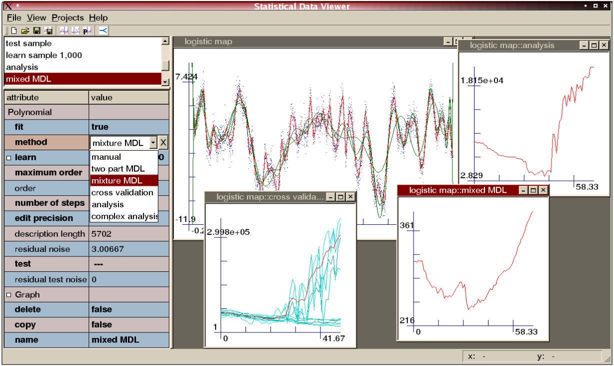
The Paradox of Overfitting

Volker Nannen

February 1, 2003

Artificial Intelligence

Rijksuniversiteit Groningen


Statistical Data Viewer
File View Projects Help
test sample
learn sample 1,000
analysis
mixed MDL
attribute
value
Polynomial
fit
true
method
mixture MDL
learn
manual
two part MDL
mixture MDL
cross validation
analysis
complex analysis
maximum order
order
number of steps
edit precision
description length
5702
residual noise
3.00667
test
---
residual test noise
0
Graph
delete
false
copy
false
name
mixed MDL
logistic map
7.424
-11.9
logistic map::analysis
1.815e+04
2.829
58.33
logistic map::cross valida...
2.998e+05
1
0
41.67
logistic map::mixed MDL
361
216
0
58.33
x: -
y: -

Contents

MDL – theory

The paradox of overfitting:

Complex models contain more
information on the training data

but less information on future data.

Machine learning uses models

to describe reality.

Models can be

- statistical distributions
- polynomials
- Markov chains
- neural networks
- decision trees
- etc.

This work uses polynomial models.

$$m_k = p_k(x) = a_0 + \cdots + a_k\, x^k \tag{1}$$

Polynomials are

- well understood
- used throughout mathematics
- suffer badly from overfitting

The mean squared error of a model m on a sample

$$s = \{(x_1, y_1) \ldots (x_n, y_n)\} \tag{2}$$

of size n is

$$\sigma_f^2 \ = \ \frac{1}{n} \sum_{i=0}^{n} \left(m(x_i) - y_i\right)^2 \tag{3}$$

The error on the training sample is called

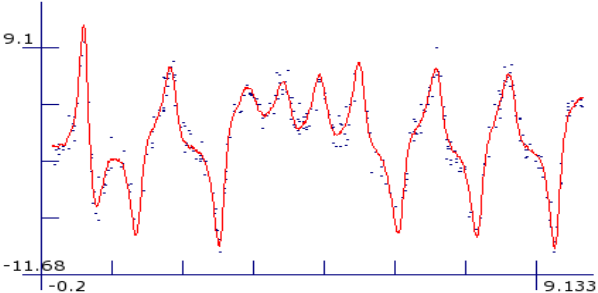
<u>training error</u>.

The error on future samples is called

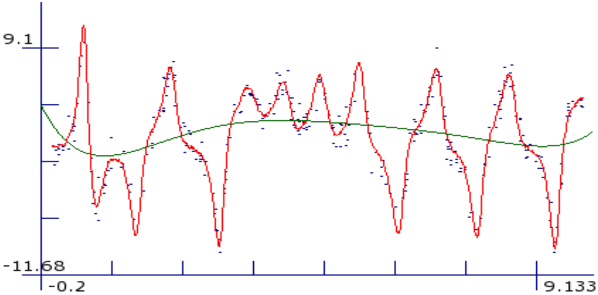
<u>generalization error</u>.

We want to minimize the generalization error.

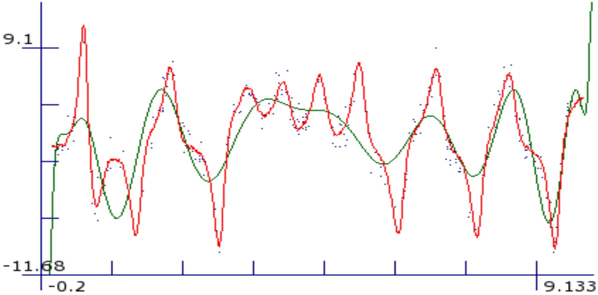
An example of overfitting: regression in the two-dimensional plane



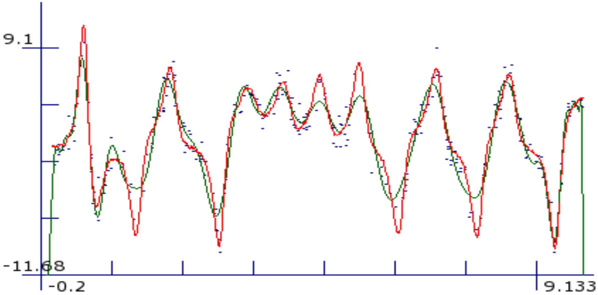
Continuous signal + noise,
300 point sample.



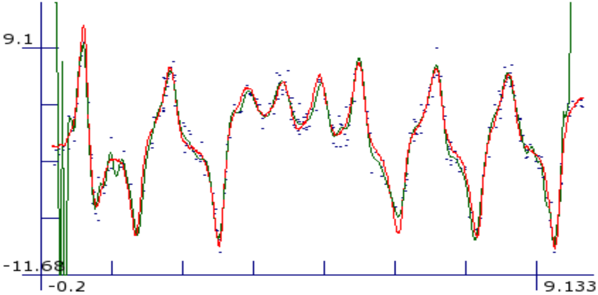
6 degree polynomial, $\sigma^2 = 13.8$



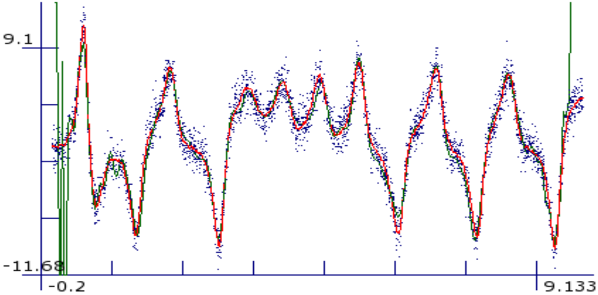
17 degree polynomial, $\sigma^2 = 5.8$



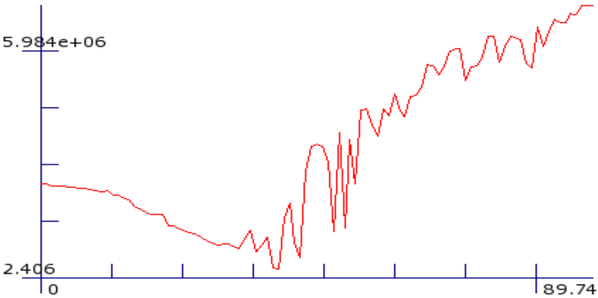
43 degree polynomial, $\sigma^2 = 1.5$



100 degree polynomial, $\sigma^2 = 0.6$



3,000 point test sample. $\sigma_t^2 = 10^{12}$



Generalization error on this 3,000 point test sample.

6 degree:	$\sigma^2 = 16$,	17 degree:	$\sigma^2 = 8.6$,
43 degree:	$\sigma^2 = 2.7$,	100 degree:	$\sigma^2 = 10^{12}$.

Rissanens hypothesis:

Minimum Description Length

prevents overfitting.

MDL minimizes the code length

$$\min_m \left[l(s|m) + l(m) \right] \tag{4}$$

This is a two-part code:

$l(m)$ is the code length of the model

and $l(s|m)$ is the code length of the data given the model.

We only look at the least square model per degree

$$\min_{k} \left[n \log \hat{\sigma}_{m_k} + \ l(m) \right] \tag{5}$$

Rissanen's original estimation:

$$\min_{k} \left[n \log \hat{\sigma}_{m_k} + k \log \sqrt{n} \right] \tag{6}$$

This is too weak.

Mixture MDL is a modern version of MDL.

$$\min_k \left[-\log \int_{m_k \in M_k} p(M_k = m_k)\, p(s|m_k)\; d\, m_k \right] \quad (7)$$

$p(M_k = m_k)$ is a prior distribution over models in M_k.

Barron & Liang provide a simple algorithm based on the uniform prior (2002).

Experimental Verification

Problems with experiments on model selection:

- shortage of appropriate data
- inefficient setup of experiments
- insufficient visualization
- few tangible results

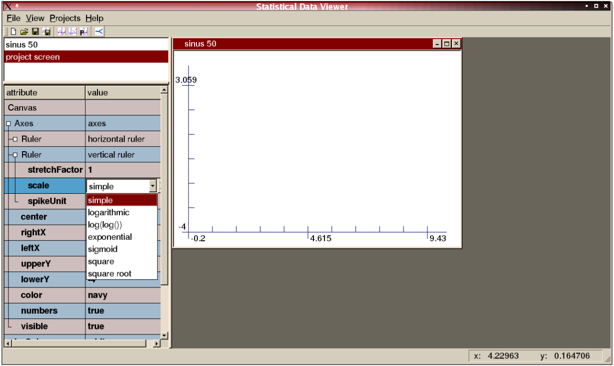
Solution:

The Statistical Data Viewer

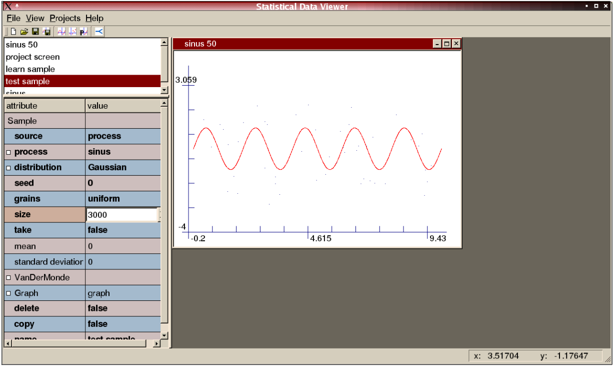
an advanced tool
for statistical experiments.

A simple experiment:

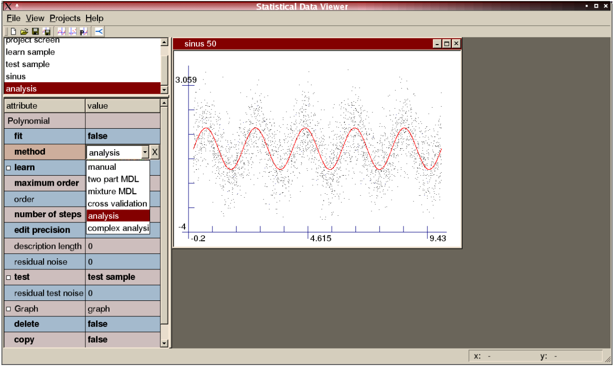
the sinus wave



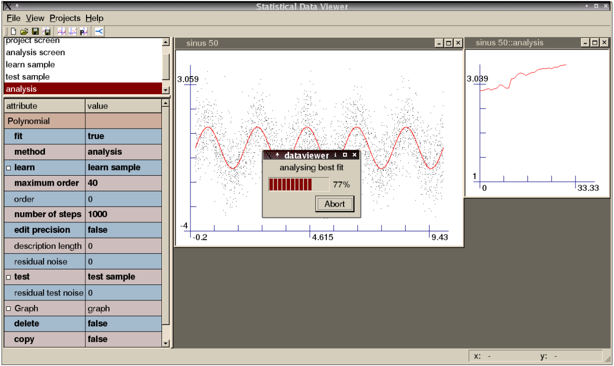
A new project



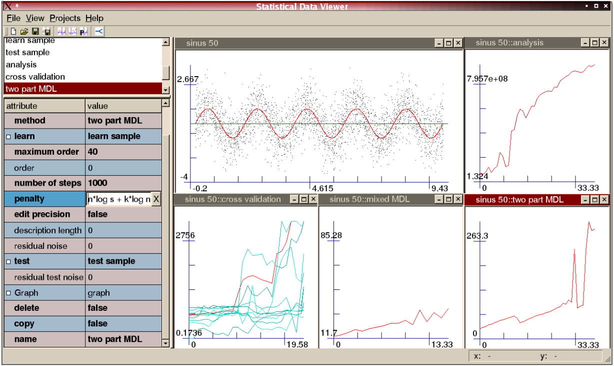
A new process and sample



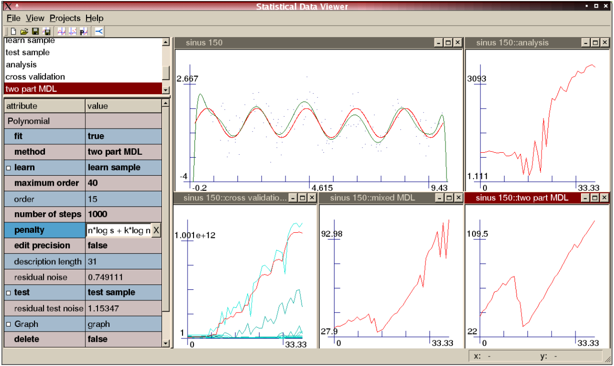
Selecting a method for a model



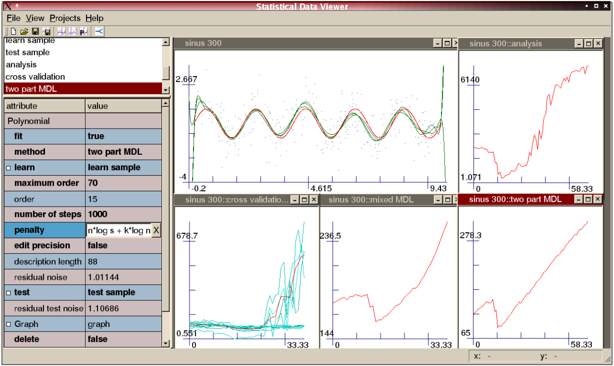
Analyzing the generalization error



Analysis, cross validation, mixture MDL and Rissanen's MDL. Optimum at 0 degrees.



150 point sample. Optimum at 17 degrees.



300 point sample. Optimum at 18 degrees.

Results

Achievements:

- generic problem space (files, broad selection of online signals, drawing by hand)
- graphical object oriented setup of experiments (no scripting)
- graphics integrated into the control structure
- simple programming interfaces

Conclusion for all experiments:

- Rissanens original version usually overfits.
- Mixture MDL can prevent overfitting.
- smoothing is important for model selection.
- Mixture MDL cannot deal with non-uniform support. (but cross validation can do it!)
- Mixture MDL can deal with different types of noise. (i.i.d. assumption can be relaxed!)
- The structure of a prediction graph contains valuable information by itself and MDL can reproduce it.

Further research:

- The structure of the generalization error
- Other types of data
- Other types of models
- Improved interfaces

volker.nannen.com/mdl