Sören Sonnenburg[†], Gunnar Rätsch[♮], Fabio De Bona[♮]

Fraunhofer FIRST Institut Rechnerarchitektur und Softwaretechnik

MAX-PLANCK-GESELLSCHAFT

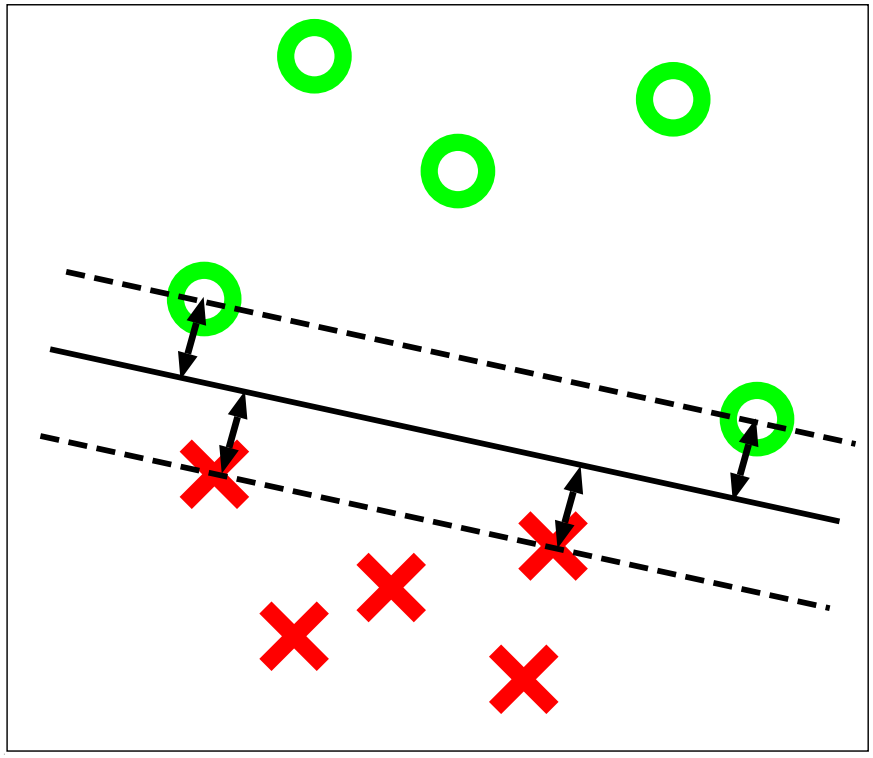
[†] Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
[♮] Friedrich Miescher Laboratory of the Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany

Machine Learning Toolbox SHOGUN features algorithms:

- to learn 2-class classification and regression problems
- to train hidden markov models
- toolbox's focus is on kernel methods esp. Support Vector Machines (SVMs) for computational biology
- also implements a number of linear methods like Linear Discriminant Analysis (LDA), Linear Programming Machine (LPM), (Kernel) Perceptrons

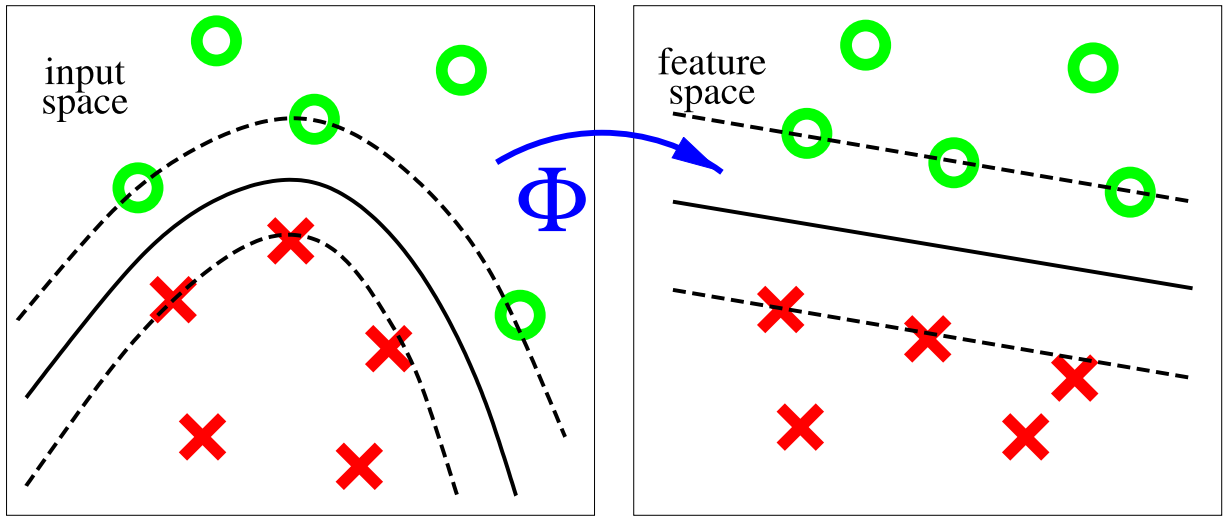
Support Vector Machine

- given: points $\boldsymbol{x}_i \in \mathcal{X}$ $(i = 1, \ldots, N)$ with respective labels $y_i \in \{-1, +1\}$
- in training hyperplane that maximizes margin is chosen



Decision function $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$

SVM with Kernels



- SVM decision function in kernel feature space:

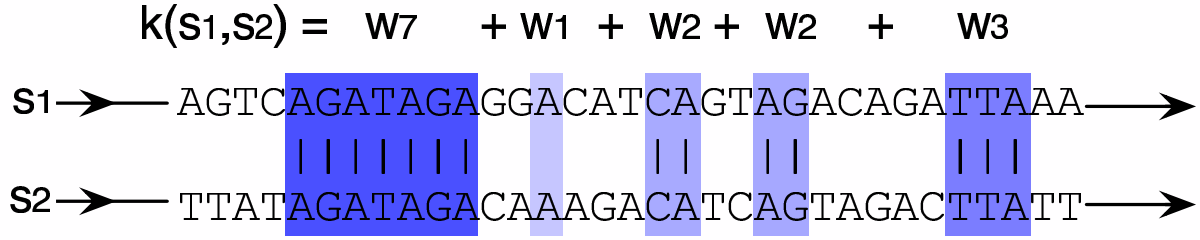
$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i \underbrace{\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x}_i)}_{=\mathrm{k}(\boldsymbol{x},\boldsymbol{x}_i)} + b \tag{1}$$

- find parameters $\boldsymbol{\alpha}$ by solving quadratic optimization problem

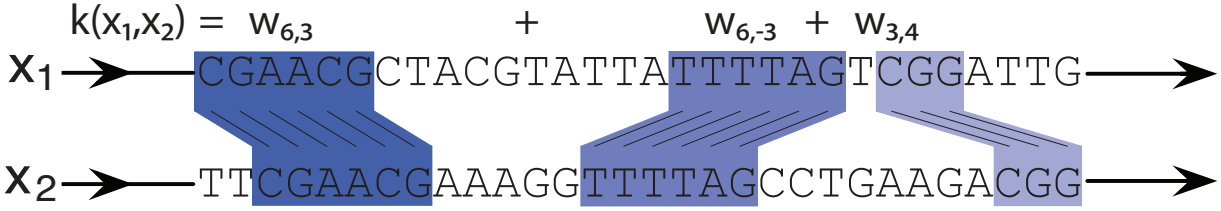
String Kernels

- Spectrum Kernel
  - count k-mers in each sequence, Spectrum Kernel is sum of product of counts

- Weighted Degree Kernel



- Weighted Degree Kernel with Shifts



Features

- SHOGUN interfaces to Matlab^{TM}, Octave and Python and R
- provides generic SVM object interfacing to several different SVM implementations, among them the state-of-the-art LibSVM and SVM^{light}
- SVMs can be trained using a variety of common kernels (Linear, Polynomial, Gaussian) and recent String Kernels (TOP, Fisher, Locality Improved, Spectrum, Weighted DegreeKernel (with shifts))
- kernels can be combined; weighting can be learned using Multiple Kernel Learning.
- input feature-objects can be dense, sparse or strings and of type int/short/double/char; can be converted into different feature types.
- multiprocessor parallelization $\Rightarrow$ able train on **10 million** examples

. . . and many more. . .

Tackled Bioinformatics Problems

- Protein Super Family classification
- Splice Site Prediction (*C.elegans*, *Drosophila*, *Human* etc.)
- Alternative Splice Site Prediction (Exon Skipping, Intron retention, alternative 3' or 5' ends)
- Interpreting the SVM Classifier
- Splice Form Prediction (Learn segmentation)
- Promoter Prediction

$\Rightarrow$ **very generic**

Bioinformatics Demo:

- Position Independent (e.g. Tissue Classification using Promotor Region)

```
AAACAAAAACGTAACTAATCTTTTAGAGAGAACGTTTCAACCATTTTGAG
AAGATTAACTCATCACAGATTTCATTACATACAGATATAATTCAAAAATT
CACTCCCCAAATCAACGATATTTAAAAATCACTAACACATCCGTCTGTGC
```

  - Task: separate DNA strings, '-' class random ACGT, '+' class contains 'AAAAA' motif

- Position Dependent (e.g. Splice Site Classification)

```
AAACAAATAAGTAACTAATCTTTTAAGAAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTTTAACATTACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTTAATTCACTAACACATCCGTCTGTGCC
```

  - Task: separate DNA strings, '-' class random ACGT, '+' class 'AA' in the middle

- Mixture Position Dependent/Independent (e.g. Promoter Classification)

```
AAACAAATAAGTAACTAATCTTTTAAAGAGAACGTTTCAACCATTTTGAG
AAGATTAAAAAAAAACAAATTTCATTAAATACAGATATAATAATCTAATT
CACTCCCCAAATCAACGATATTTAAATTTCACTAACACATCCGTCTGTGC
```

  - Task: separate DNA strings, '-' class random 'ACGT', '+' class 'AAA' in the middle shifted ± 15

Generic Demo:

- Support Vector Classification
  - Task: separate 2 clouds of gaussian distributed points in 2D
- Support Vector Regression
  - Task: learn a sine function
- Hidden Markov Model
  - Task: 3 loaded dice are drawn 1000 times, find out when which dice was drawn

Summary

- SHOGUN is a large scale machine learning toolbox
  $\Rightarrow$ able to train on **10 million** examples

- unified SVM framework + many string kernels suitable for comp. biology

- Algorithms: HMM, LDA, LPM, Perceptron, SVM, SVR + many kernels

We need your help:

- Documentation
- Examples
- Testing
- Test Suite

Source Code is freely available under the GPLv2.

`http://www.fml.tuebingen.mpg.de/raetsch/projects/shogun`