1, \dots, w_n

Compute $w_1x_1 + \dots + w_nx_n$ and compare to threshold θ .

- if $w_1x_1 + \dots + w_nx_n > \theta$ then class “good”
- if $w_1x_1 + \dots + w_nx_n < \theta$ then class “bad”
- perhaps: if $w_1x_1 + \dots + w_nx_n \approx \theta$ then “no decision possible”
- can be extended to more than 2 classes

How to determine the weights?

Scoring Systems



Classes are separated by hyperplane $\frac{w}{|w|}x - \frac{\theta}{|w|} = 0$ with $x = (x_1, \dots, x_n)$ and $w = (w_1, \dots, w_n)$

Statistical Techniques for the Weights

Discriminant Analysis:

- Bayes-optimal decisions (Fisher discriminant)

Regression models:

- minimize

$$\sum_i (k_i - (wx - \theta))^2$$

wrt. to w and θ

$k_i \in \{-1, 1\}$ target value for example i , e.g. 1 = good, and
-1 = bad

Bayes optimal decisions

The **conditional density** is $p(x|k) = p(x, k)/p(k)$.

$p(k)$ is the **a priori density** of the classes and $p(x, k)$ is the **common density** of patterns and classes.

Bayes Rule: Decide for class k_j , if

$$p(k_j|x) > p(k_i|x)$$

equivalently: $p(k_j)p(x | k_j) > p(k_i)p(x | k_i)$

Bayes optimal decisions

1-Dim. Case:

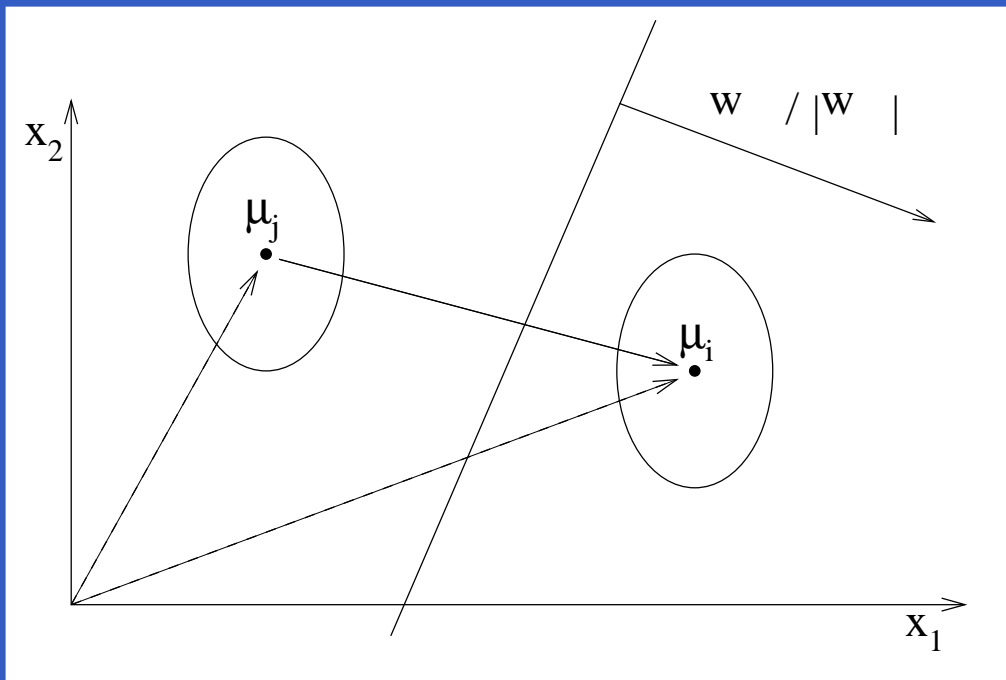
Assume that all classes have a Gaussian distribution:

$$p(x | k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

Decide for class j , if

$$2 \log\left(\frac{p(j)}{p(i)}\right) - \log\left(\frac{\sigma_i}{\sigma_j}\right) - \frac{(x - \mu_j)^2}{\sigma_j^2} + \frac{(x - \mu_i)^2}{\sigma_i^2} > 0 \quad \forall i \neq j$$

Bivariate Gaussian Distribution



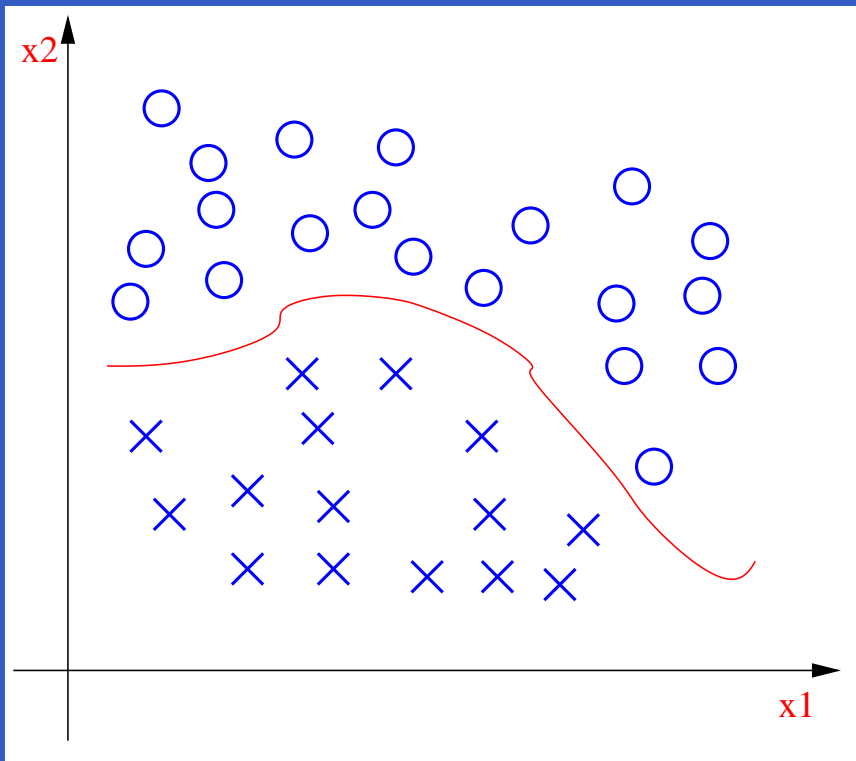
2-Dim. Case:

equal covariance matrices

...

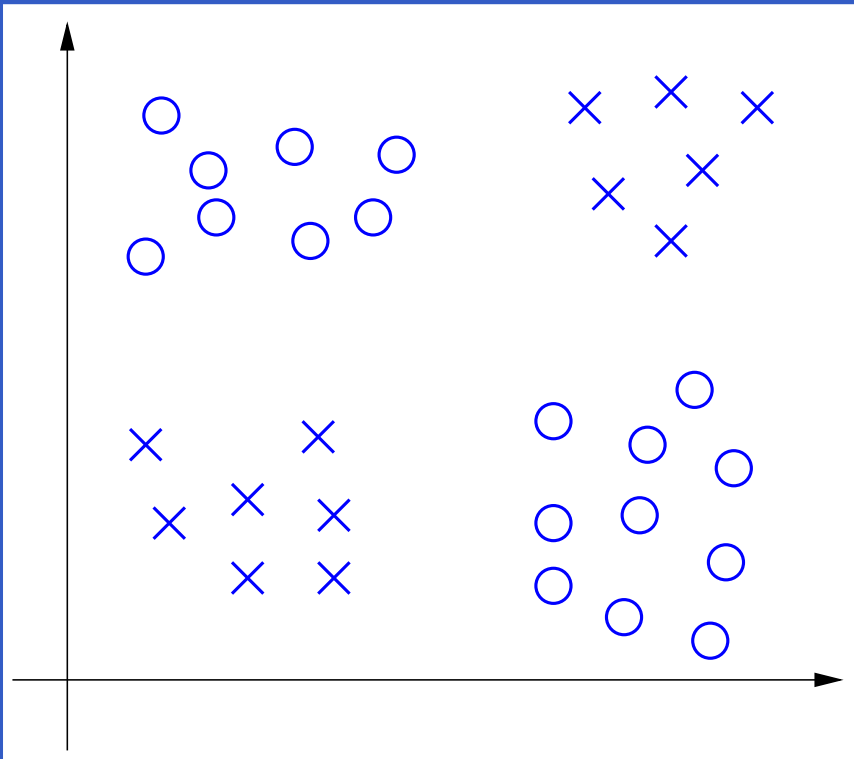
weight vector w for scoring systems can be determined explicitly

Scoring Systems – Problems



only **non-linear** separation possible!

Scoring Systems – Problems



How to treat **multi-modal** distributions?

Intermediate Summary

statistical methods for determining weights for scoring systems:

- no assumption on distribution
- non-linear separations of classes possible
- can treat multi-modal distributions

CAL5

- decision tree algorithm
- interpretable results

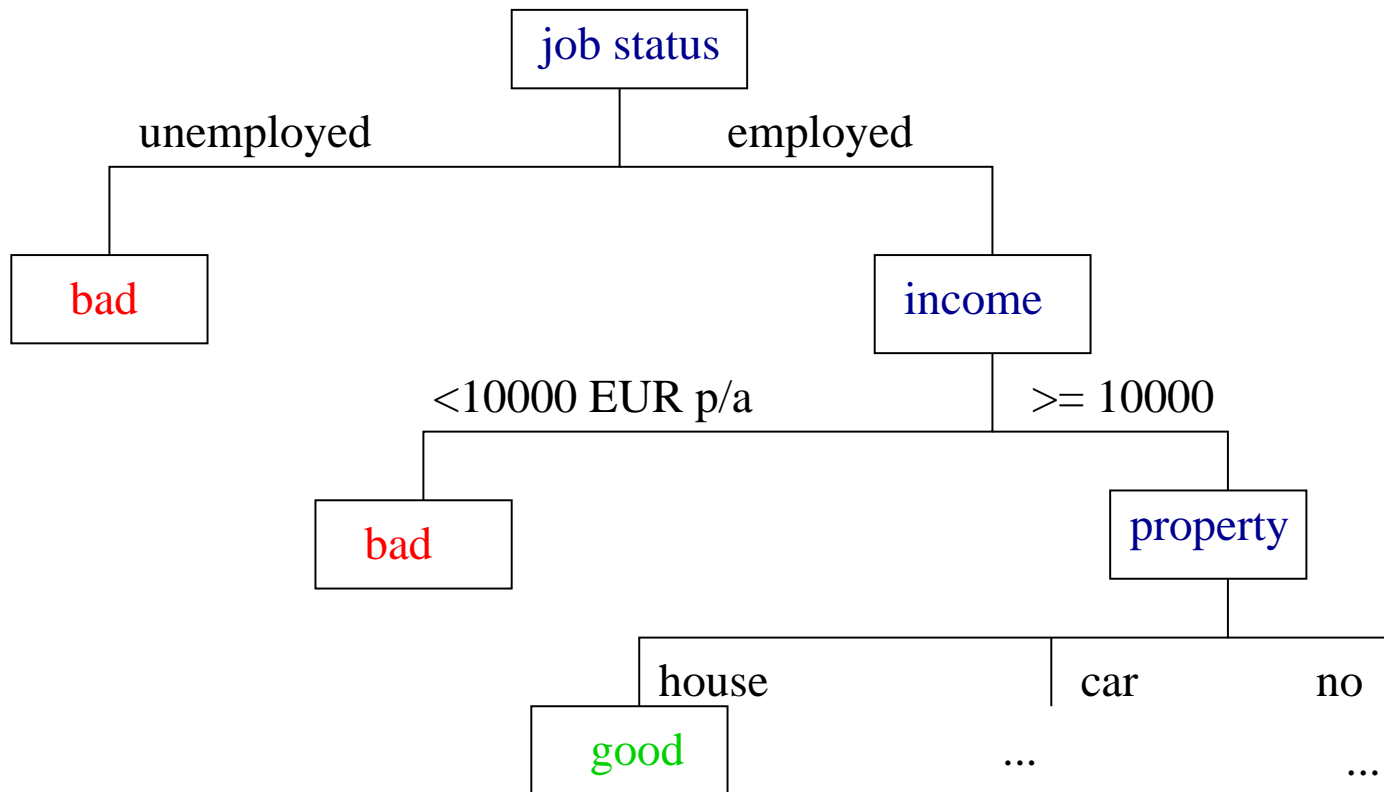
DIPOL

- neural network
- can be seen as extension of scoring method

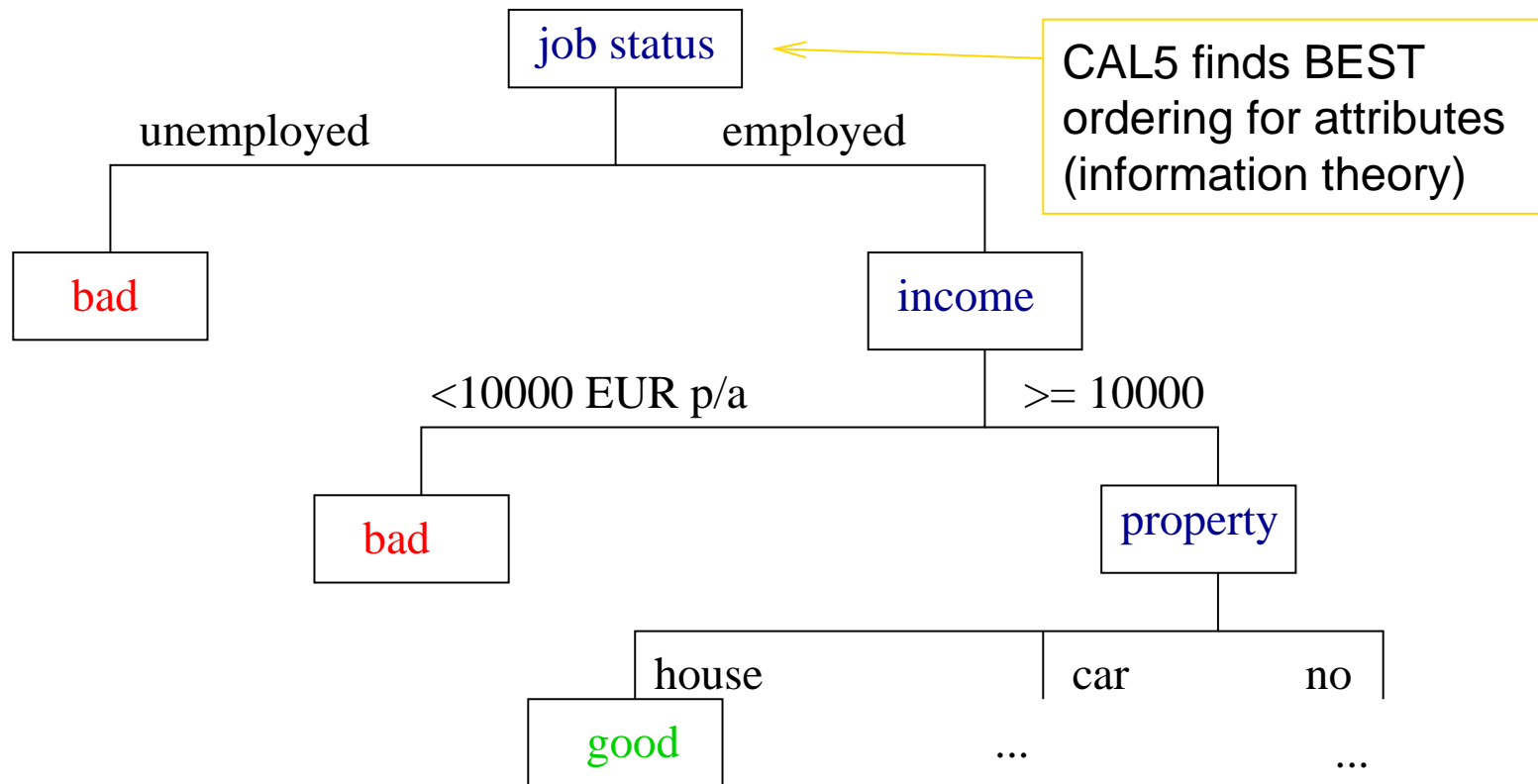
CAL5

- **decision tree** algorithm for continuous and categorical attributes
- invented by Fritz Wysotzki in 1981
- STATLOG project: performs well on the credit risk problem
- automated handling of data errors: **noise** is suppressed
- extension to **missing values** possible
- understandable rules can be extracted from the tree

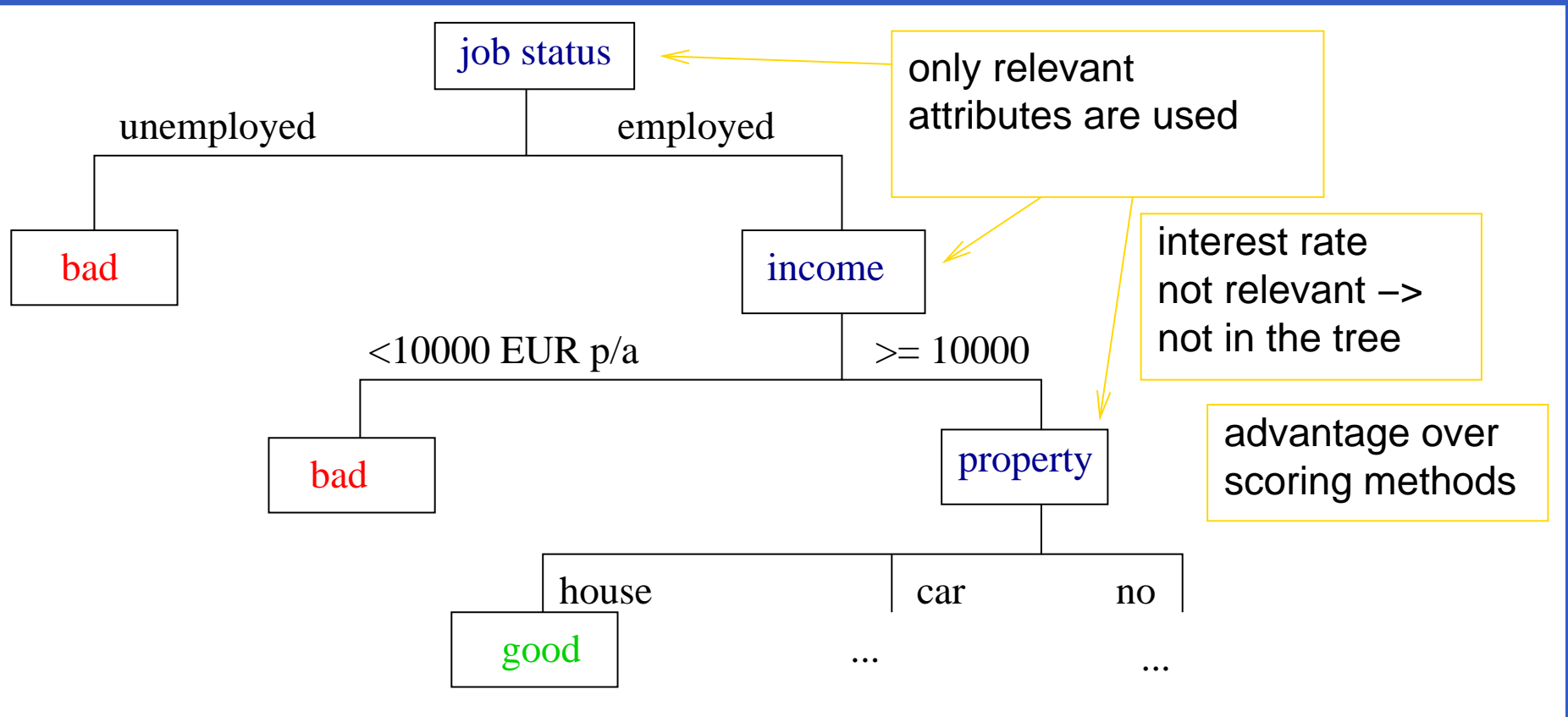
CAL5



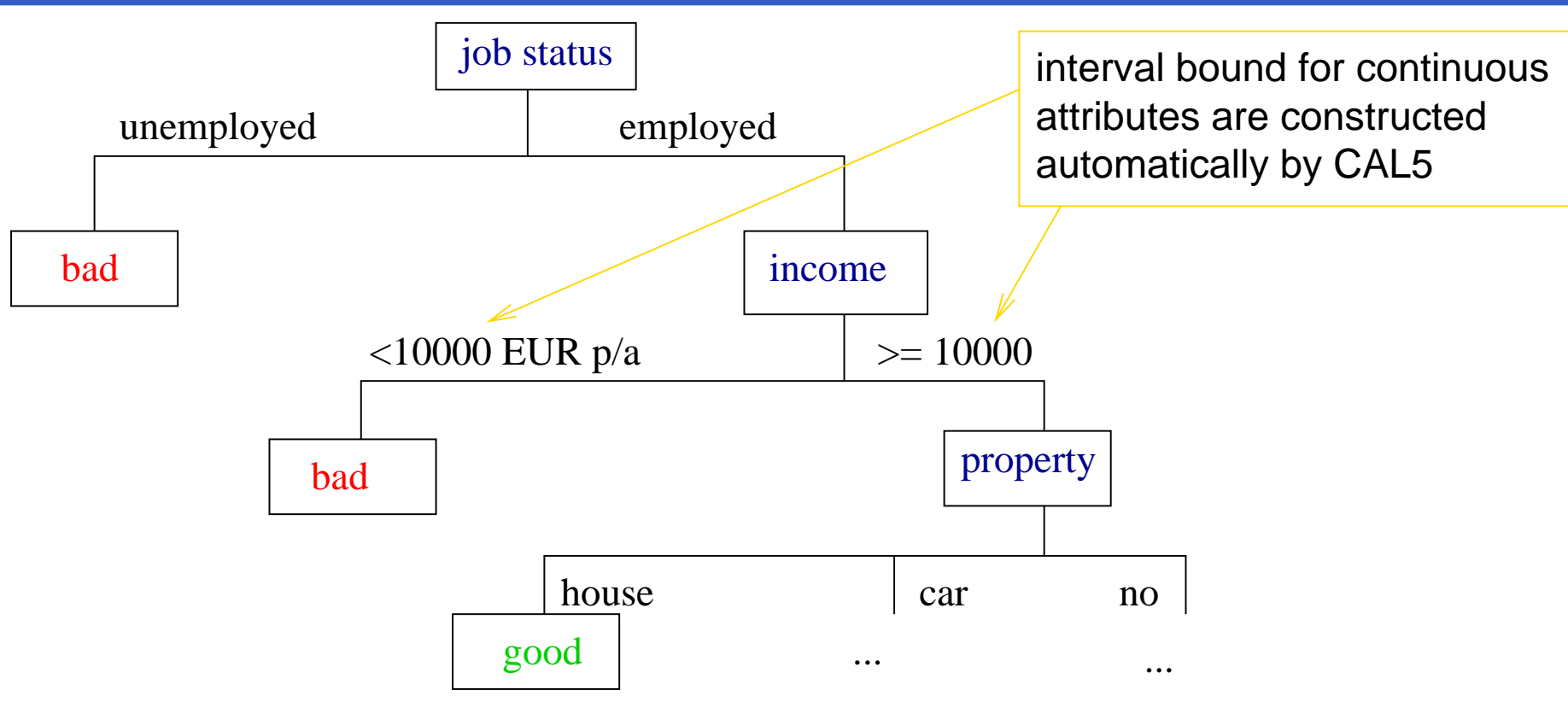
CAL5



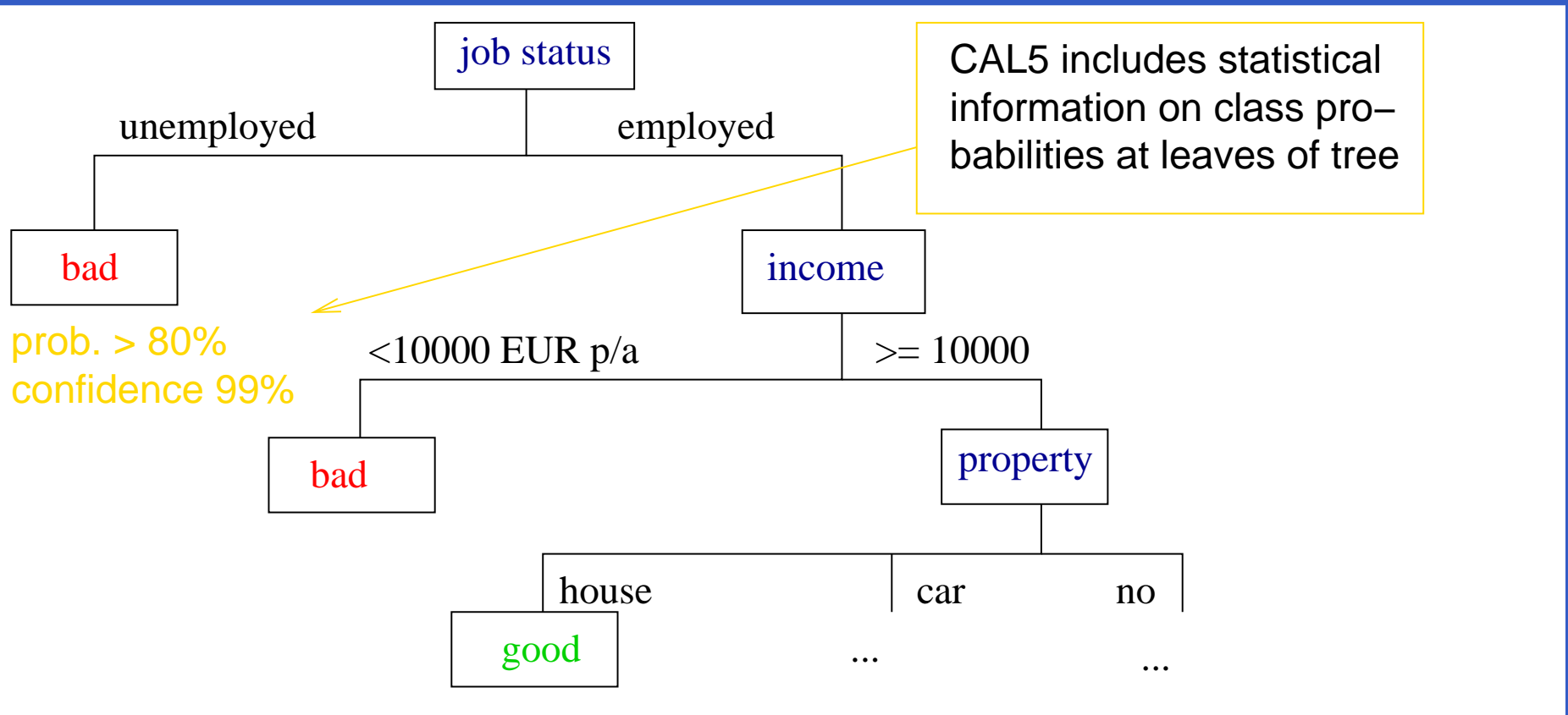
CAL5



CAL5



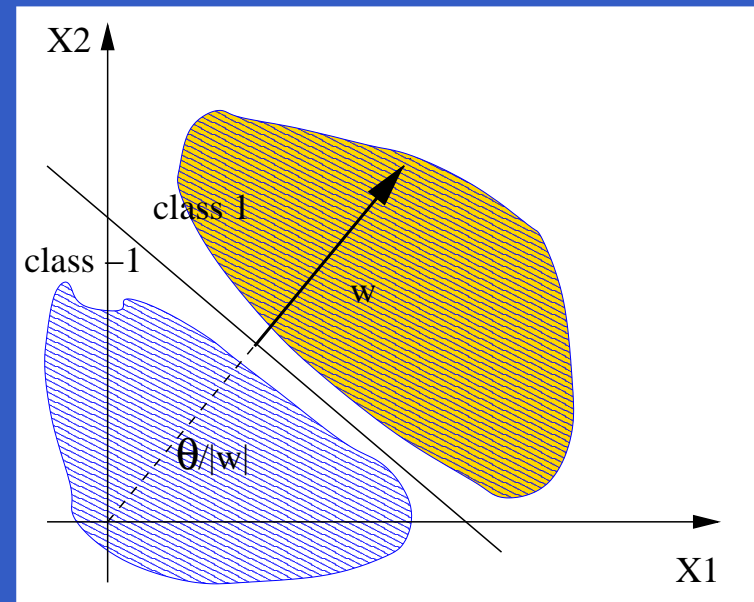
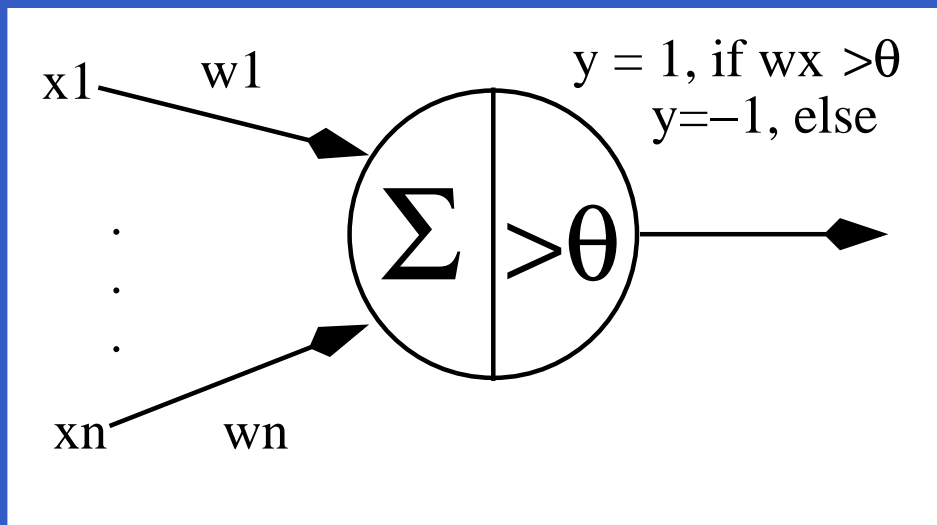
CAL5



DIPOL

- B. Schulmeister and F. Wysotzki, 1994
- extension of perceptron algorithm
- perceptron \approx learning weights for scoring system
 - perceptron is successfully applied in German Bank (Deutsche Postbank)
 - saves several Mio. EUR
 - good customers vs. **hard to distinguish** bad customers
 - obviously bad customers are a priori discarded

Perceptron

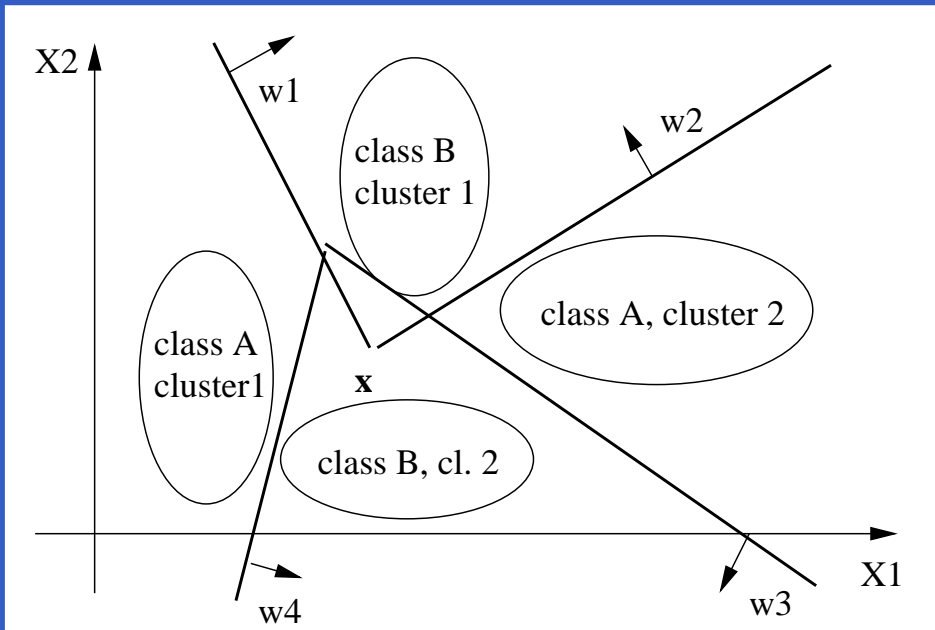


incremental learning algorithm (distribution free)

$w := w + x$ or $w := w - x$ in case of error

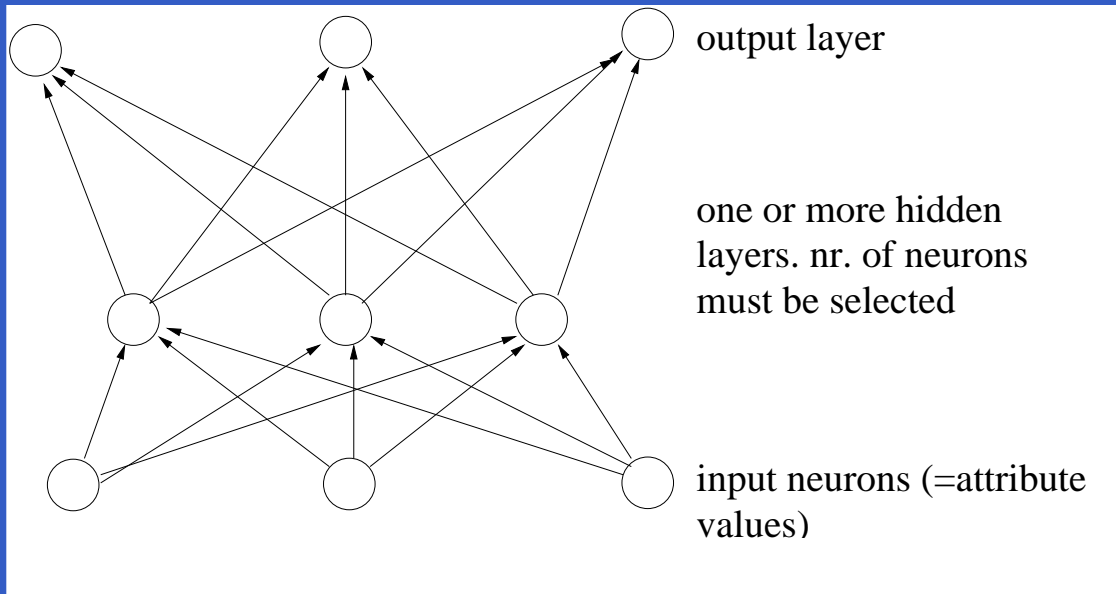
can be extended to linearly non-separable classes

Two Classes/Two Clusters



one hyperplane per pair of classes/clusters

Neural Architecture for DIPOL

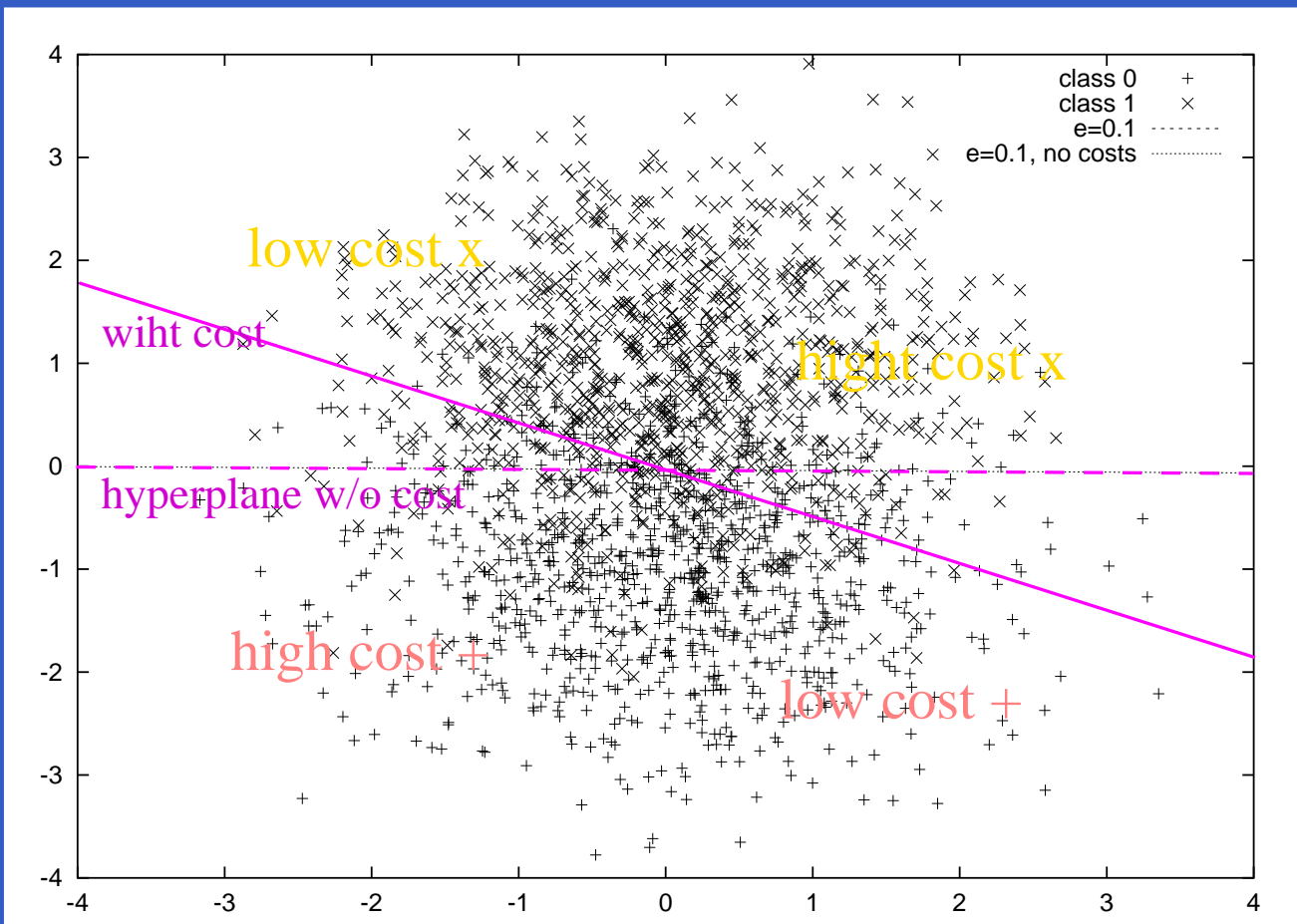


output layer= classes

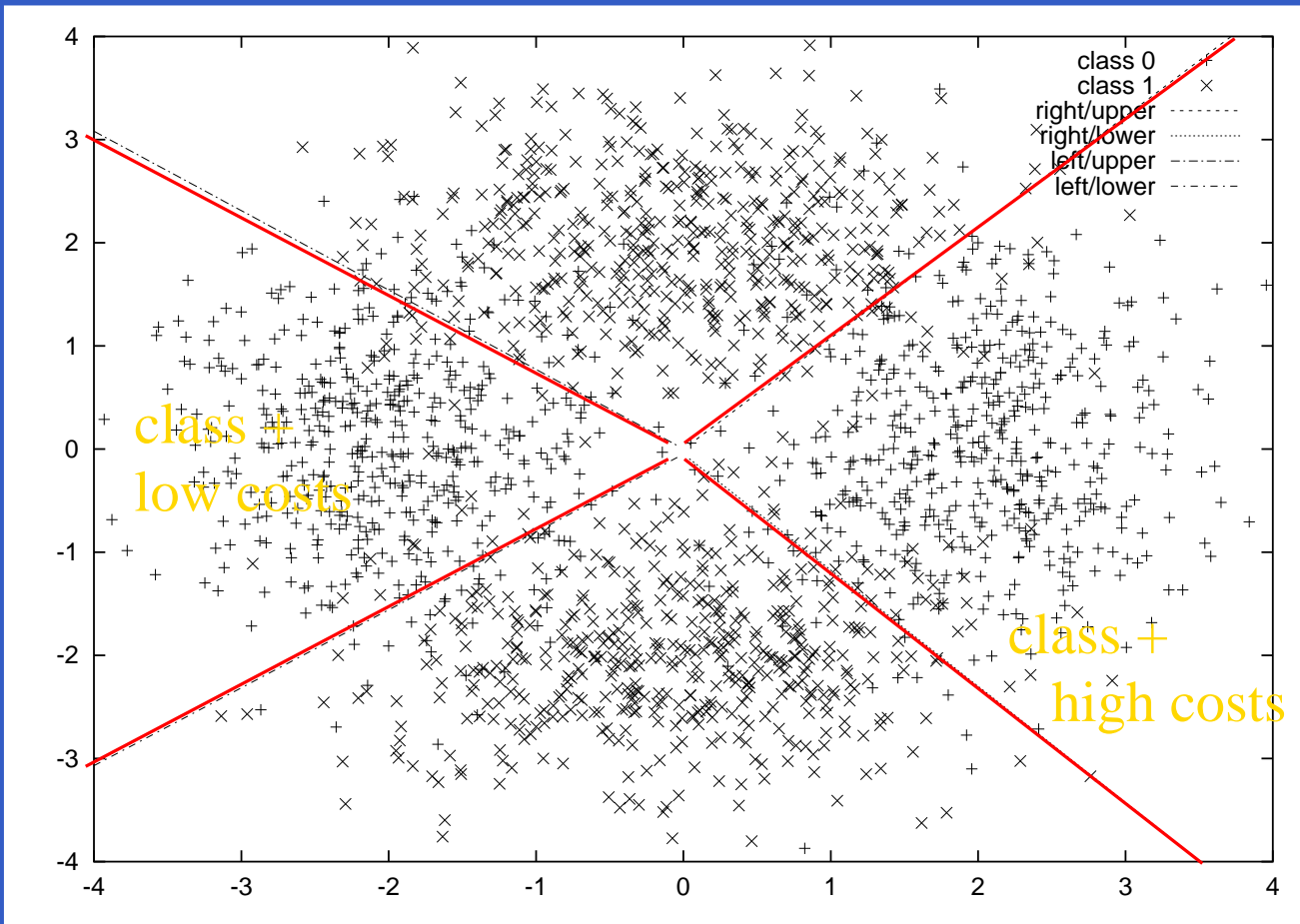
input layer= attribute values

hidden layer= hyperplanes

Costs



Costs



Summary

- scoring methods suffer from several drawbacks:
 - weight adaptation by hand or
 - assumptions on distribution
 - only linear
 - only unimodal
- machine learning can help:
 - CAL5 learns decision trees, interpretable rules
 - DIPOL learns neural networks. Increased Accuracy
- machine learning saves money

Application in Chinese Banks

- CAL5 and DIPOL are **general methods** for data analysis
- they are **not specialized** for german data set
- CAL5 and DIPOL incorporate techniques to deal with **data errors**
- its possible to deal with **incomplete information**
- CAL5 and DIPOL can be used in Chinese Banks