Meta-parameters of kernel methods and their optimization

Petra Vidnerová Roman Neruda

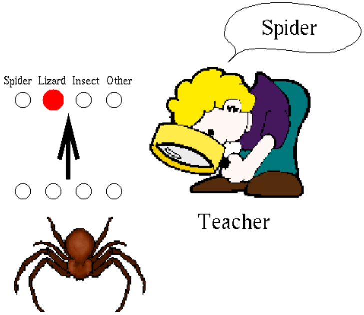
Institute of Computer Science
Academy of Sciences of the Czech Republic

ITAT 2014

Motivation

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.

Motivation

Learning methods

- wide range of methods available
  - statistical approaches
  - neural networks (MLP, RBF networks, etc.)
  - kernel methods (SVM, etc.)

Learning steps

- data preprocessing, feature selection
- model selection
- parameter setup

Motivation

Aim of this work

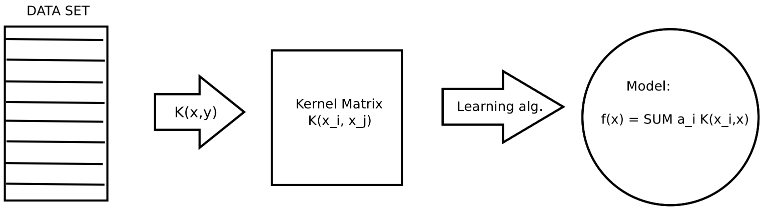
- some experience needed to achieve best results
- our ultimate goal - automatic setup
  - model recomendation
  - meta-parameters setup
- in this talk: meta-parameters setup for the family of kernel models

Outline

- brief overview SVM, RN
- role of kernel function
- meta-parameters optimisation methods
- some experimental results

Kernel methods

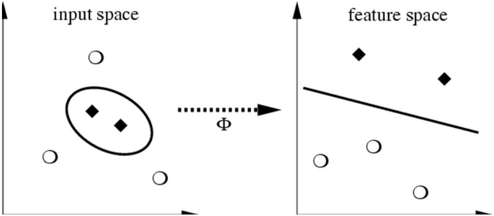
- family of models, became famous with SVM
- learning schema
  1. data is processed into a kernel matrix
  2. learning algorithm applied using only the information in the kernel matrix
- resulting model - linear combination of kernel functions



Kernel methods - basic idea

- choose a mapping to some (high dimensional) dot-product space - *feature space*

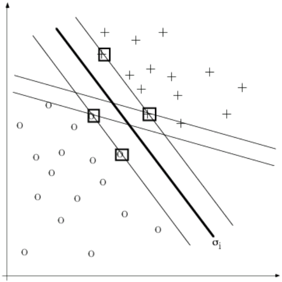
  $\Phi : \mathcal{X} \rightarrow \mathcal{H}$



- work in feature space
- dot product in feature space given by kernel fucntion $K(\cdot, \cdot)$

Support Vector Machine

- classification task
- input points are mapped to the feature space
- classification via separating hyperplane with maximal margin
- such hyperplane is determined by support vectors

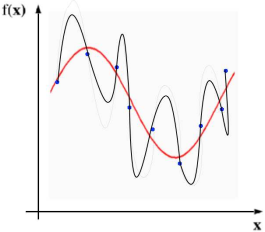


- many implementations available, i.e. libSVM
- parameter setup includes:
  - kernel function
  - *C* trade-of between maximal margin and minimum training error

Regularization Networks

approximation tasks, neural networks with one hidden layer

- given $\{(\vec{x}_i, y_i) \in R^d \times R\}_{i=1}^N$, recover the unknown function
- 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
- penalize functions that oscillate too much

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

$\tilde{f}$ Fourier transform of f

$\tilde{G}$ positive function

$\tilde{G}(\vec{s}) \to 0$ for $||s|| \to \infty$

$1/\tilde{G}$ high-pass filter

- for a wide class of stabilizers the solution has a form

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

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

- meta-parameters: G kernel function, γ

Role of Kernel Function

Choice of Kernel Function

- choice of a stabilizer
- choice of a function space for learning (hypothesis space)
- geometry of the feature space
- represent our prior knowledge about the problem
- should be chosen according to the given problem

Frequently used kernel functions

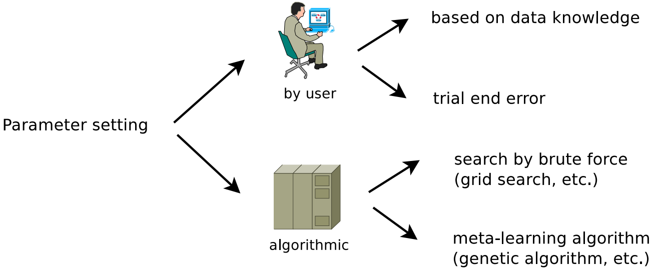
- linear $K(\vec{x}, \vec{y}) = \vec{x}^T \vec{y}$
- polynomialial $(\vec{x}, \vec{y}) = (\gamma \vec{x}^T \vec{y} + r)^d, \gamma > 0$
- radial basis function $(\vec{x}, \vec{y}) = exp(-\gamma ||\vec{x} - \vec{y}||^2), \gamma > 0$
- sigmoid $(\vec{x}, \vec{y}) = tanh(\gamma \vec{x}^T \vec{y} + r)$

Toy example - image approximation

Meta-parameters setup

Parameters of kernel learning algorithms

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

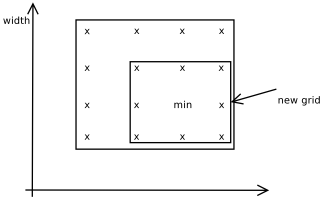


Search for optimal meta-parameters

- minimization of cross-validation error
- winning parameters used for training on the whole data set

Grid search

- extensive search, various couples of parameters tried
- time consuming
- start with coarse grid, than make finer
- quite standard way, implemented for example in libSVM



b)

Search for optimal meta-parameters

Genetic algorithm

- robust optimisation technique
- often used in combination with learning algorithms or NNs
- individuals coding kernel function, its parameters, regularization parameter $I = \{K, p, \gamma\}$

Simulated annealing

- stochastic optimisation method
- search
- least number of evaluations

Thank you! Questions?