Model validation

T-61.6040, 03.10.2006, Heli Hiisilä

Testing Neural Models: How to Use Re-Sampling Techniques?
A. Lendasse
&
Fast bootstrap methodology for model selection,
A. Lendasse, G. Simon, V. Wertz, M. Verleysen,
Neurocomputing, Elsevier, Vol. 64,
March 2005, pp. 161-181.

Problem setting

- The model is often chosen a priori based on the user experience
  - Sometimes bad choices
- How to select objectively the best model when the amount of data is limited?
  - Cross validation and bootstrap methods
  - Fast bootsrap methodology

Learning and validation sets

- Data is split into learning and validation sets (different ways in different methods)
  - Learning set is used to calculate the optimal parameters of a model once structure is established
  - Validation set is used to select specific structure in a family
    - Also test set should be used to estimate correctly the generalization error once model is built

Learning error

- The unknown relation between the input and output data is $y_t = g(x_t) + e_t$
- Model: $\hat{y}_t = h^q(x_t, \theta(q))$
  - – q identifies model structure and
  - – θ represents the model parameters
- Learning error:

$$E_{learn}(q,\theta) = \frac{\sum_{t=1}^{N} \left(h^q(x_t, \theta(q)) - y_t\right)^2}{N}$$

Genralization error

- An infinite number of data is never available
  - Generalization error is approximated on the basis of mean value derived from a finite set of validation data
  - $\theta^*(q)$ is the best set of model parameters

$$E_{gen}(q,\theta^*(q)) = \frac{\sum_{t=1}^{M}\left(h^q(x_t,\theta^*(q)) - y_t\right)^2}{M}$$

Validation methods

- Simple validation
- Monte-Carlo Cross-validation
- K-fold Cross-Validation
- Leave-one-out
- Bootstrap and Bootstrap variants

Simple validation

- Divide data to learning and validation sets
  - keep 2/3 of the data in the learning set and 1/3 in the validation set
- Learning and validation sets are selected at random among the available data
- Results depend closely on the random split of the data
  - Bad performance, still used in many cases
  - Should not be used unless maybe when the number of available data is very large with respect to the dimension of the space

Monte-Carlo Cross-Validation

- Take several random drawings of learning and validation sets and evaluate the mean of generalization error
  - Do for example 10, 100 or 1000 random drawings
  - Generalization error estimate is the mean of the errors computed on the validation sets
  - If the learning process is slow, the number of draws J is limited
  - Because of the random drawings, it is not certain that each data is used as often for learning as for the validation

K-fold Cross-Validation

- Variant of Monte-Carlo Cross-Validation
  - Divide the set of *N* available data into *K* disjoined subsets having almost the same size
  - Use *K-1* data sets in learning and one for the validation; learning and validation done *K* times
  - Generalization error is an average of the results with different K learning and validation sets

Leave-one-out

- K-fold Cross-Validation carried to extremes: K=N
  - N models are built on basis of N-1 data
  - Tested with one data
  - Gives good results with only large data sets

Bootstrap

- Learning set: draw N data points randomly from the original data set
  - some samples may be left out, some may appear multiple times
  - process is called re-sampling
- Validation set: origial data
- Re-sampling is done *J* times
  - *J for example 10, 100, 1000 or 10000*

Bootstrap: optimism

- Difference between the learning error and generalization error is called optimism
  - Optimism measures the difference between the error obtained on the initial data set and the error that would be obtained on an infinite group of data
- Global estimate of the optimism is calculated as an average of J runs
- Generalization error is then calculated using the averaged optimism and learning error for each model structure q

Bootstrap 632

- Bootstrap method gives biased estimations
  - method systematically overestimates the errors
- Bias is corrected through weighting that is different for the terms representing the learning and the generalization errors

Fast bootstrap (FB)

- Assumes that the computationally expensive term estimated by the bootstrap, **the optimism**, is usually a smooth function (low-order polynomial) of the complexity parameter
- Approximating the optimism term makes it possible to considerably reduce the necessary number of simulations

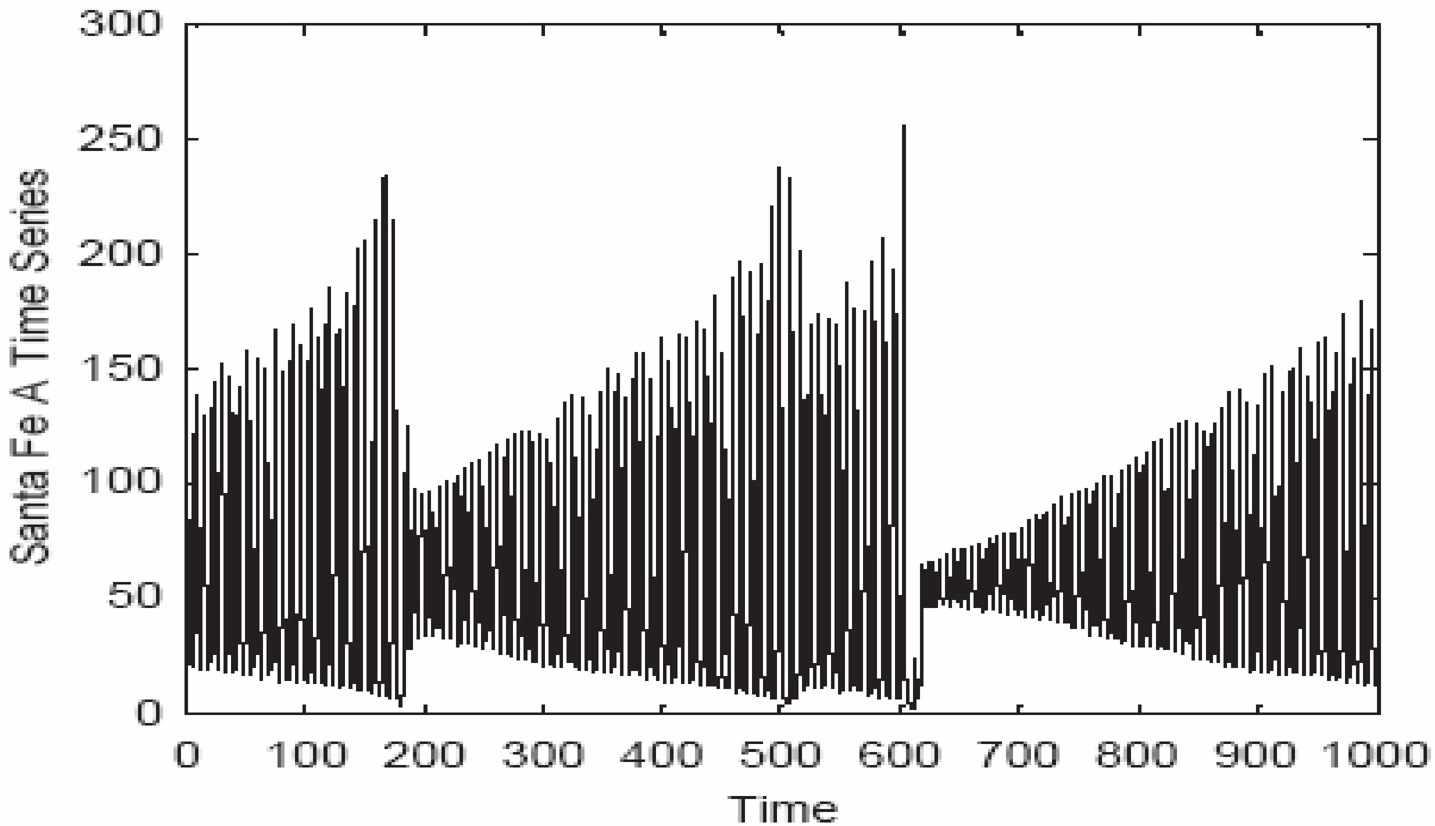
How to approximate Optimism?

- Optimism can be approximated by a lower-order function
  - for example a linear, quadratic or exponential
- Hypothesis of the approximation function of optimism
  - optimism is a polynomial of certain order
  - use statistical testing to test hypothesis
  - if hypothesis not accepted, try something else

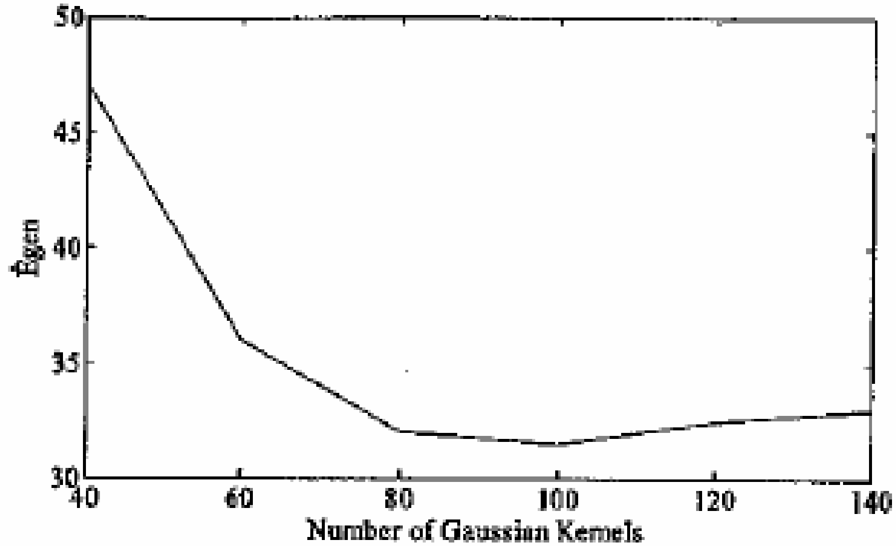
Fast bootsrap

- Computational load of resampling metods is eased using an approximation of optimism
  - approximation of generalization error with considerably less iterations
  - we are able to use bootstrap resampling with computationally heavy real life applications

Example: SantaFe



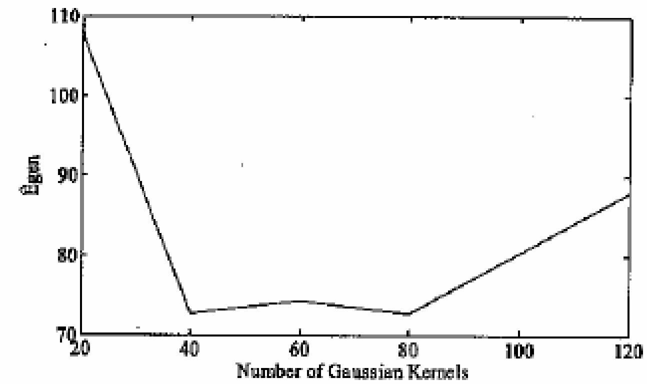
Best structure

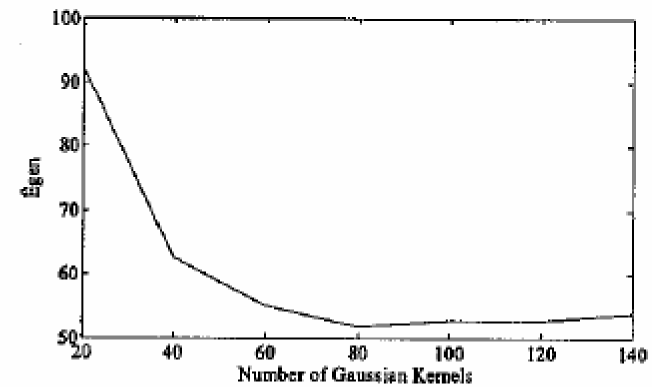


- All data used to find the best structure
  - 100 Gaussian kernels is the optimal solution
- Does the validation methods lead to similar results?
  - only first 1000 data used

Monte-Carlo Cross-Validation and Leave-One-Out

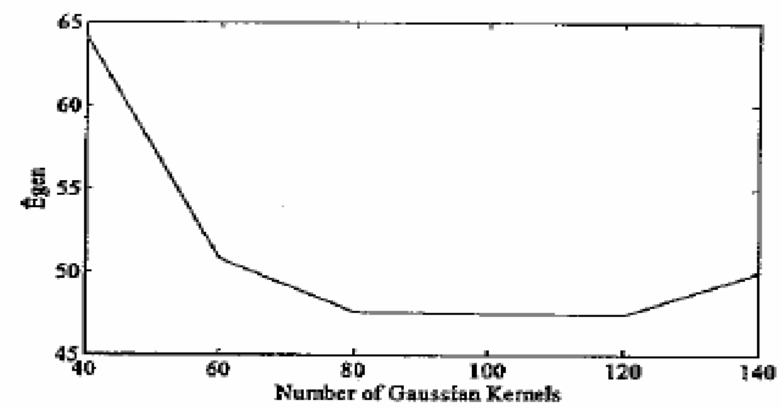
- Monte-Carlo Cross-Validation
  - ideal models 40 or 80 Gaussian Kernels, usually 40 selected
  - very different from the expected value 100
  - number of iterations 100
- Leave-One-Out
  - 80 kernels
  - number of iterations 1000!

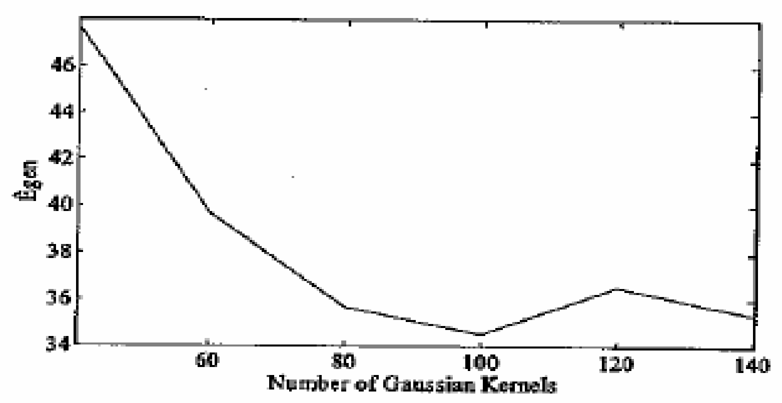




Bootstrap and Bootstrap 632

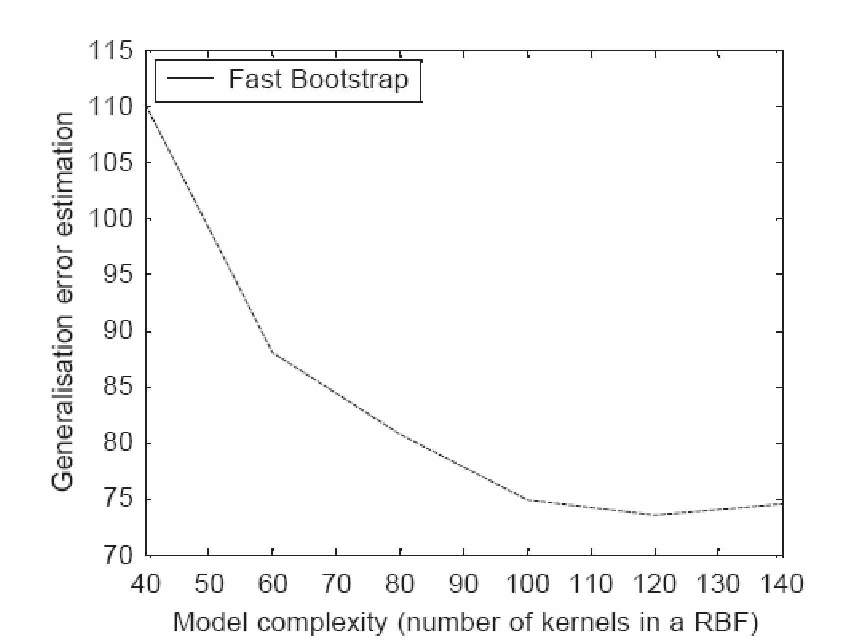
- Bootstrap
  - 120 kernels
  - 100 iterations
- Bootstrap 632
  - 100 kernels selected
  - Best results!





Fast Bootsrap

- From the second paper
  - Are the results comparable?
- 120 kernels selected
- only 20 Bootstrap replications



Comparison of the methods

- For a very large number of data, each of the methods will select the same model structure
  - In reality, the amount of data is limited
- Long calculation time
  - Leave-One-Out, K-fold Cross-Validation (less)
- Bias
  - Bootstrap, corrected in Bootstrap 632
- Variance
  - Monte-Carlo Cross-Validation, K-fold Cross-Validation, Leave-One-Out
- Bootstrap methods seems to be the best ones!