Systems Neuroscience
& Neurotechnology Unit

Kernel Machines
Support Vector Machine Classification

Prof. Dr. Dr. Daniel J. Strauss

Overview

Support Vector Machine Classification

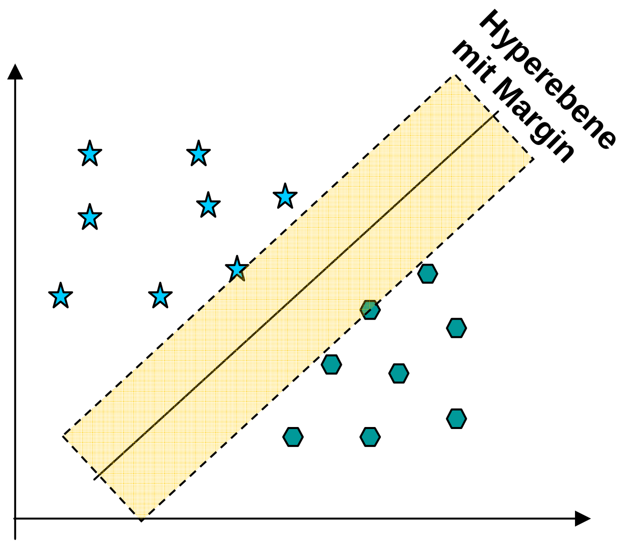
- Inducing Feature Spaces by Reproducing Kernels
- Regularization in RKHS and the Optimal Hyperplane
- Solution of the RKHS Regularization Problem

Support Vektor Machines - SVMs

- Developed by Vladimir Vapnik & Aleksei Chervonenkis in 1995
- Firmly grounded in the framework of statistical learning theory
- „Based on Support Vectors (SV)"
- Idea: Seperation of a dataset into two classes

  Hyperplane + Margin
  - High accuracy and low error probability (*„Hard margin"*)

- Classification → Boolean value (true/false; 1/-1)
- Regression → Real value or number



Initial Situation

- Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ containing the data to be classified.
  We are given a training set

$$\mathcal{A} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, 1\} : i = 1, \ldots, M\}$$

  of M associations.
- Suppose that there exists unknown *target function t* mapping $\mathcal{X}$ to $\{-1, 1\}$
- Interested in „good approximation „ of *t* (i.e. use sgn(f)) which classifies the training data correctly.

→ *Search for the Hypothesis function f in some reproducing kernel Hilbert spaces (RKHS)*

Why Positive definite kernel functions K ??

- Rate of computation can be highly reduced

Thus, we are interested in functions K arising from RBFs (e.g. Gaussian) so that

$$K(\mathbf{x}, \mathbf{y}) = k(||\mathbf{x} - \mathbf{y}||_2),$$

1.1 RKHS

- For a given K, there exists a RKHS

$$\mathcal{H}_K = \overline{\text{span}\{K(\tilde{\mathbf{x}}, \cdot) : \tilde{\mathbf{x}} \in \mathcal{X}\}}$$

of real valued functions on X with inner product. K is the reproducing kernel on $\mathcal{H}_K$

•→ Given a p.d. kernel K over X, we can find a Hilbert space $\mathcal{H}$ with reproducing kernel K

Mercer's Theorem

This theorem allows us to transfer the euclidian space into the high dimensional Feature-space.
Note: K has to be p.d. (that is continuous with finite trace), then there exists an infinite sequence of eigenfunctions $<\varphi_j>_{i=1}^{\infty}$
and eigenvalues $\eta_j \geq 0$, and that we can write

$$K(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{\infty} \eta_j \varphi_j(\mathbf{x}) \varphi_j(\mathbf{y}).$$

Feature Map & Space

- Let us introduce the **feature map** $\Phi : X \subset IR^n \to \ell^2$ by

$$\Phi(\cdot) = \left(\sqrt{\eta_i}\,\alpha(\cdot)\right)_{i \in IN}$$

which is induced by a kernel *K* of a reproducing kernel Hilbert space

$$K(\mathbf{x},\mathbf{y}) = \sum_{i \in IN} \eta_i \alpha(\mathbf{x})\alpha(\mathbf{y}) \qquad \|\Phi(\mathbf{x})\|^2_{\ell^2} = \sum_{i \in IN} \eta_i \alpha_i(\mathbf{x}) = K(\mathbf{x},\mathbf{x})$$

- We define the **feature space** $F_K \subset \ell^2$ by

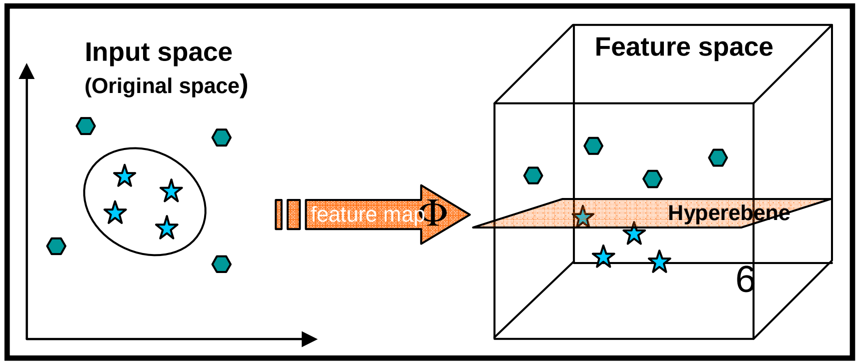
$$F_K = \overline{\text{span}\{\Phi(\mathbf{x}) : \mathbf{x} \in X\}} \quad \Leftrightarrow \quad K(\vec{x},\vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_F$$

Kernel evaluation of x,y is equal an inner product in the high-dimensional Feature space.

The map Φ is just induced by the reproducing kernel

→ Map does not have to be calculated, solution of the kernel is know and is lying in the high dimensional space

We do not know the map Φ
we only know the reproducing kernel



Unconstrained Optimization Problem

Let us turn to our classification task. For a given training set we intend to construct a function $f \in \mathcal{H}_K$ which minimizes

„Cost function“ — Smothness term

$$\lambda \sum_{i=1}^{M} (1 - y_i f(\mathbf{x}_i))_+ + \frac{1}{2} ||f||^2_{\mathcal{H}_K},$$

Regularisation-parameter

Term evaluates how much error we produce (empirical risk)

Smothness in original space means pushing points away in the Featurespace

The goal here: Find a function f which minimize the Cost (error) and smothness

⟺ Constrained OP $\lambda \left(\sum_{i=1}^{M} u_i \right) + \frac{1}{2} ||f||^2_{\mathcal{H}_K},$

Keep in mind

Feature-space formulation of the OP

Every function $f \in \mathcal{H}_K$ corresponds uniquely to a sequence $\mathbf{w} \in \mathcal{F}_K$. Thus
The optimization problem can be rewritten

$$\lambda\left(\sum_{i=1}^{M} u_i\right) + \frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K},$$

Soft margin $u_i \neq 0$

Optimal Hyperplane

- If the above mentioned OP, however, is fullfilled with $u_i = 0\ (i = 1, \ldots, M),$, then we say that our training set is linearly seperable in $\mathcal{F}_K$

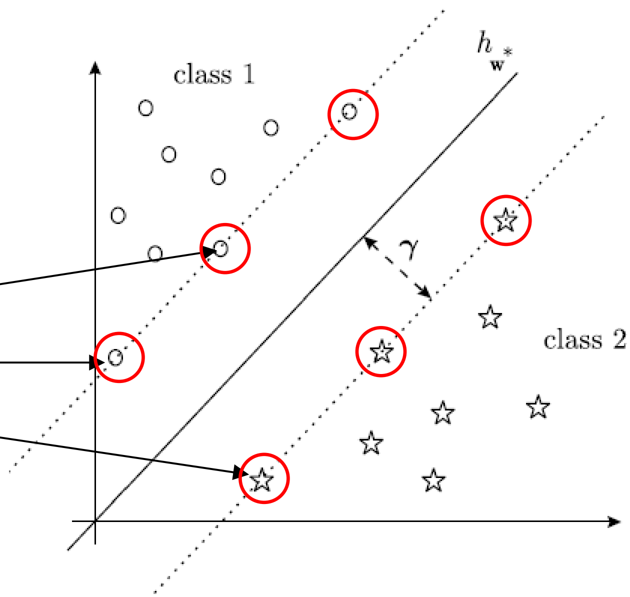
→ The OP can be further simplified to: find $\mathbf{w} \in \mathcal{F}_K$ to minimize

$$\frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K}$$

Hard margin

What does soft and hard margin (γ) mean??

- **„Support Vectors (SVs)“**
  - Define the orientation and position of the hyperplane



Large margin classifiction

Solution of the RKHS Regularization Problem

Here the notation *support vector (SV)* comes into the play.

By the <u>*Representer Theorem*</u>, the minimizer of $\lambda\left(\sum_{i=1}^{M} u_i\right) + \frac{1}{2}||\mathbf{w}||^2_{\mathcal{F}_K}$, i.e., the hyphotesis function, has the form

$$f(\mathbf{x}) = \sum_{j=1}^{M} c_j K(\mathbf{x}, \mathbf{x}_j).$$

What do you think could be done to solve the regularization problem regarding this formular ?

Can we play around with our reproducing kernel ?

What remains ?

„Here we have to rotate the knob“
→Search cj so that the resulting function can solve the
Optimization problem (minimize Cost-function and smoothness-term)

Figure: Points defining our SVs are those points having $c_i \neq 0$ (c don‘t vanish)

SV count theorem

- The fewer the number of SVs the better generalization of the SVM can be expected
- The fewer SVs the better can be classified (capacity decreases)