The Role of Kernel Function in Regularization Network

Petra Kudová

Department of Theoretical Computer Science
Institute of Computer Science
Academy of Sciences of the Czech Republic

ITAT 2006

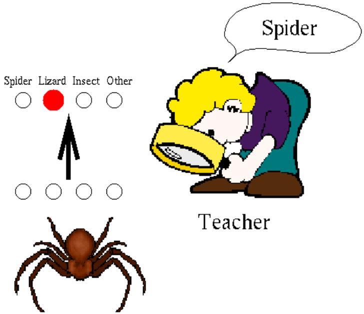
Outline

- Introduction
  - supervised learning
- Regularization Networks
  - regularization theory
  - RN learning algorithm
  - role of kernel function
- Experimental Results
  - the role of kernel function and regularization parameter
  - comparison of different kernel functions
- Summary and Future Work

Supervised Learning

Learning

- given set of data samples
- find underlying trend, description of data



Supervised Learning

- data – input-output patterns
- create model representing IO mapping
- classification, regression, prediction, etc.

Regularization Networks

Regularization Networks

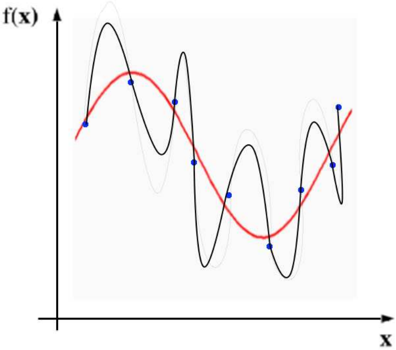
- method for supervised learning
- a family of feed-forward neural networks with one hidden layer
- derived from regularization theory
- very good theoretical background

Our Focus

- we are interested in their real applicability
- setup of explicit parameters – choice of kernel function

Learning from Examples – Problem Statement

- **Given:** set of data samples $\{(\vec{x}_i, y_i) \in R^d \times R\}_{i=1}^N$
- **Our goal:** recover the unknown function or find the best estimate of it



Regularization Theory

Empirical Risk Minimization:

- find f that minimizes $H[f] = \sum_{i=1}^{N}(f(\vec{x}_i) - y_i)^2$
- generally ill-posed
- choose one solution according to a priori knowledge (*smoothness, etc.*)

Regularization Approach

- add a **stabiliser** $H[f] = \sum_{i=1}^{N}(f(\vec{x}_i) - y_i)^2 + \gamma\Phi[f]$

Derivation of Regularization Network

Stabilizer Based on Fourier Transform

[Girosi, Jones, Poggio, 1995]

- reflects some knowledge about the target function (usually smoothness, etc.)
- penalize functions that oscillate too much
- stabilizer in a form:

$$\Phi[f] = \int_{R^d} d\vec{s} \frac{|\tilde{f}(\vec{s})|^2}{\tilde{G}(\vec{s})}$$

$\tilde{f}$	Fourier transform of f	$\tilde{G}(\vec{s}) \to 0$ for $\|\|s\|\| \to \infty$
$\tilde{G}$	positive function	$1/\tilde{G}$ high-pass filter

Derivation of Regularization Network

Form of the Solution

- for a wide class of stabilizers (G positive semi-definite) the solution has a form

$$f(x) = \sum_{i=1}^{N} w_i G(\vec{x} - \vec{x}_i)$$

- where weights w_i satisfy

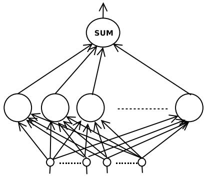
$$(\gamma I + G)\vec{w} = \vec{y}$$

- represented by a feed-forward neural network with one hidden layer

Regularization Network

Network Architecture

$$f(x) = \sum_{i=1}^{N} w_i G(\vec{x} - \vec{x}_i)$$



- function G called basis or kernel function

Basic Algorithm

1. set the centers of kernel functions to the data points
2. compute the output weights by solving linear system

$$(\gamma I + K)\vec{w} = \vec{y}$$

Model Selection

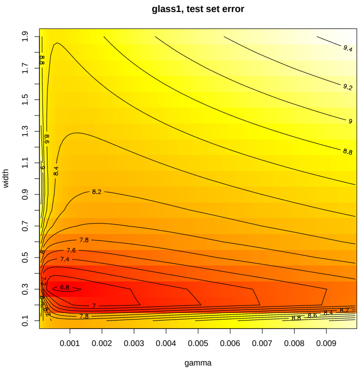
Parameters of the Basic Algorithm

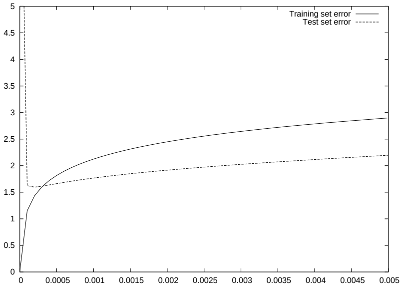
- kernel type
- kernel parameter(s) (i.e. width for Gaussian)
- regularization parameter γ

How we estimate these parameters?

- kernel type by user
- kernel parameter and regularization parameter by grid search and cross-validation
- speed-up techniques: grid refining

Role of Regularization Parameter





Role of Kernel Function

Choice of Kernel Function

- choice of a stabilizer
- choice of a function space for learning (hypothesis space)

Role of Kernel Function

- represent our prior knowledge about the problem
- *no free lunch* in kernel function choice
- should be chosen according to the given problem
- what functions are good first choice?

Experiments

Data

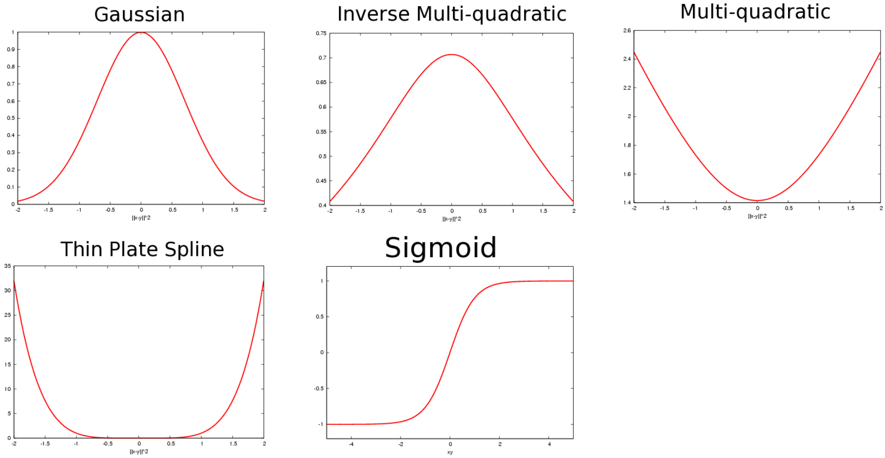
- Lenna image – approximation
- benchmark data sets – Proben1 data repository

Methodology

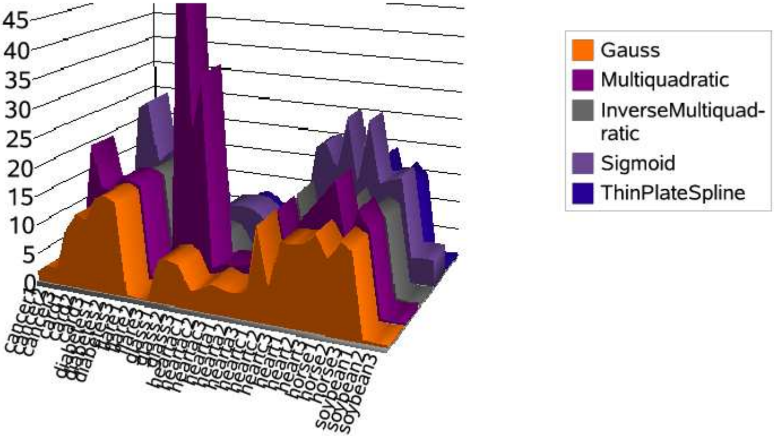
- separate data for training and testing
- find suitable γ on training set by cross-validation
- learn on training set (estimation of weights w)
- evaluate error on testing set – generalization ability

Lenna – Approximation

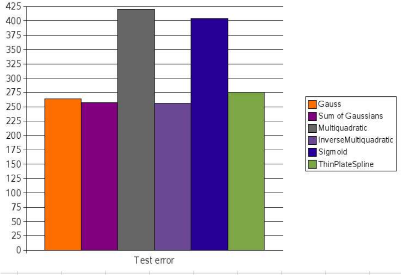
Proben1 – Comparison of Kernel Functions



Proben1 – Comparison of Kernel Functions

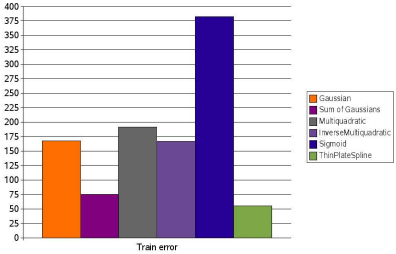


Comparison of Test Errors



- inverse multi-quadratic (20 tasks)
- Gaussian function
- local response

Comparison of Training Errors



- thin plate spline
- multi-quadratic
- sum of two Gaussians

- good generalization without reg. member

Summary and Future Work

Summary

- learning with RN networks described
- role of kernel function discussed
- impact of kernel function choice demonstrated
- different kernel functions compared

Work in Progress and Future Work

- kernel functions for other data types (categorical data, etc.)
- composite types of kernels

Thank you! Questions?