Modern Machine Learning Regression Methods

Igor Baskin

Igor Tetko

Machine Learning Regression Methods

- Multiple Linear Regression (MLR)
- Partial Least Squares (PLS)
- Support Vector Regression (SVR)
- Back-Propagation Neural Network (BPNN)
- K Nearest Neighbours (kNN)
- Decision Trees (DT)

Machine Learning Regression Methods

- **Multiple Linear Regression (MLR)**
- *Partial Least Squares (PLS)*
- *Support Vector Regression (SVR)*
- *Back-Propagation Neural Network (BPNN)*
- *K Nearest Neighbours (kNN)*
- *Decision Trees (DT)*

Multiple Linear Regression

Y=CX

$$C = (X^T X)^{-1} X^T Y$$

M < N !!!

$$C = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_M \end{pmatrix} \qquad Y = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1^1 & \cdots & x_M^1 \\ 1 & x_1^2 & \cdots & x_M^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & \cdots & x_M^N \end{pmatrix}$$

Regression coefficients

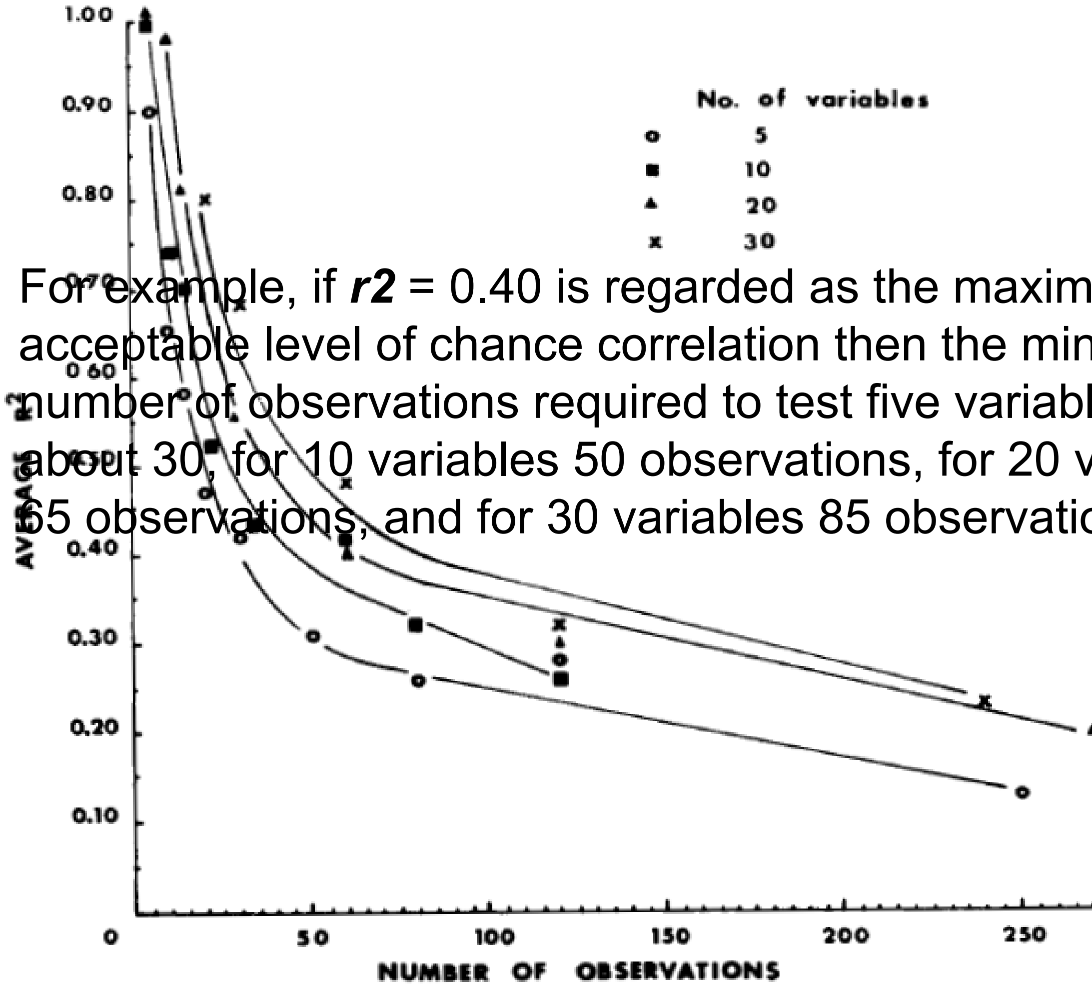
Experimental property values

Descriptor values

Topless: M < N/5 for good models

Mistery of the “Rule of 5”

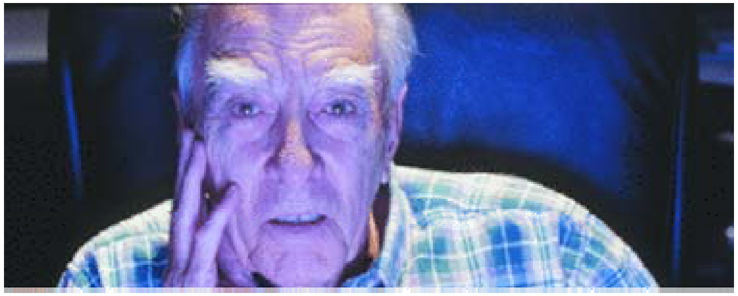
John G. Topliss



For example, if ***r2*** = 0.40 is regarded as the maximum acceptable level of chance correlation then the minimum number of observations required to test five variables is about 30, for 10 variables 50 observations, for 20 variables 65 observations, and for 30 variables 85 observations.

J.H.Topliss & R.L.Costello, JMC, 1972, Vol. 15, No. 10, P. 1066-1068.

Mistery of the "Rule of 5"


Hansch

C.Hansch, K.W.Kim, R.H.Sarma, JACS, 1973, Vol. 95, No.19, 6447-6449

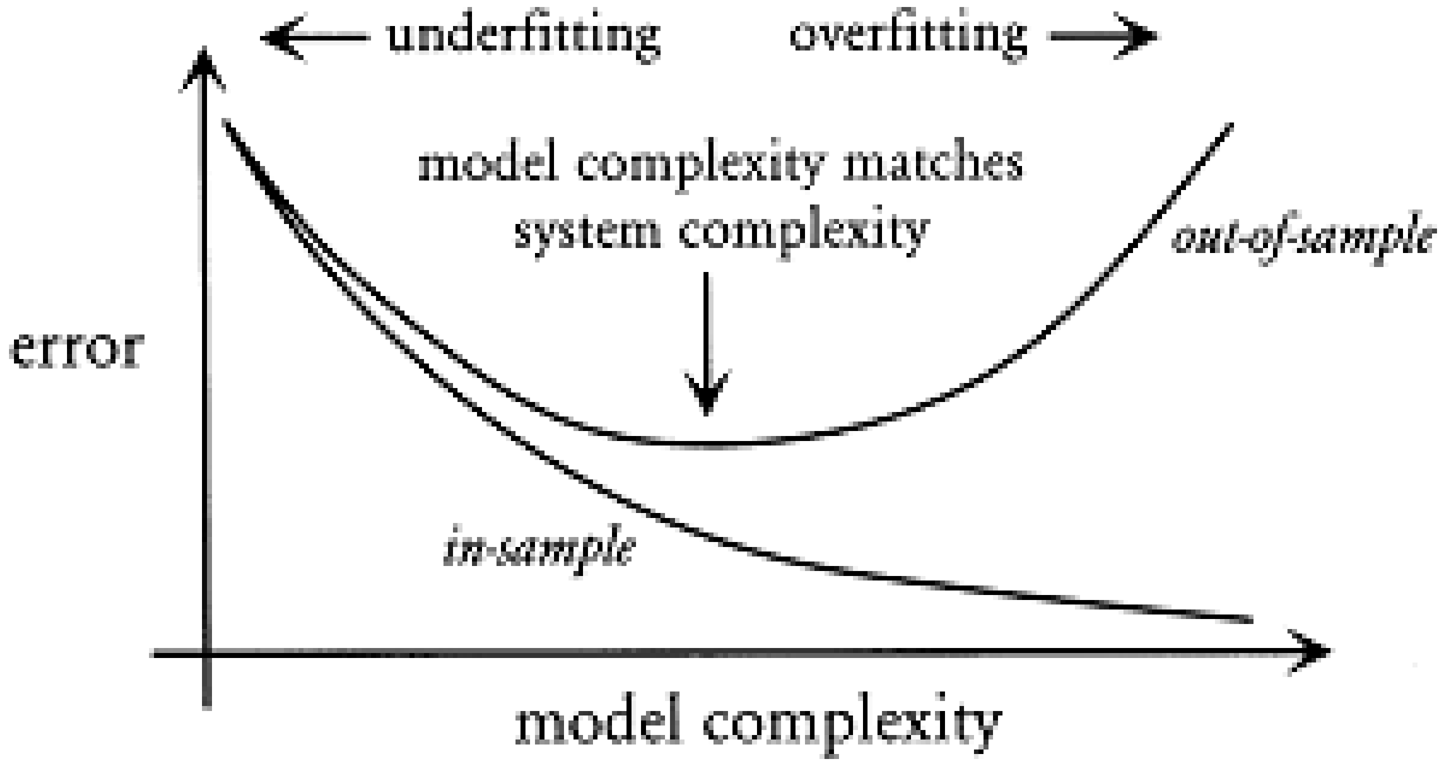
$$\log \frac{1}{K_{ER,I}} = 0.453(\pm 0.28)\pi\text{-}4 - 0.804(\pm 0.30)\sigma - 0.232(\pm 0.17)E_s\text{-}4 - 2.369(\pm 0.20)$$

14 0.953 0.168

Topliss and Costello[2] have pointed out the danger of finding meaningless chance correlations with three or four data points per variable.

The correlation coefficient is good and there are almost <u>five data points per variable</u>.

Model Overfitting for the Multiple Linear Regression



Model complexity ~ the number of descriptors

Machine Learning Regression Methods

- *Multiple Linear Regression (MLR)*
- **Partial Least Squares (PLS)**
- *Support Vector Regression (SVR)*
- *Back-Propagation Neural Network (BPNN)*
- *K Nearest Neighbours (kNN)*
- *Decision Trees (DT)*

Partial Least Squares (PLS)
Projection to Latent Structures

$$y^j = \sum_{k=1}^{K} a_k s_k^j \qquad s_k^j = \sum_{k=l}^{M} l_{ik} \underline{x}_i^j$$

$$y \propto c_0 + c_1 x_1 + \ldots + c_M x_M$$

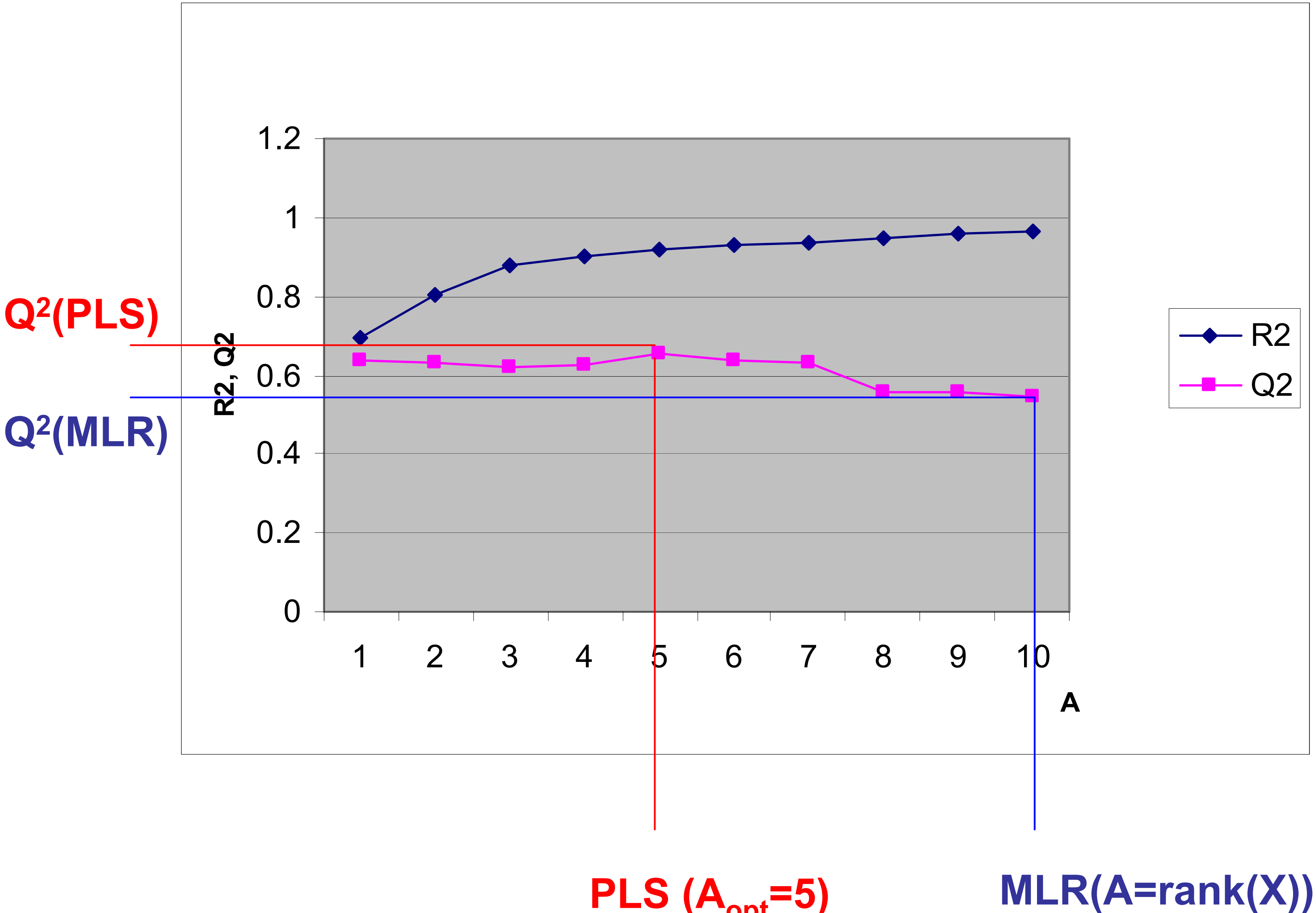
Principal Component Analysis (PCA)

$$
\begin{cases}
\vec{l}_i = \arg\max\{\operatorname{var}(\vec{l}_i^{\,T} X)\} \\
(\vec{l}_i, \vec{l}_k) = 0,\ i \neq k \\
(\vec{l}_i, \vec{l}_i) = 1
\end{cases}
$$

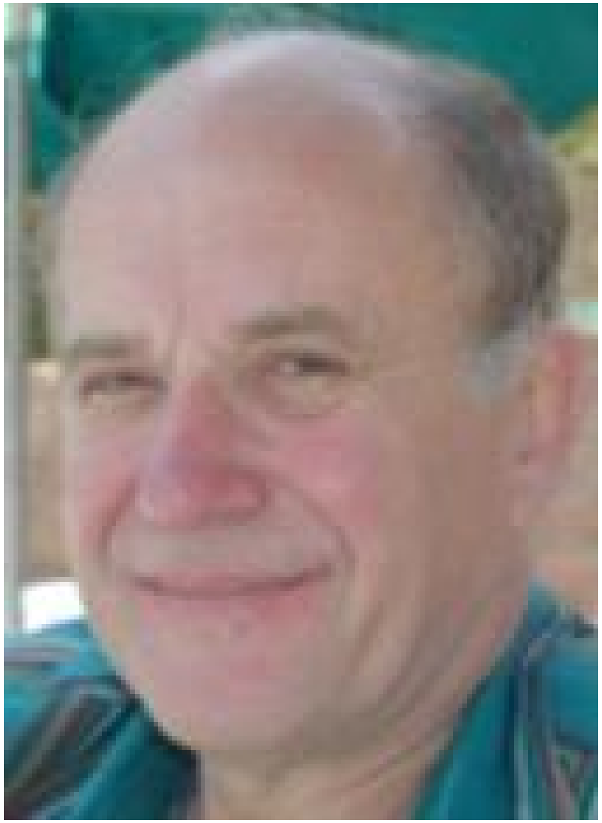
Partial Least Squares (PLS)

$$
\begin{cases}
\vec{l}_i = \arg\max\{\operatorname{cov}(\vec{y}, \vec{l}_i^{\,T} X)\} \\
(\vec{l}_i, \vec{l}_k) = 0,\ i \neq k \\
(\vec{l}_i, \vec{l}_i) = 1
\end{cases}
$$

Dependence of R^2, Q^2 upon the Number of Selected Latent Variables A

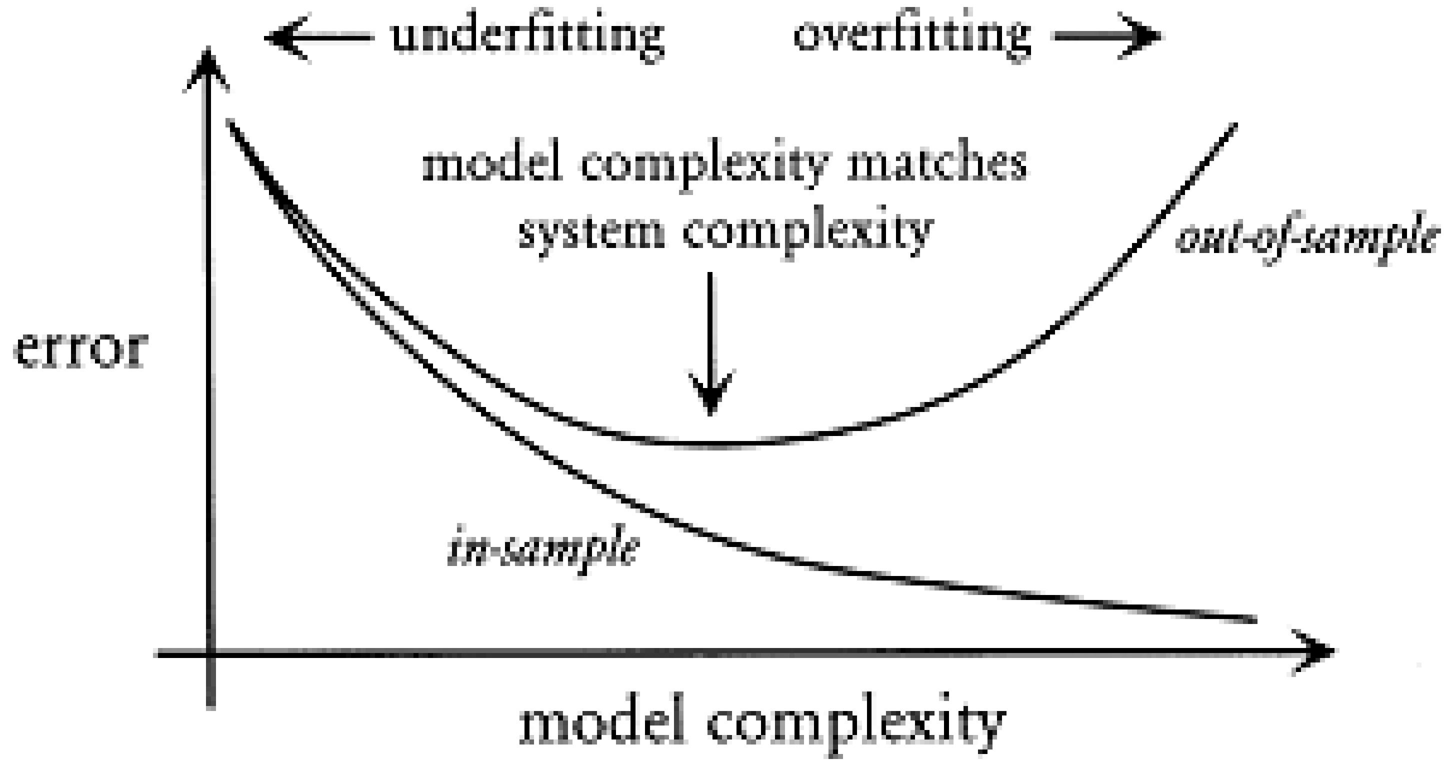


Herman Wold (1908-1992)



Swante Wold

Model Overfitting for the Partial Least Squares

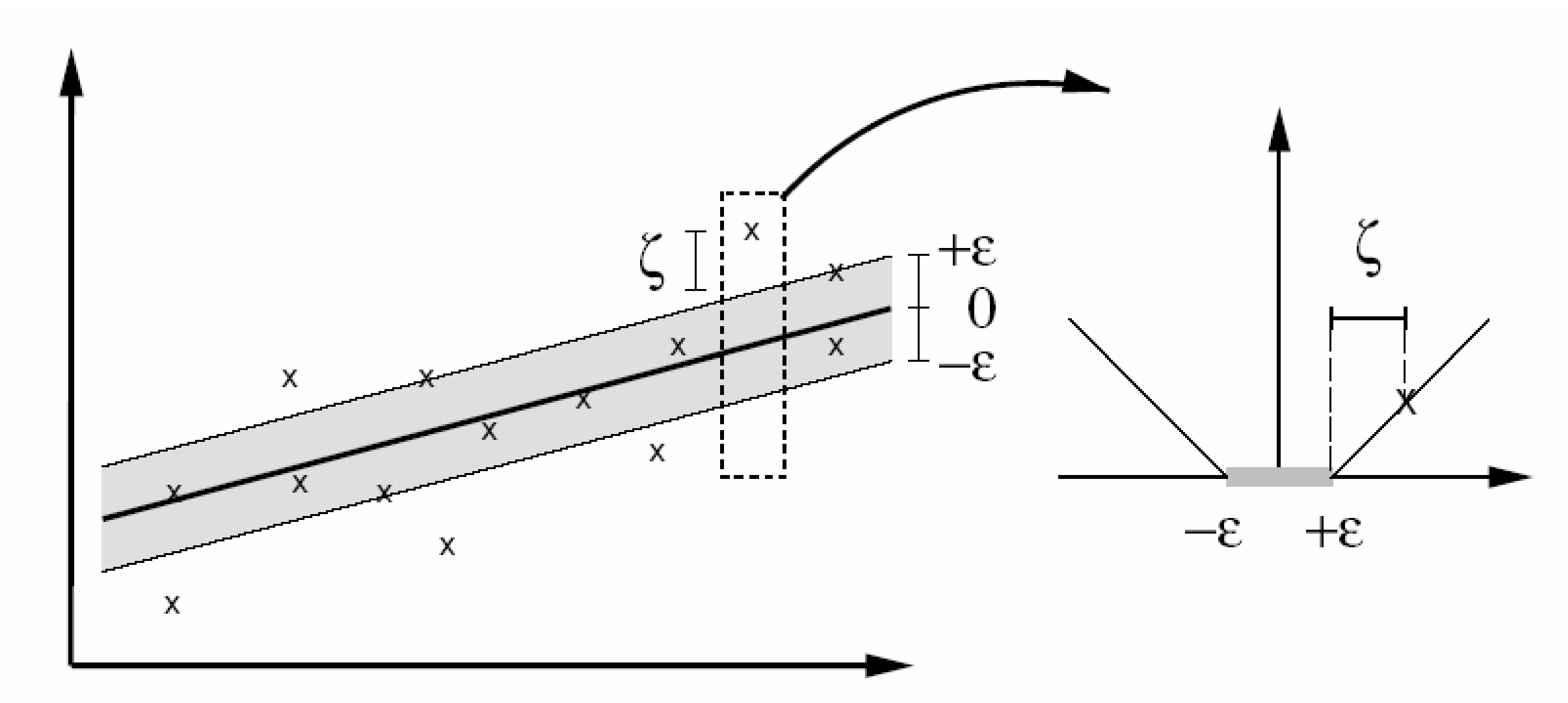


Model complexity ~ the number of latent variables

Machine Learning Regression Methods

- *Multiple Linear Regression (MLR)*
- *Partial Least Squares (PLS)*
- **Support Vector Regression (SVR)**
- *Back-Propagation Neural Network (BPNN)*
- *K Nearest Neighbours (kNN)*
- *Decision Trees (DT)*

Support Vector Regression. ε-Insensitive Loss Function



Only the points outside the ε-tube are penalized in a linear fashion

$$|\xi|_{\varepsilon} := \begin{cases} 0 & \text{if } |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases}$$

Linear Support Vector Regression. Primal Formulation

Complexity term

Penalty term

Task for QP

$$\arg\min_{\omega,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

subject to

$$\begin{cases} y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

Points should lie below the upper bound of ε-tube

Points should lie above the lower bound of ε-tube

Regression function

$$f(x) = \langle w, x \rangle + b$$

C – trade-off between the flatness (and complexity) of f and the amount up to which deviations larger than ε are tolerated

Linear Support Vector Regression. Dual Formulation

Task for QP

$$\arg\min_{\alpha,\alpha^*}\begin{cases}-\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)<x_i,x_j> \\ -\varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*)\end{cases}$$

subject to

$$\begin{cases}\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0 \\ \alpha_i,\alpha_i^*\in[0,C]\end{cases}$$

these parameters should be optimized

Regression function

$$f(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)<x_i,x>+b$$

In reality, only several objects, for which $\alpha_i-\alpha_i^*>0$ take part in this summation. Such points are called **support vectors**.

Dualism of QSAR/QSPR Models

Ordinary method

Primal formulation

$$f(x) = \langle w, x \rangle + b$$

Dual formulation

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$$

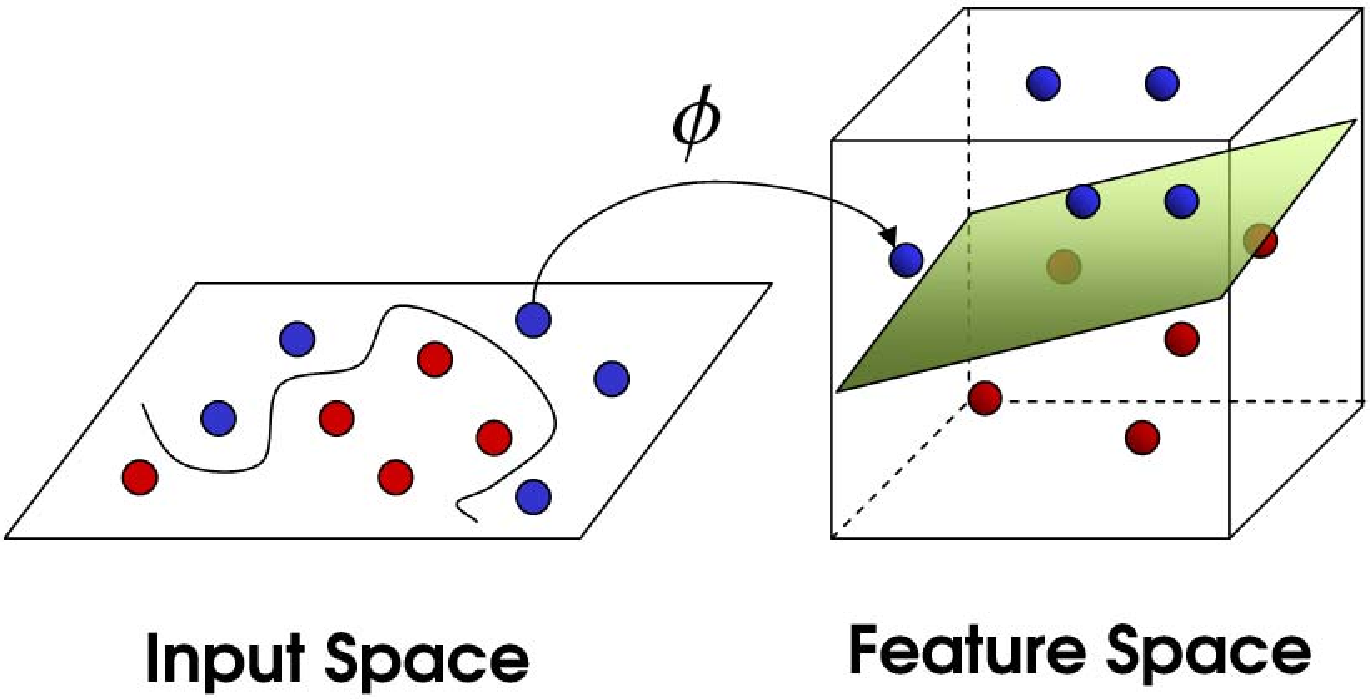
Similarity-based method

Dualism of QSAR/QSPR Approaches

Vector-Based Methods	Similarity-Based Methods
Multiple linear regression, partial least squares, backpropagation neural networks, regression trees, etc.	K nearest neighbours, RBF neural networks
Support vector regression in primal formulation	Support vector regression in dual formulation

The Lagrange's methods builds a bridge between both types of approaches

Kernel Trick



Any non-linear problem (classification, regression) in the original input space can be converted into linear by making non-linear mapping Φ into a feature space with higher dimension

Kernel Trick

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) < x_i, x > + b$$

In high-dimensional feature space →

$$< \Phi(x), \Phi(x') > = K(x, x')$$

← **In low-dimensional Input space**

Kernel

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

In order to convert a linear statistical method to a powerful non-linear kernel-based counterpart it is sufficient to substitute all dot products in the dual formulation of the linear method for a kernel

Common Kernel Functions

Gaussian RBF $$K(x,x') = \exp(\frac{-\|x-x'\|^2}{2\sigma^2})$$

Polynomial $$K(x,x') = (< x, x' > +\theta)^d$$

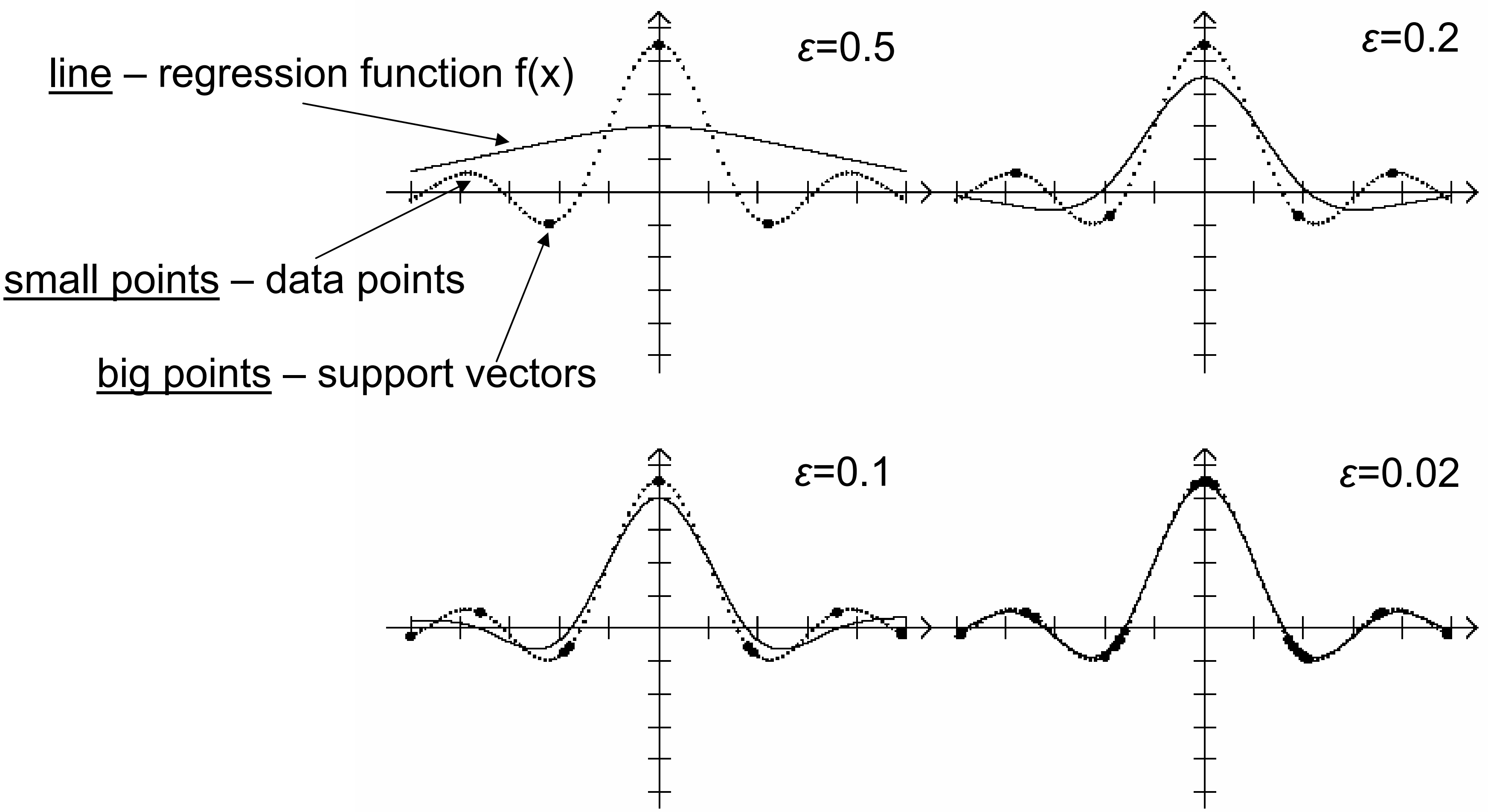
Sigmoidal $$K(x,x') = \tanh(\kappa < x, x' > +\theta)$$

Inverse multi-quadratic $$K(x,x') = \frac{1}{\sqrt{(x-x')^2 + c^2}}$$

So, all these kernel functions are functions of dot products or distance between points.

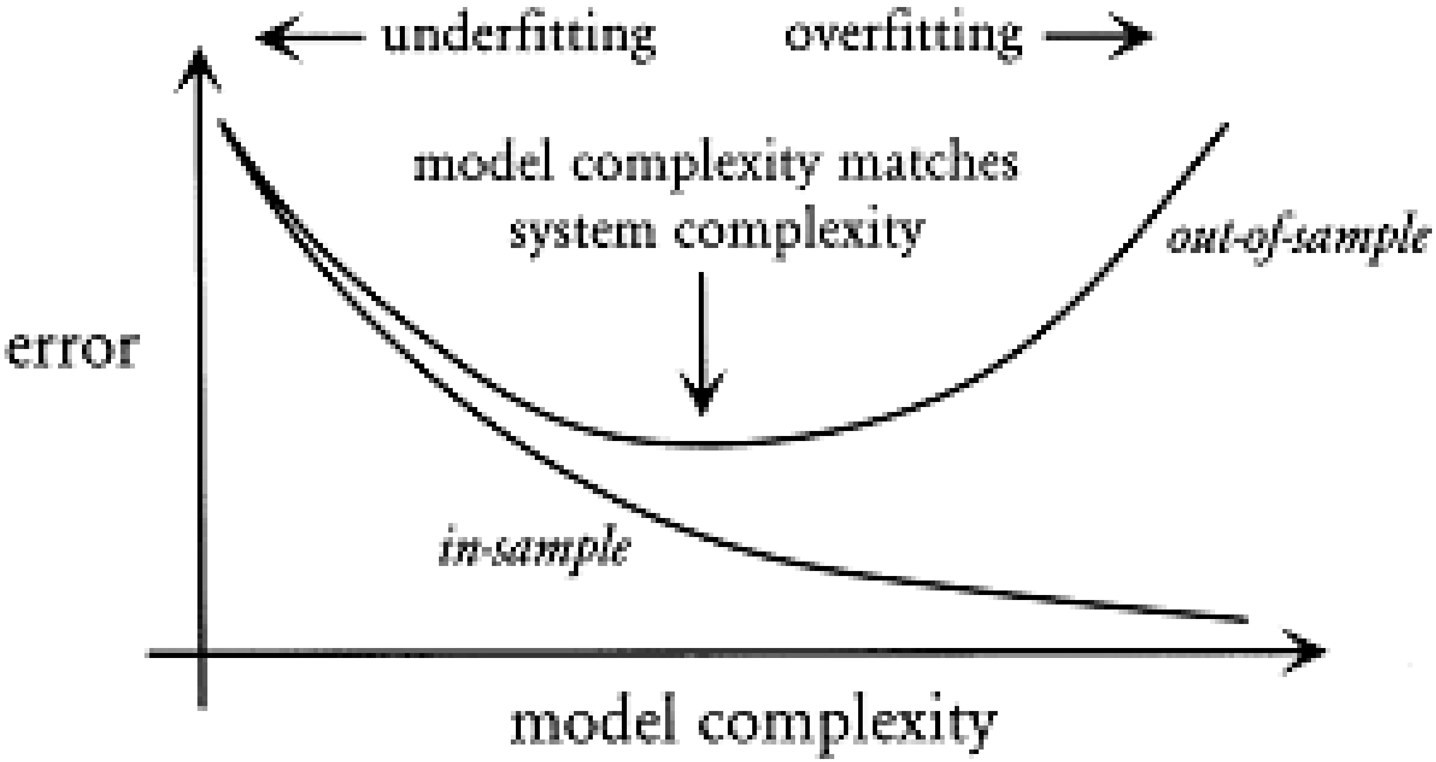
Therefore, kernels can be viewed as **nonlinear similarity measures** between objects

Function Approximation with SVR with Different Values of ε



So, the number of support vectors increases with the decrease of ε

Model Overfitting for the Support Vector Regression

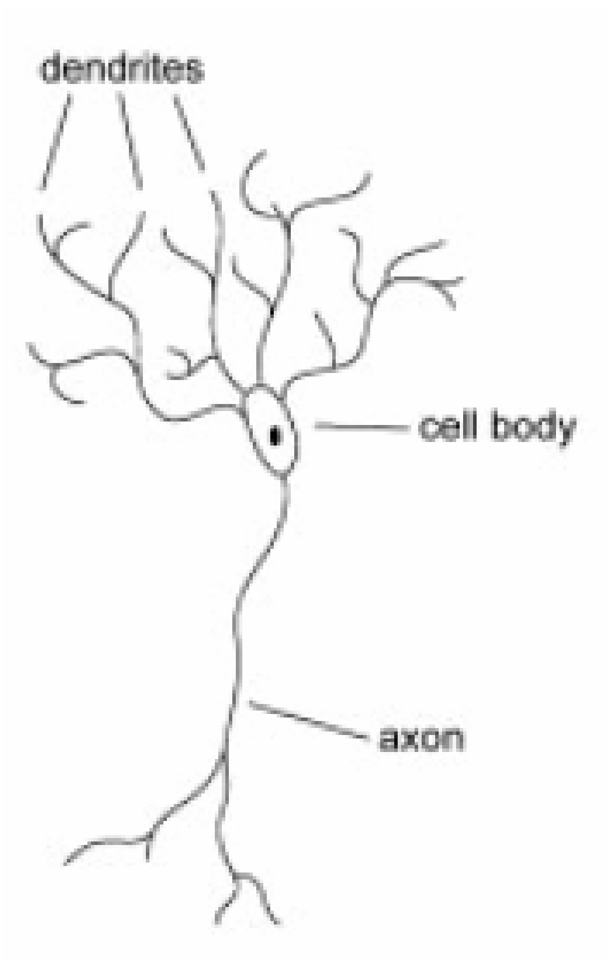


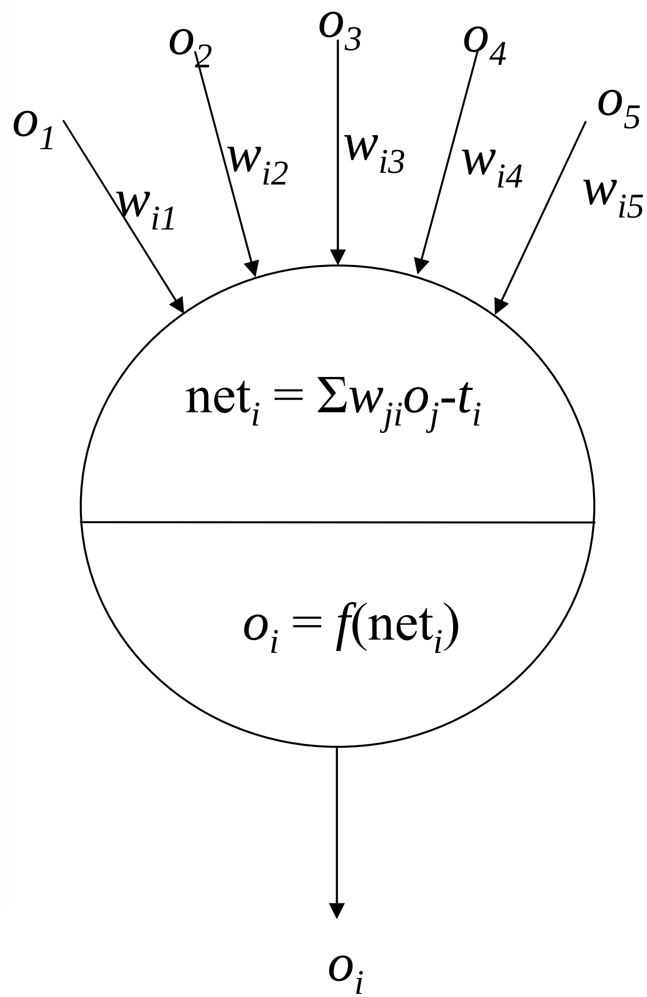
Model complexity ~ 1/ε ~ the number of support vectors

Machine Learning Regression Methods

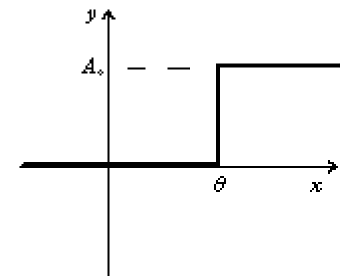
- *Multiple Linear Regression (MLR)*
- *Partial Least Squares (PLS)*
- *Support Vector Regression (SVR)*
- **Back-Propagation Neural Network (BPNN)**
- *K Nearest Neighbours (kNN)*
- *Decision Trees (DT)*

Artificial Neuron

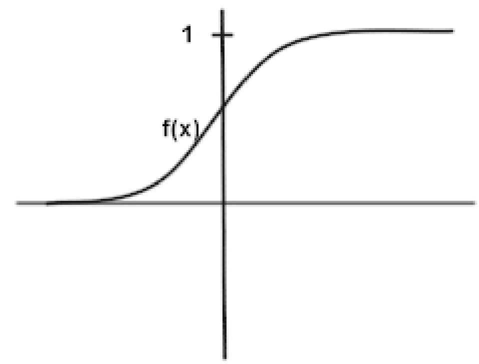




Transfer function

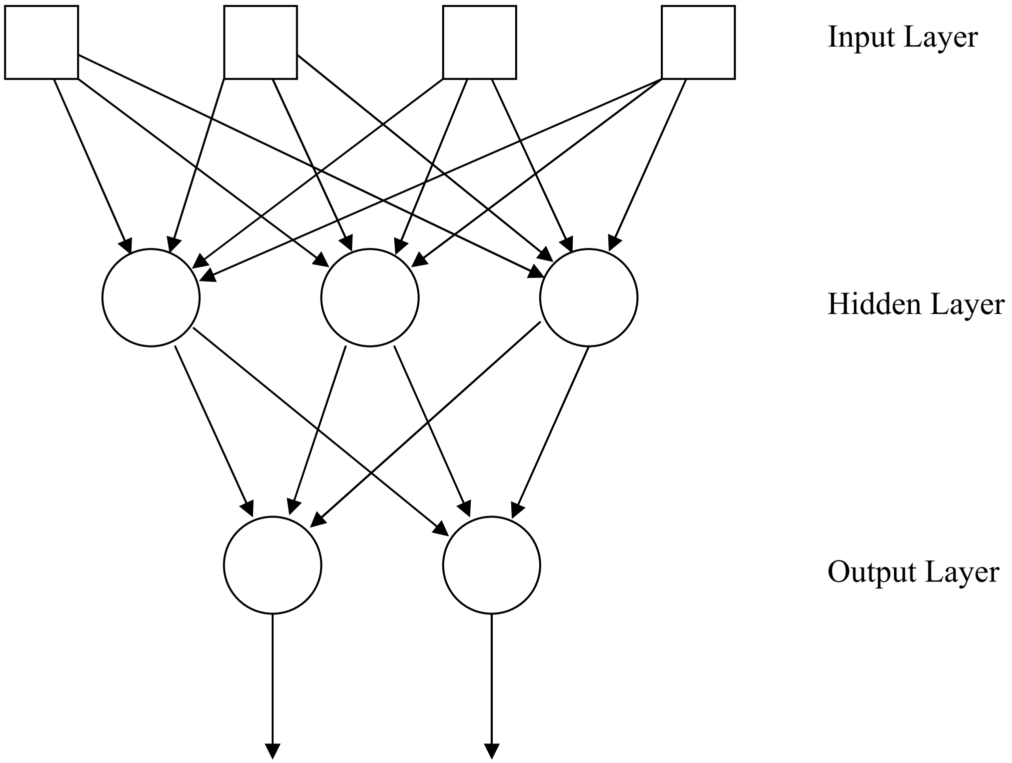


$$f(x) = \begin{cases} 1, & x \geq \theta \\ 0, & x < \theta \end{cases}$$



$$f(x) = 1/(1 + e^{-x})$$

Multilayer Neural Network



Neurons in the input layer correspond to *descriptors*, neurons in the output layer – to *properties* being predicted, neurons in the hidden layer – to *nonlinear latent variables*

Generalized Delta-Rule

This is application of the steepest descent method to training backpropagation neural networks

$$\Delta w_{ij} = -\eta \frac{\partial R_{emp}}{\partial w_{ij}} = -\eta y_i \delta_j \frac{df_j(e)}{de} = -\eta y_i \delta_j'$$

η – learning rate constant

1974

Paul Werbos

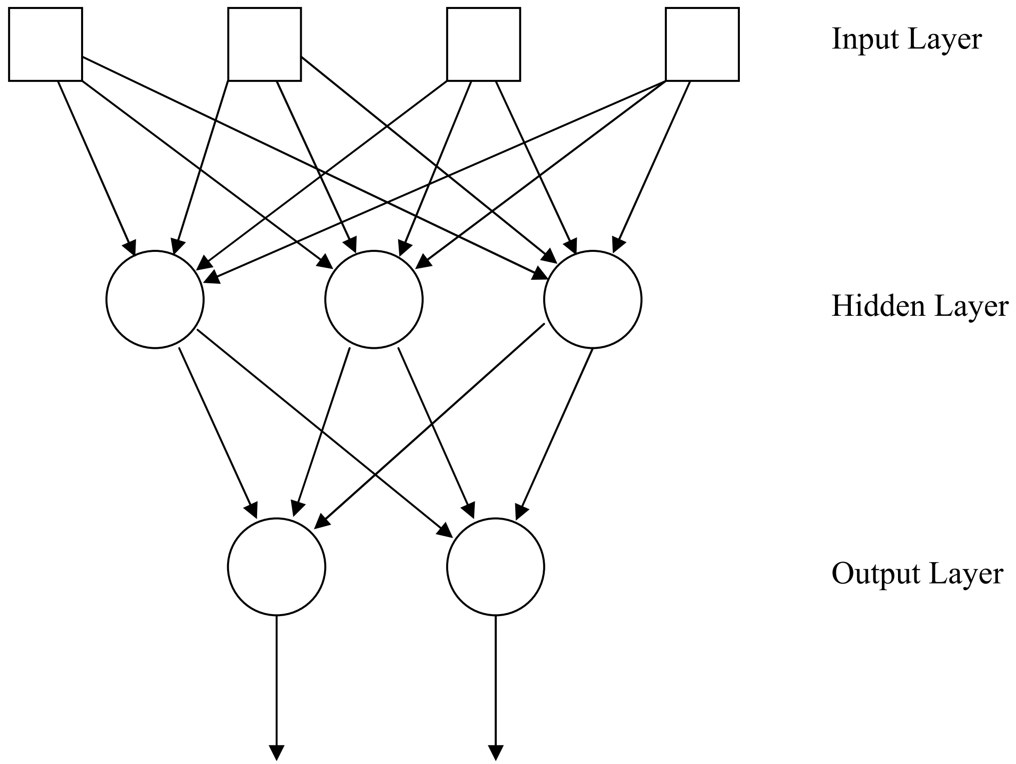
1986

David Rummelhard

James McClelland

Geoffrey Hinton

Multilayer Neural Network



The number of weights corresponds to the number of adjustable parameters of the method

Origin of "Rule of 2"

The number of weights (adjustable parameters) for the case of one hidden layer

W = (I+1)H + (H+1)O

Parameter ρ: $$\rho = \frac{N}{W}$$

$$1.8 \le \rho \le 2.2$$

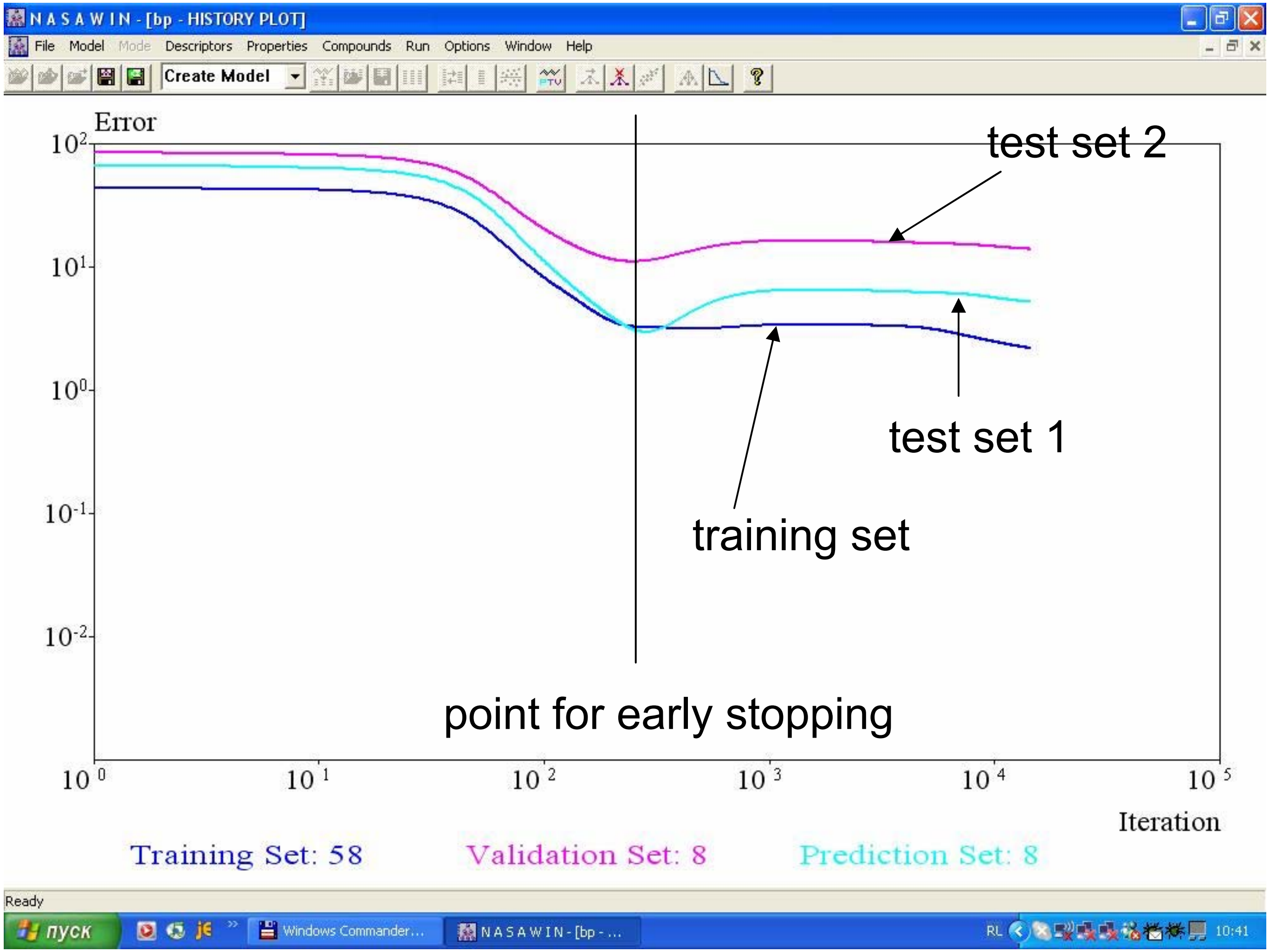
T.A. Andrea and H. Kalayeh, *J. Med. Chem*., **1991**, *34*, 2824-2836.

End of "Rule of 2"

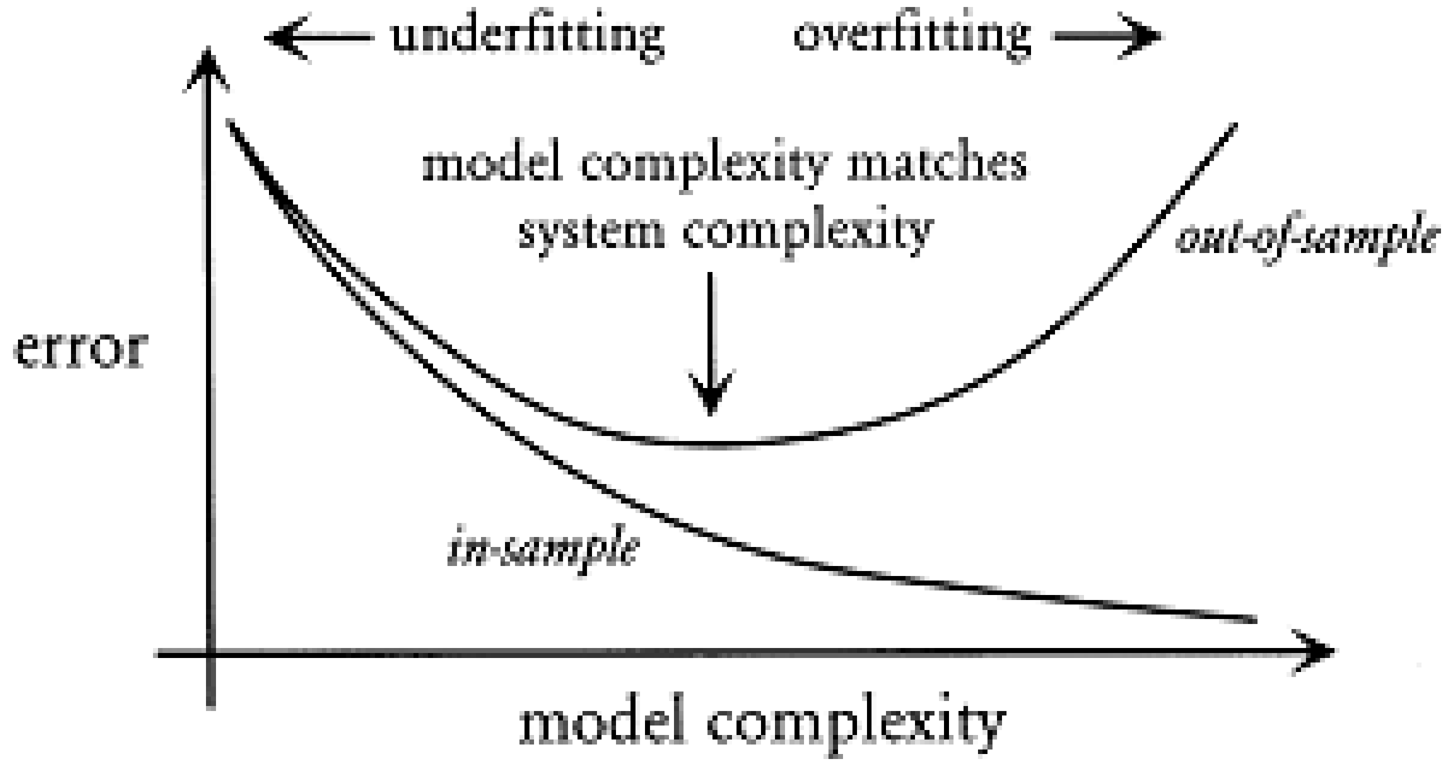
I.V. Tetko, D.J. Livingstone, Luik, A.I. *J. Chem. Inf. Comput. Sci*., **1995**, *35*, 826-833.

Baskin, I.I. et al. *Foundations Comput. Decision. Sci.* **1997**, v.22, No.2, p.107-116.

Overtraining and Early Stopping



Model Overfitting for the Backpropagation Neural Network

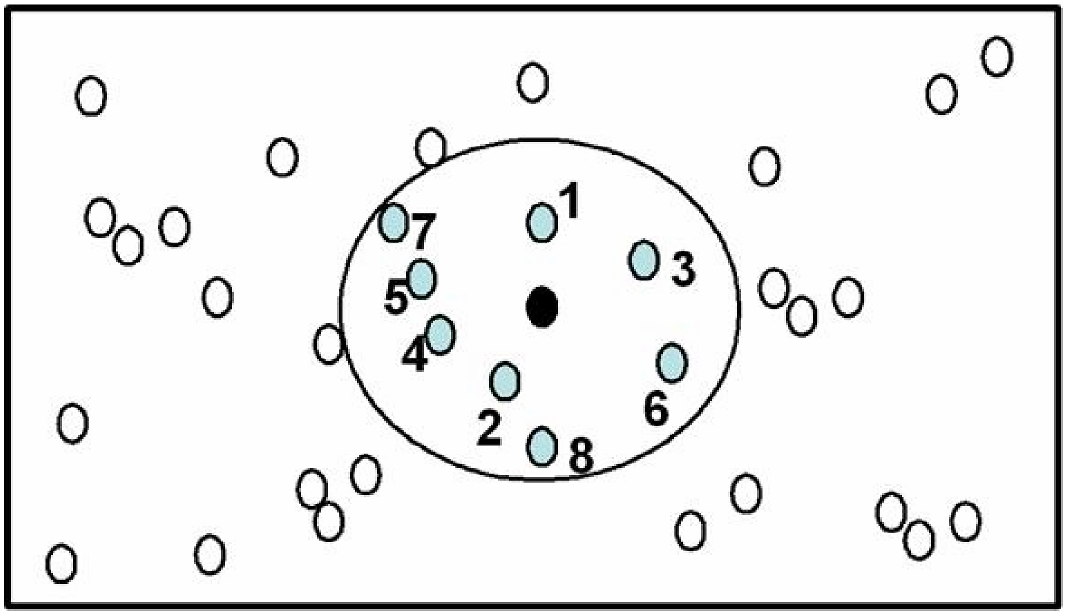


Model complexity ~ number of epochs & weights

Machine Learning Regression Methods

- *Multiple Linear Regression (MLR)*
- *Partial Least Squares (PLS)*
- *Support Vector Regression (SVR)*
- *Back-Propagation Neural Networks (BPNN)*
- **K Nearest Neighbours (kNN)**
- *Decision Trees (DT)*

K Nearest Neighbours



$$D_{ij}^{Euclid} = \sqrt{\sum_{k=1}^{M} (x_k^i - x_k^j)^2}$$

$$D_{ij}^{Manhat\tan} = \sum_{k=1}^{M} \left| x_k^i - x_k^j \right|$$

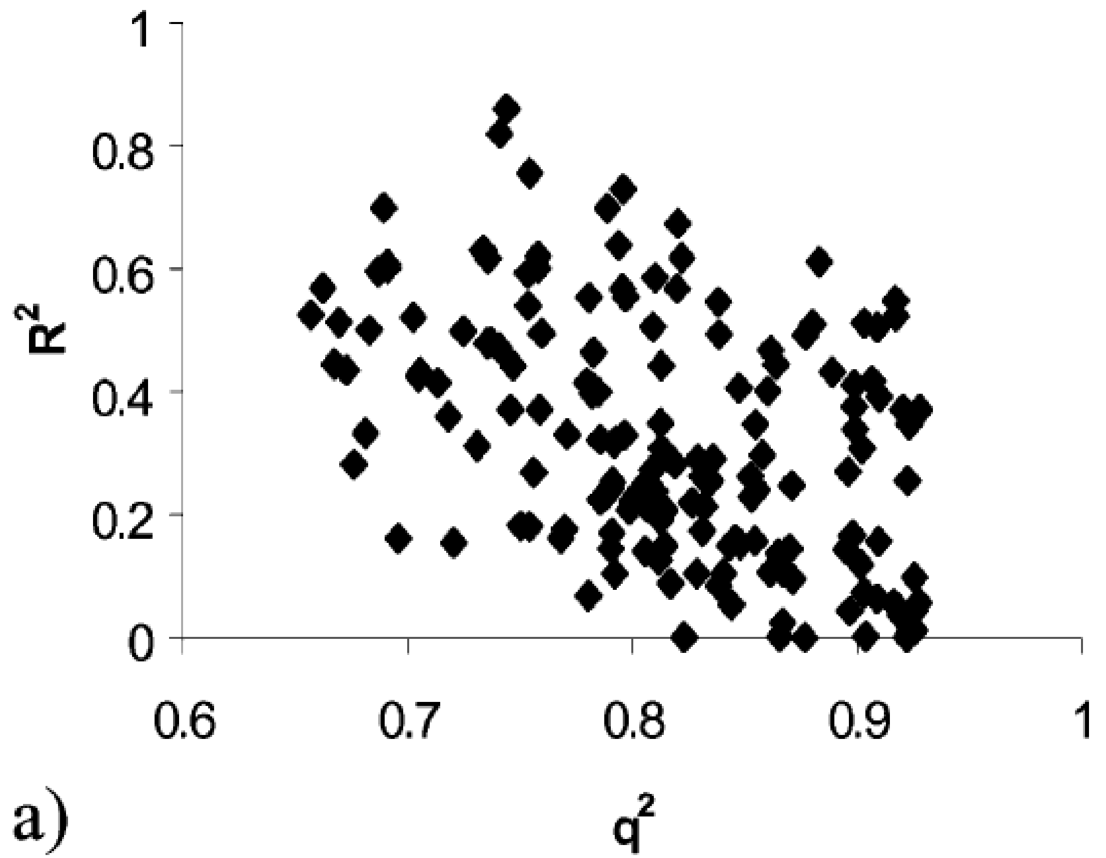
Non-weighted

$$y_i^{pred} = \frac{1}{\mathrm{k}} \sum_{j \in k-neighbours} y_j$$

Weighted

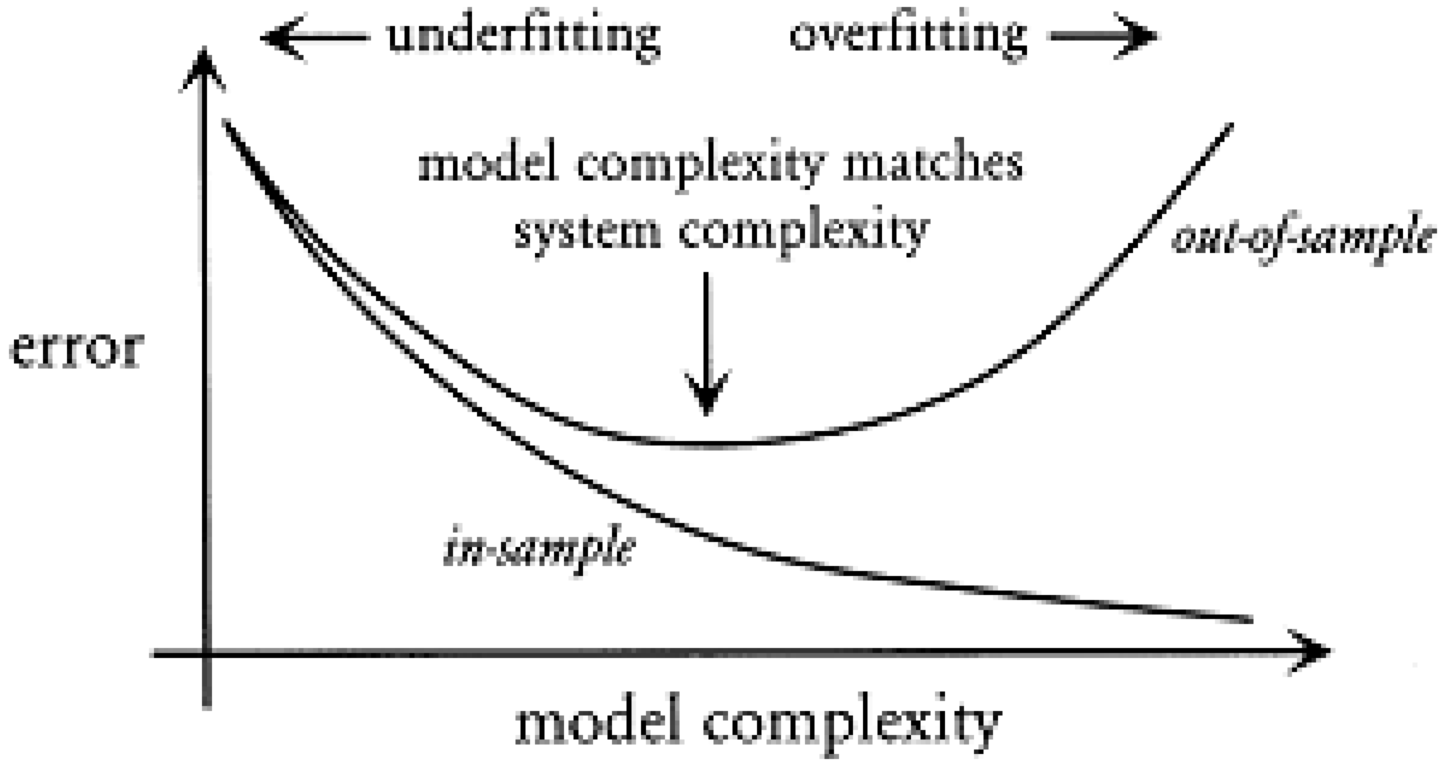
$$y_i^{pred} = \frac{1}{\sum_{j \in k-neighbours} \frac{1}{D_{ij}}} \cdot \frac{1}{D_{ij}} \sum_{j \in k-neighbours} y_j$$

Overfitting by Variable Selection in *k*NN



Golbraikh A., Tropsha A. Beware of q^2! *JMGM*, **2002**, *20*, 269-276

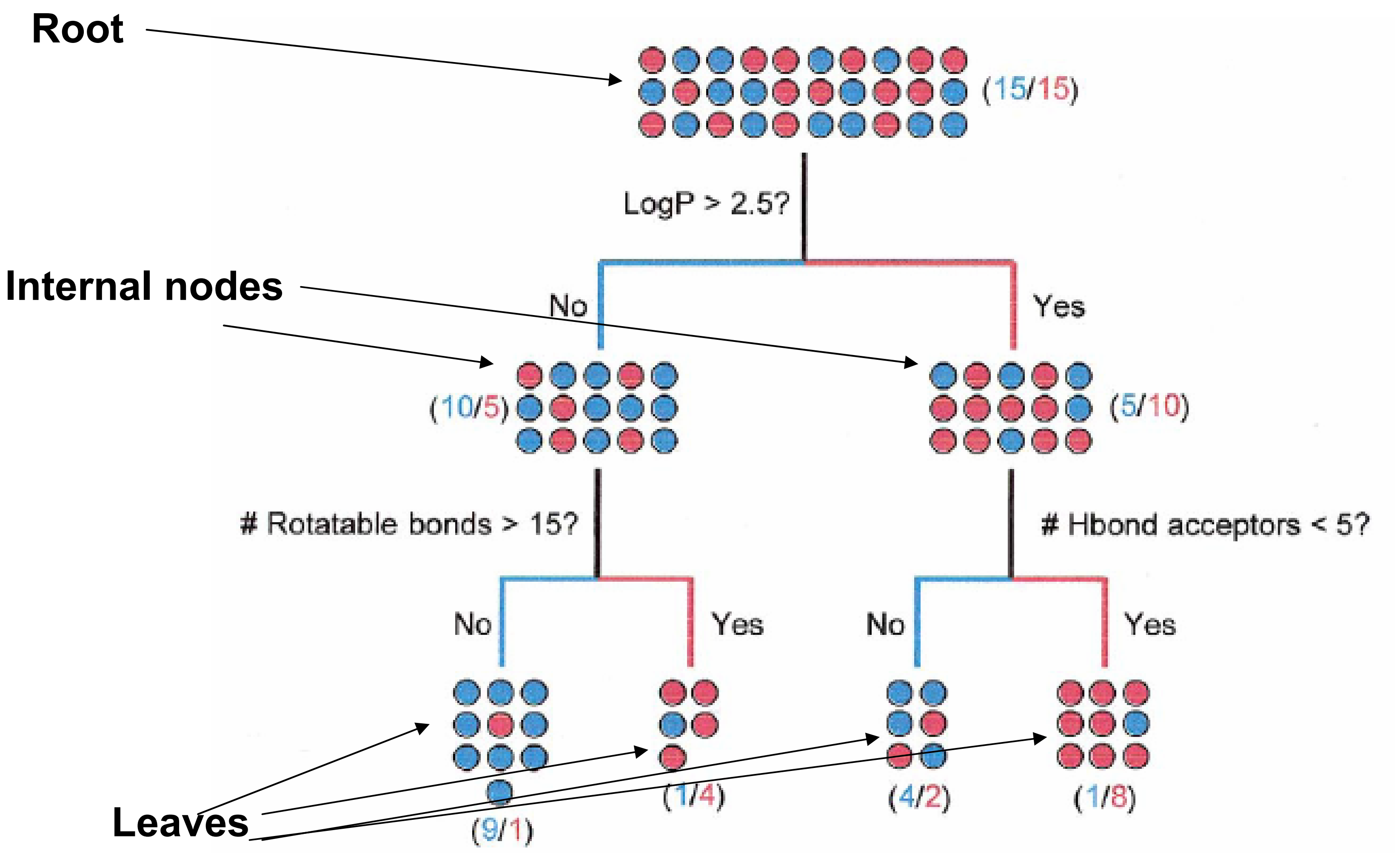
Model Overfitting for the *k* Nearest Neighbours



Model complexity ~ selection of descriptors ~1/k

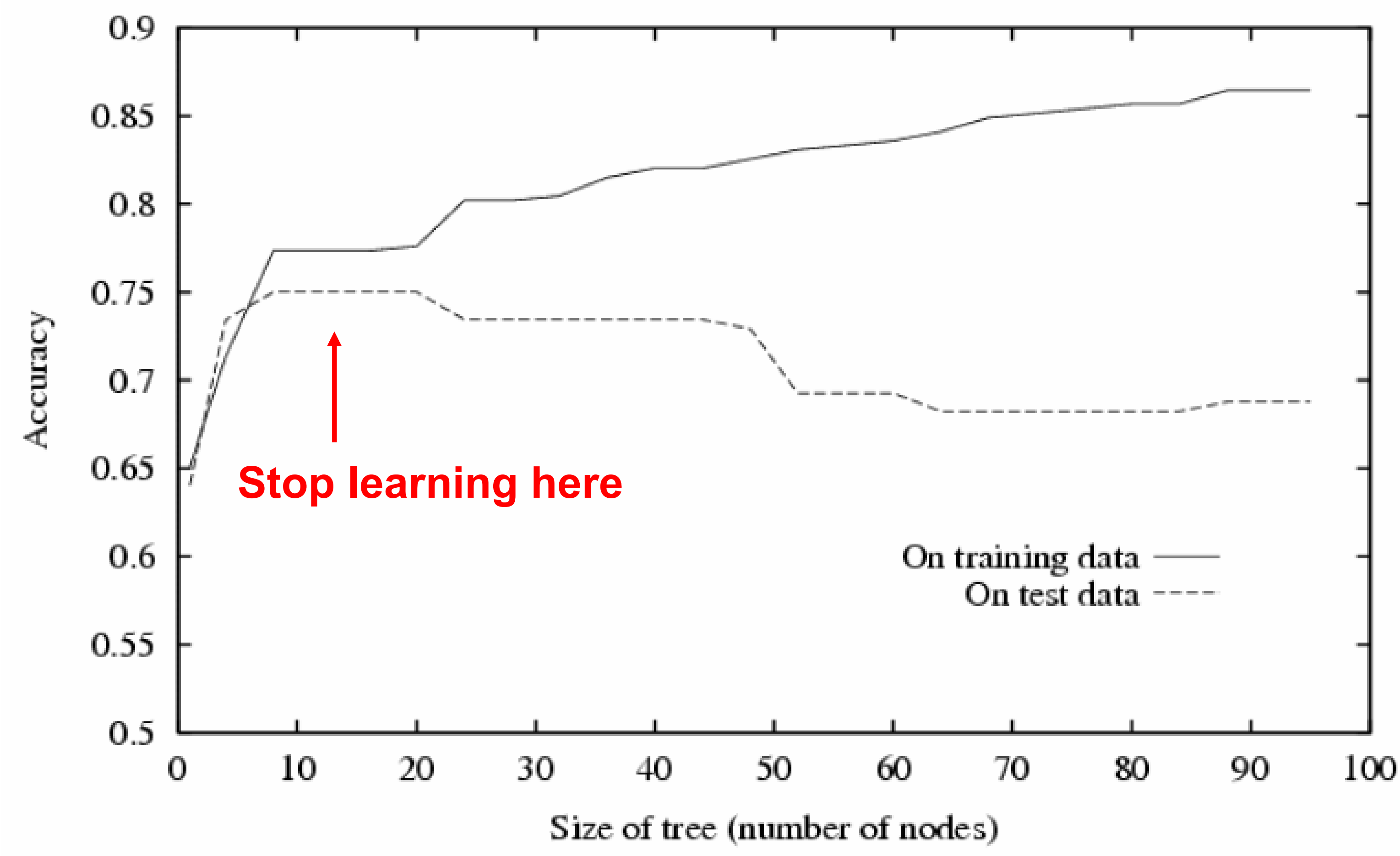
Machine Learning Regression Methods

- *Multiple Linear Regression (MLR)*
- *Partial Least Squares (PLS)*
- *Support Vector Regression (SVR)*
- *Back-Propagation Neural Networks (BPNN)*
- *K Nearest Neighbours (kNN)*
- **Decision Trees (DT)**



A decision tree splits a set of objects into subsets (usually 2 subsets in so-called binary trees) that are purer in composition. After that splitting is applied recursively to these subsets. Splitting starts from a root, and the tree growth continues while some statistical criterion allows it.

Overfitting and Early Stopping of Tree Growth

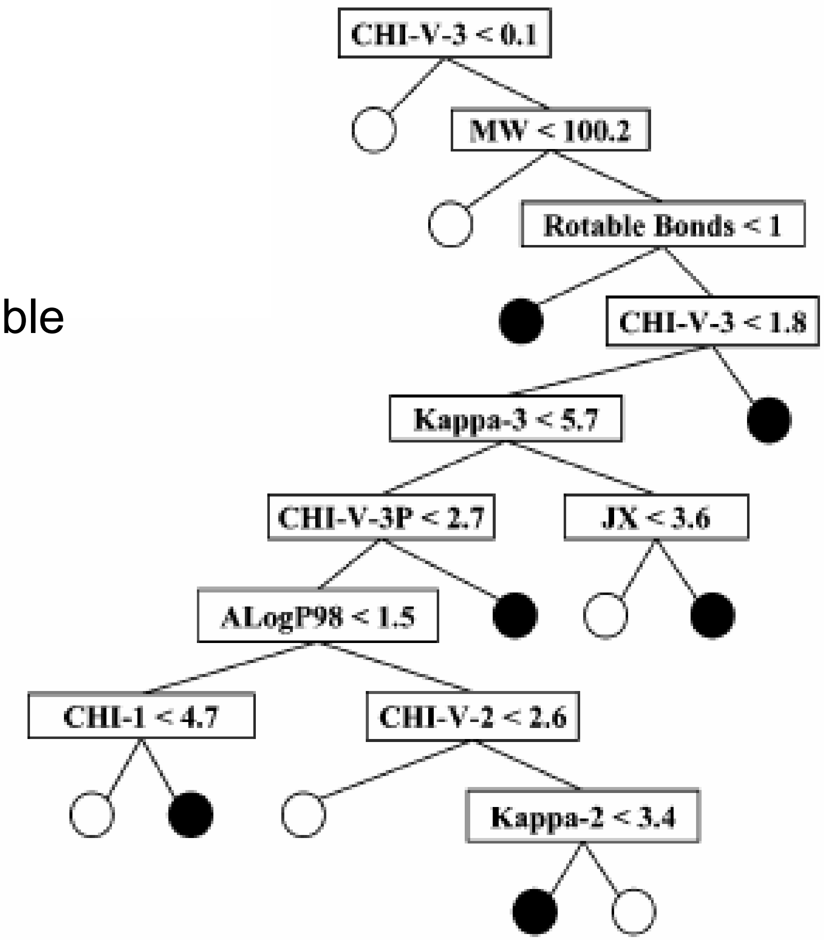


Decision trees can overfit data. So, it is necessary to use an external test set in order to stop tree growth at the optimal tree size

Decision Tree for Biodegradability

○ - biodedagrable

● - nonbiodedagrable

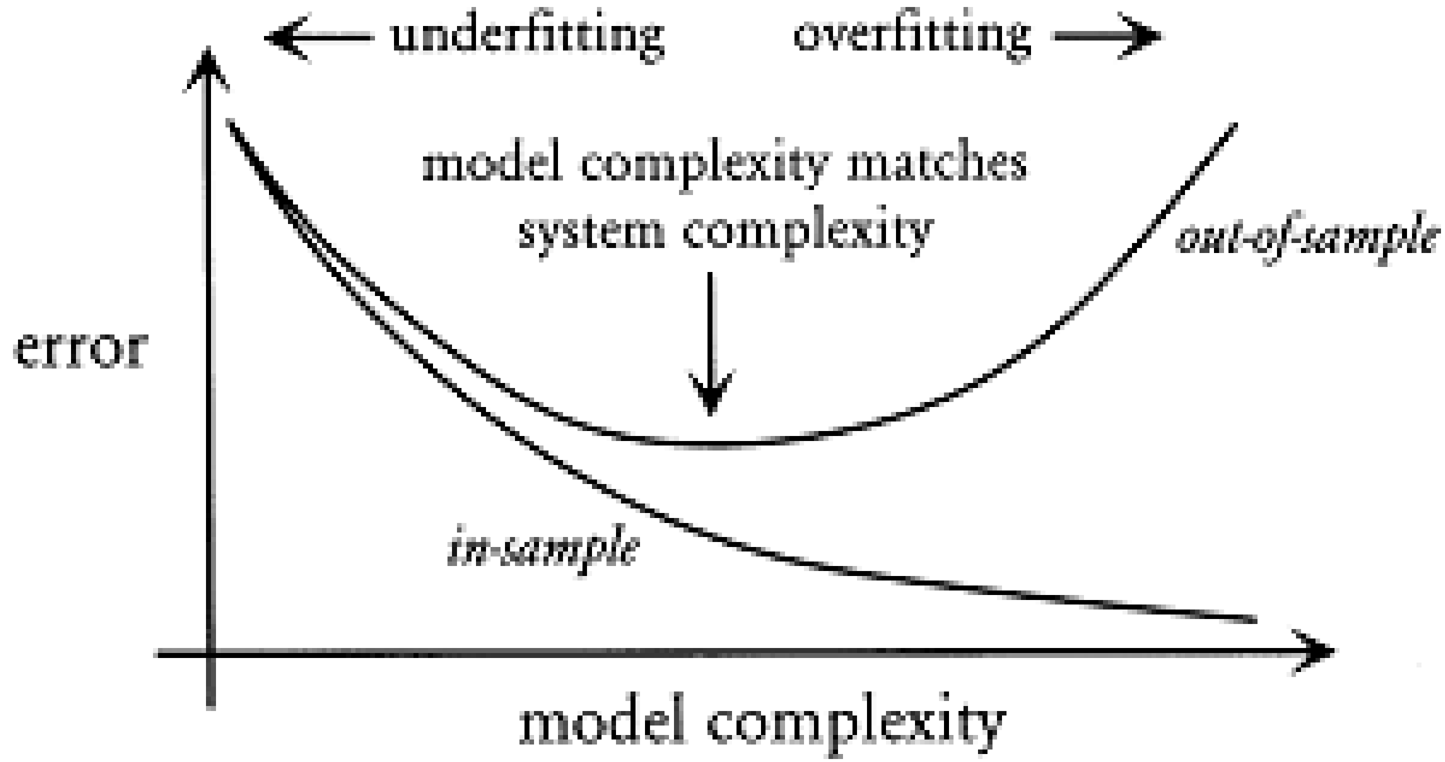


JCICS, **2004**, *44*, 862-870

Tasks for Decision Trees

- Classification – in accordance with the class of objects dominating at the leaves of decision trees (classification trees)
- Regression – in accordance with the average values of properties or MLR model built at the leaves of decision trees (regression trees)

Model Overfitting for the Decision Trees Regression

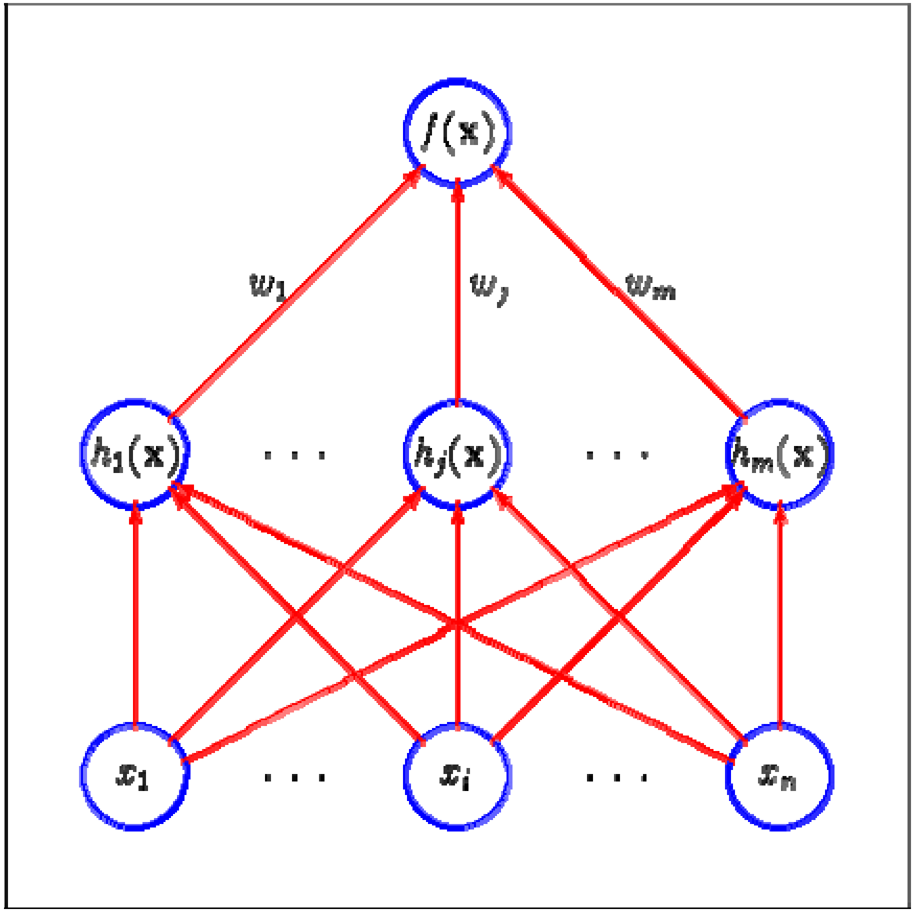


Model complexity ~ the number of nodes

Conclusions

- There are many machine learning methods
- Different problems may require different methods
- All methods could be prone of overfitting
- But all of them have facilities to tackle this problem

Exam. Question 1



What is it?

1. Support Vector Regression
2. Backpropagation Neural Network
3. Partial Least Squares Regression

Exam. Question 2

Which method is not prone to overfitting?

1. Multiple Linear Regression
2. Partial Least Squares
3. Support Vector Regression
4. Backpropagation Neural Networks
5. K Nearest Neighbours
6. Decision Trees
7. Neither