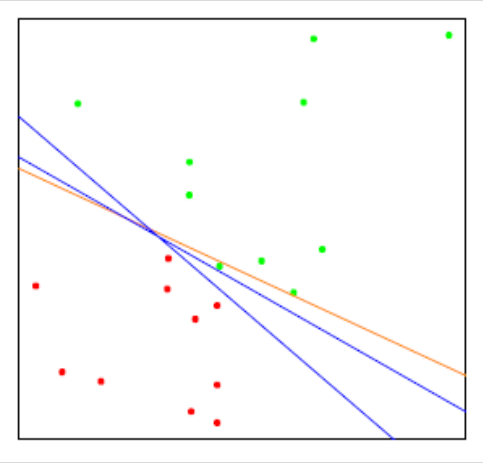


Big Data Analytics

Basic concepts of analyzing very large amounts of data

Dr. – Ing. Morris Riedel
Adjunct Associated Professor
School of Engineering and Natural Sciences, University of Iceland
Research Group Leader, Juelich Supercomputing Centre, Germany

LECTURE BDA1

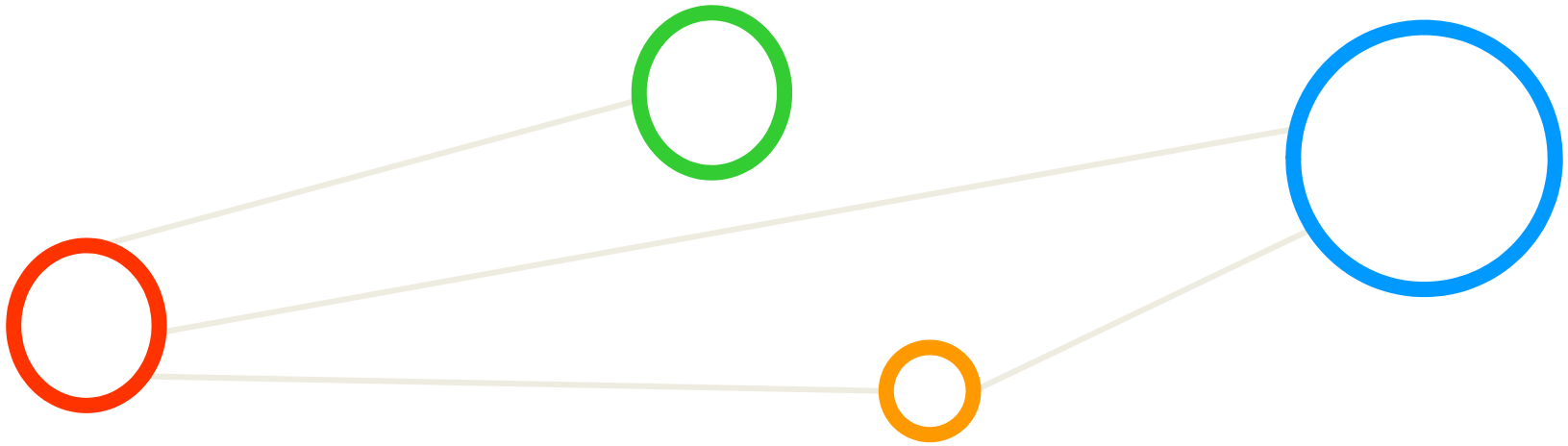
Support Vector Machines

2014-01-24 (v4)
Online Material

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

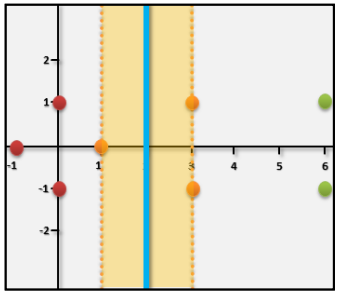
JÜLICH
FORSCHUNGSZENTRUM

Outline

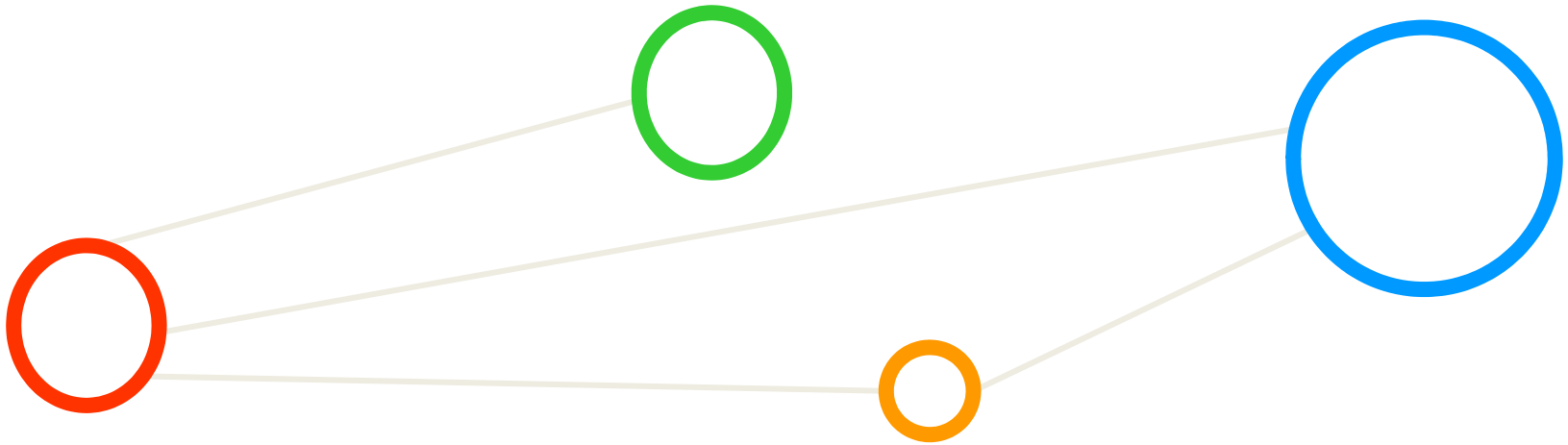


Outline

- SVMs
  - General Information
  - Need for Labelled Data
  - Advantages / Disadvantages
  - Linear Example
- Step-Wise Approach
  - Categorical feature transformations
- Application Areas
  - Text Categorization
  - Bioinformatics
  - Image Processing



Support Vector Machines (SVMs)



General Information

- Support Vector Machines (SVMs)
  - Appeared in the early 1990s
  - Standard methodology in computer science and engineering communities
- Optimal margin classfiers
  - In the context of 'statistical learning theory'
- Operate within the framework of 'regularization theory'
  - Minimizing an 'empirical risk' in a well-posed and consistent way
- Compromise and Kernel Functions
  - Between the parametric and the pure nonparametric approaches
  - As in linear classiers: SVMs estimate a linear decision function
  - With the particularity that a previous mapping of the data into a 'higher-dimensional feature space' may be needed
  - Mapping is characterized by the choice of a class of functions (known as kernels)

Need for Labelled Data

- Sample Dataset $\{(\mathbf{x}_i, \mathbf{y}_i) \in X \times Y\}_{i=1}^{n}$

- Goal is to learn the relationship between the **x** and **y** variables
- The main goal in this context usually is 'predictive accuracy'
  - In most cases it is not possible to assume a parametric form for the probability distribution p(**x**, **y**)

Advantages / Disadvantages

- Advantages
  - Sparse solutions to classication and regression problems are obtained
  - Only a few samples are involved in the determination of the classication (or regression functions)

Literature

- Books
  - ‘Introduction to SVMs’, Cristianini and Shawe-Taylor

[2] Introduction to SVMs

SVM Foundations

- **SVM make use of a (nonlinear) mapping function ϕ that transforms data in input space to data in feature space in order to render a problem linearly separable**
- **SVM then automatically discovers the optimal separating hyperplane**
- **When mapped back into input space via ϕ^{-1} , this plane can be a complex decision surface**

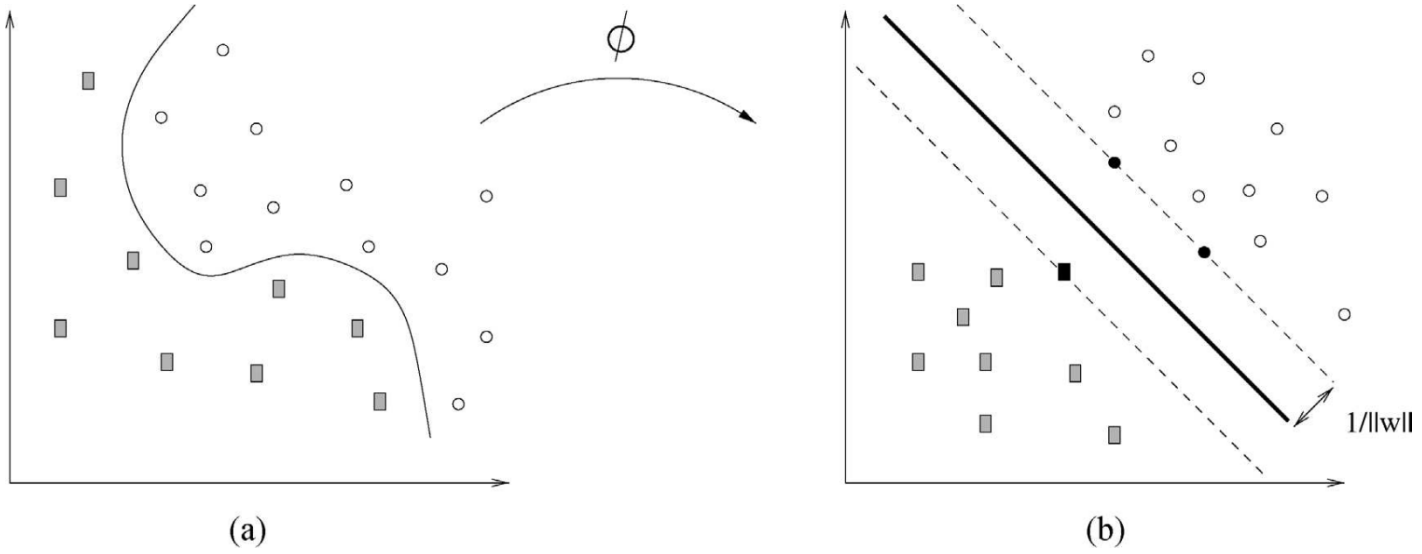
- Support Vector Machines (SVMs)
  - Represent a popular classification technique
  - Based on a sound theoretical basis
  - Have state-of-the-art success in real-world applications

SVM Principles

- Classification problem
  - Discriminant function is nonlinear
  - (a) Original data in the input space
  - (b) Mapped data in the feature space – e.g. two possible classes (mapping function ϕ such that the data becomes linearly seperable)

$$\{(\Phi(\mathbf{x}_i), y_i)\}_{i=1}^{n}$$

$$y_i \in \{-1, +1\}$$

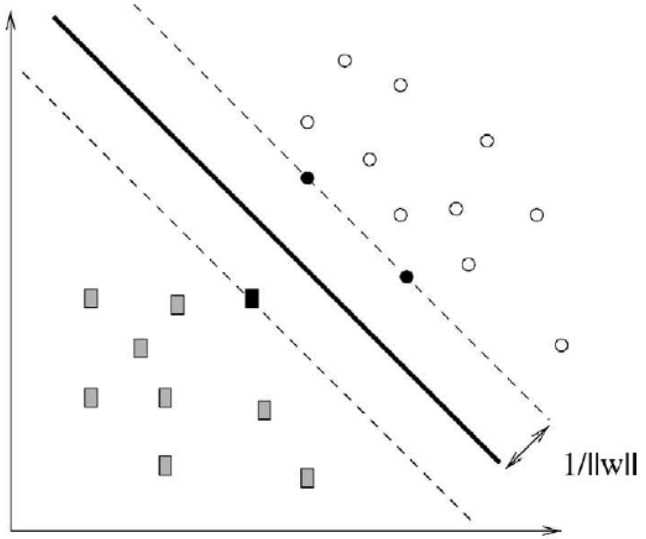


[1] SVMs with Applications

Optimal (Maximal) Margin Hyperplane

- There is an infinite number of existing separating hyperplanes
- Separating hyperplane in the space of the mapped data:
  - (equidistant to the nearest point in each class)

$$\mathbf{w}^T \Phi(\mathbf{x}) + b = 0$$



- Rescale

$$|\mathbf{w}^T \Phi(\mathbf{x}) + b| = 1$$

every $i \in \{1, \ldots, n\}$

  - (assuming separability) for those points in each class nearest to the hyperplane

$$\mathbf{w}^T \Phi(\mathbf{x}_i) + b \begin{cases} \geq 1, & \text{if } y_i = +1 \\ \leq -1, & \text{if } y_i = -1. \end{cases}$$

- **SVMs pick the separating hyperplane that lies furthermost both clases**
- **The separating plane is called optimal (maximal) margin hyperplane**

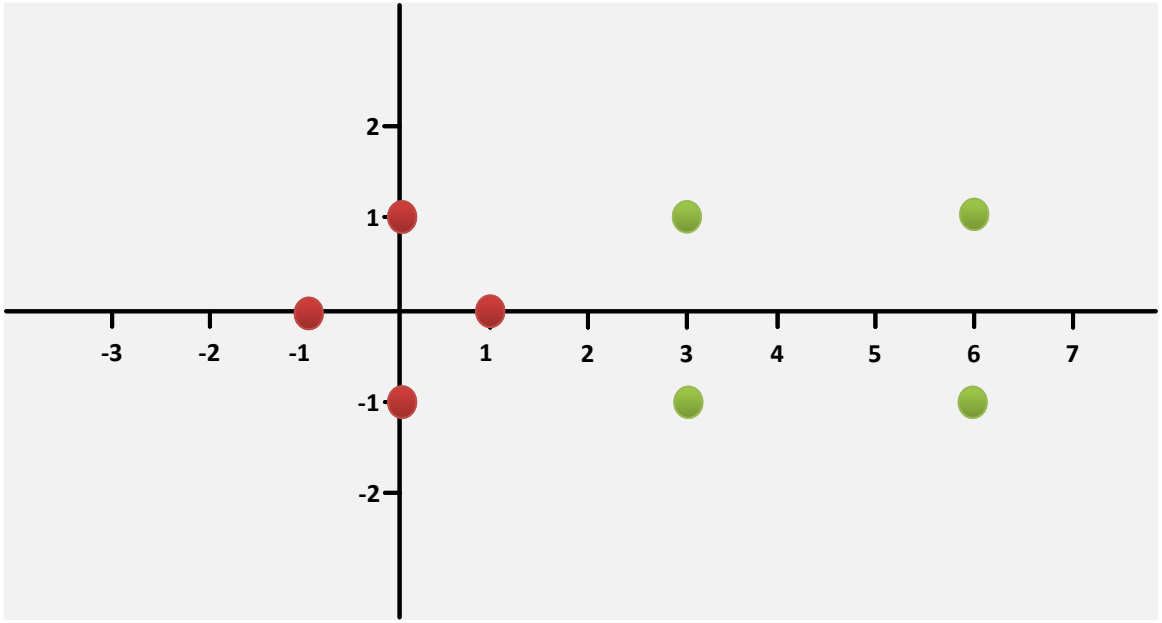
Linear SVM Example (1)

- Given: positively labelled data points in $\mathbb{R}^2$: red (class -1)

$$\left\{\begin{pmatrix}3\\1\end{pmatrix},\begin{pmatrix}3\\-1\end{pmatrix},\begin{pmatrix}6\\1\end{pmatrix},\begin{pmatrix}6\\-1\end{pmatrix}\right\}$$

- Given: negatively labelled data points in $\mathbb{R}^2$: green (class +1)

$$\left\{\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}-1\\0\end{pmatrix}\right\}$$



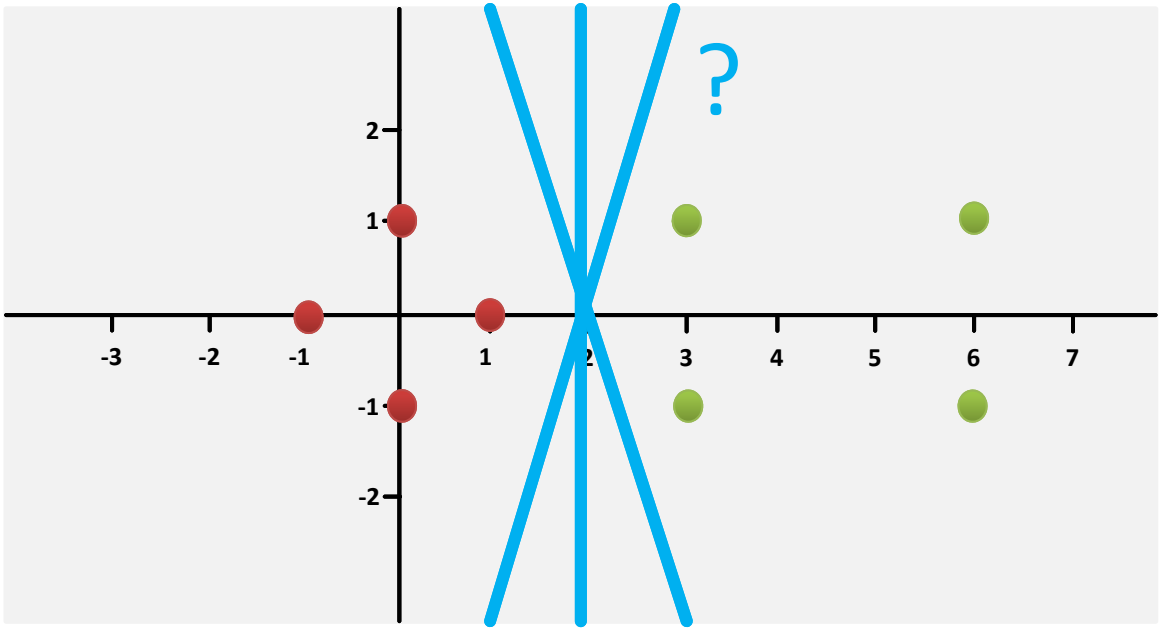
- $\mathbb{R}$: **Real numbers can be represented as points on an infinitely long number line**
- **Real numbers include all the rational numbers, irrational numbers, & Pi**

[3] SVM Example

Linear SVM Example (2)

- **A 'Linear SVM' is a SVM whose mapping function ϕ() is the identity function**
- **The 'identity function' is a function that always returns the same value that was used as ist argument: e.g. f (x) = x**

- Goal: Discover a SVM that accurately discriminates two classes
- Simple Analysis of Input Space:
  - Data is linear separable → Linear SVM should be ok (finds optimal plane)



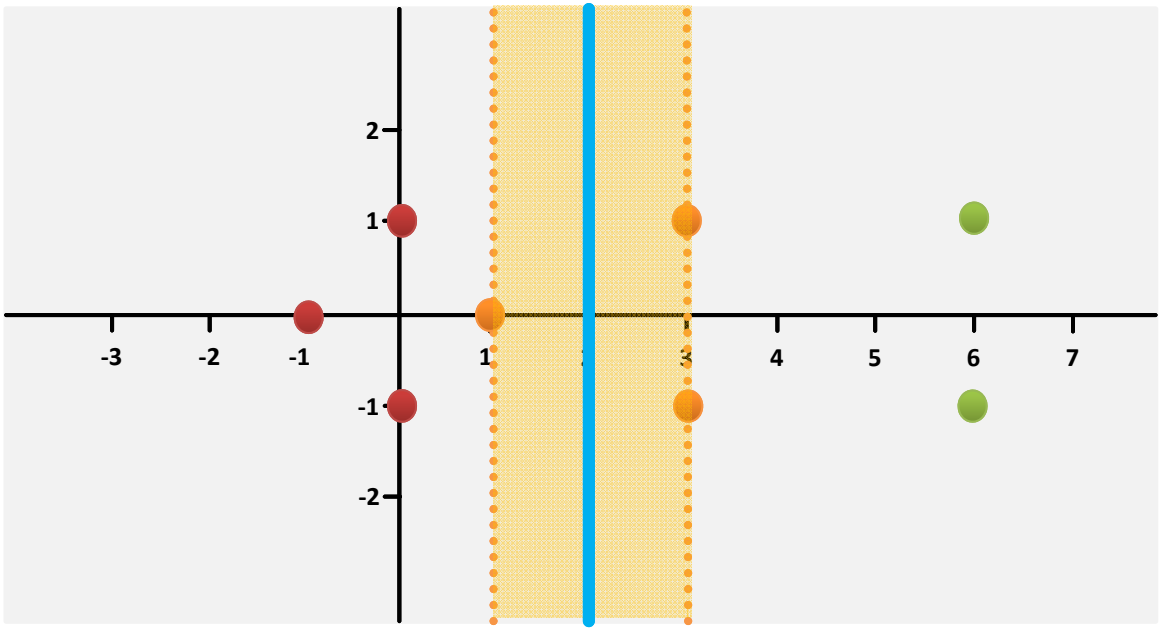
[3] SVM Example

Linear SVM Example (3)

- **Support Vectors lie directly on the margin and satisfy $y_i = (w \cdot x_i - b) = 1$**

- (Manual) Inspection of the data
  - Three support vectors (orange)

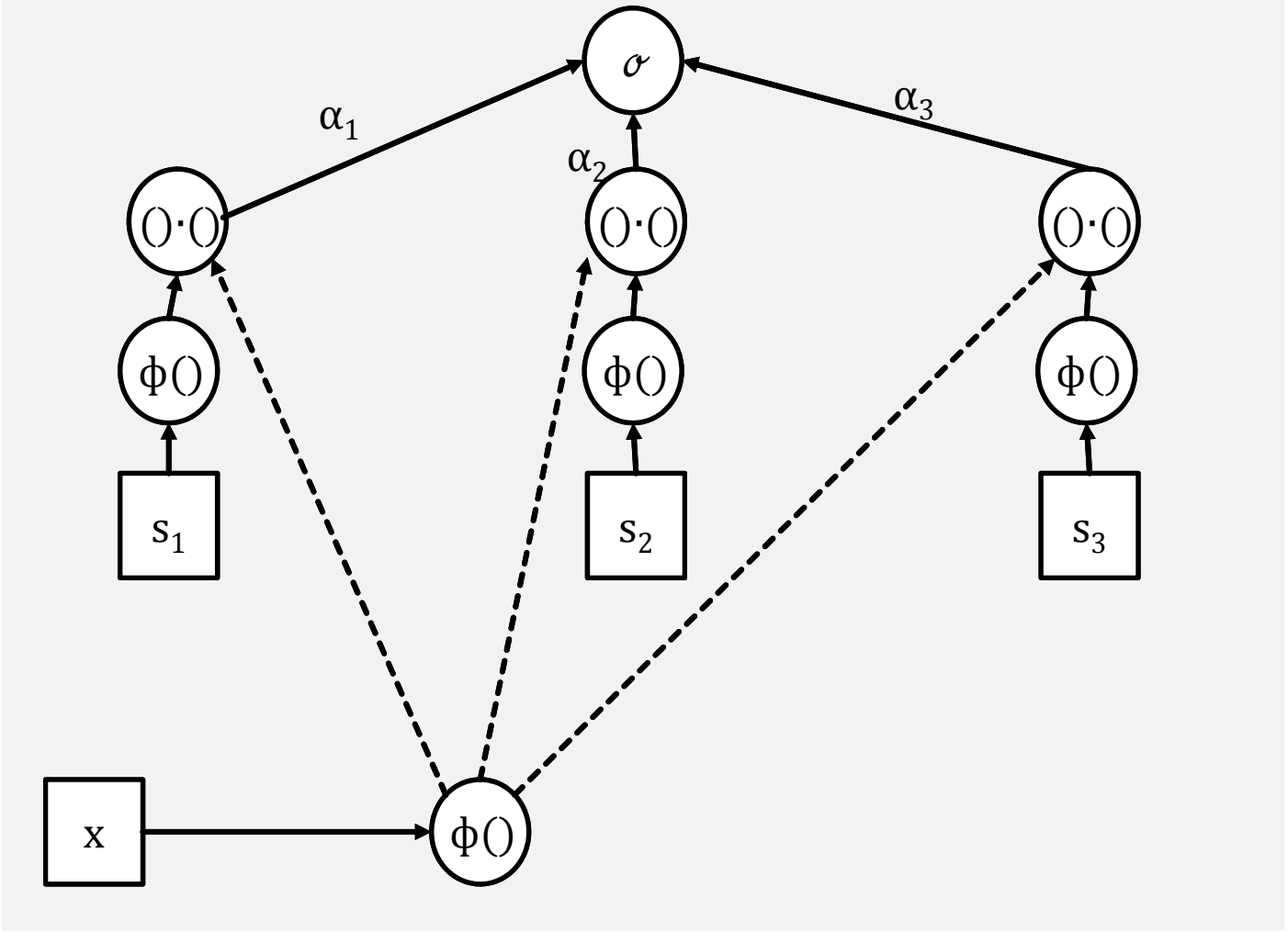
$$\left\{s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}\right\}$$



[3] SVM Example

Linear SVM Example (4)

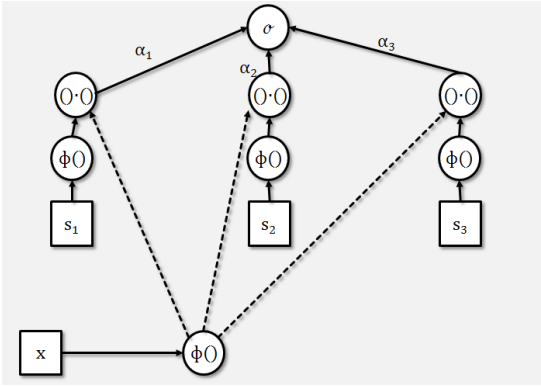
- The ‘SVM Architecture‘



[3] SVM Example

Linear SVM Example (5)

- Use vectors augmented with a 1 as a bias input
  - Differentiated with over-tilde
  - E.g. $s_1 = (10)$, with bias $\tilde{s}_1 = (101)$



- Goal: Find values for α_i so that:

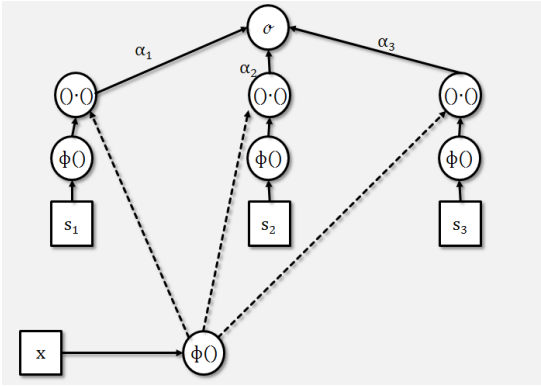
$$\begin{aligned}
\alpha_1 \phi(s_1) \cdot \phi(s_1) + \alpha_2 \phi(s_2) \cdot \phi(s_1) + \alpha_3 \phi(s_3) \cdot \phi(s_1) &= -1 \\
\alpha_1 \phi(s_1) \cdot \phi(s_2) + \alpha_2 \phi(s_2) \cdot \phi(s_2) + \alpha_3 \phi(s_3) \cdot \phi(s_2) &= +1 \\
\alpha_1 \phi(s_1) \cdot \phi(s_3) + \alpha_2 \phi(s_2) \cdot \phi(s_3) + \alpha_3 \phi(s_3) \cdot \phi(s_3) &= +1
\end{aligned}$$

- **Sometimes it is required that the hyperplane pass through the origin of the coordinate system for simplicity ('unbiased hyperplane')**
- **General hyperplanes not necessarily passing through the origin ('biased hyperplane')**

[3] SVM Example

Linear SVM Example (6)

- Using $\phi() = I$
  (Identity function, because linear SVM)
  - Adding bias for $s\sim_i$



$$\begin{aligned}
\alpha_1 s\sim_1 \cdot s\sim_1 + \alpha_2 s\sim_2 \cdot s\sim_1 + \alpha_3 s\sim_3 \cdot s\sim_1 &= -1 \\
\alpha_1 s\sim_1 \cdot s\sim_2 + \alpha_2 s\sim_2 \cdot s\sim_2 + \alpha_3 s\sim_3 \cdot s\sim_2 &= +1 \\
\alpha_1 s\sim_1 \cdot s\sim_3 + \alpha_2 s\sim_2 \cdot s\sim_3 + \alpha_3 s\sim_3 \cdot s\sim_3 &= +1
\end{aligned}$$

- Recall: Given three points with added bias

$$\left\{ \boldsymbol{s}\sim_1 = \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{1} \end{pmatrix}, \boldsymbol{s}\sim_2 = \begin{pmatrix} \mathbf{3} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix}, \boldsymbol{s}\sim_3 = \begin{pmatrix} \mathbf{3} \\ \mathbf{-1} \\ \mathbf{1} \end{pmatrix} \right\}$$

- Computing the dot product

$$\begin{aligned}
2\,\alpha_1 + 4\,\alpha_2 + 4\,\alpha_3 &= -1 \\
4\,\alpha_1 + 11\,\alpha_2 + 9\,\alpha_3 &= +1 \\
4\,\alpha_1 + 9\,\alpha_2 + 11\,\alpha_3 &= +1
\end{aligned}
\quad \xrightarrow{\textit{linear algebra}} \quad
\begin{aligned}
\alpha_1 &= -3.50 \\
\alpha_2 &= +0.75 \\
\alpha_3 &= +0.75
\end{aligned}$$

[3] SVM Example

Linear SVM Example (7)

$$\tilde{w} = \sum_i \alpha_i \tilde{s}_i$$

- α_i relates to the discriminating hyperplane with w:

$$w^T \phi(x) + b = 0$$

- Computing w~ using given points with bias and α_i :

$$\tilde{w} = -3.5 \begin{pmatrix}1\\0\\1\end{pmatrix} + 0.75 \begin{pmatrix}3\\1\\1\end{pmatrix} + 0.75 \begin{pmatrix}3\\-1\\1\end{pmatrix}$$

$$\tilde{w} = \begin{pmatrix}1\\0\\-2\end{pmatrix}$$

$$\alpha_1 = -3.50$$
$$\alpha_2 = +0.75$$
$$\alpha_3 = +0.75$$

$$\left\{ \tilde{s}_1 = \begin{pmatrix}1\\0\\1\end{pmatrix}, \tilde{s}_2 = \begin{pmatrix}3\\1\\1\end{pmatrix}, \tilde{s}_3 = \begin{pmatrix}3\\-1\\1\end{pmatrix} \right\}$$

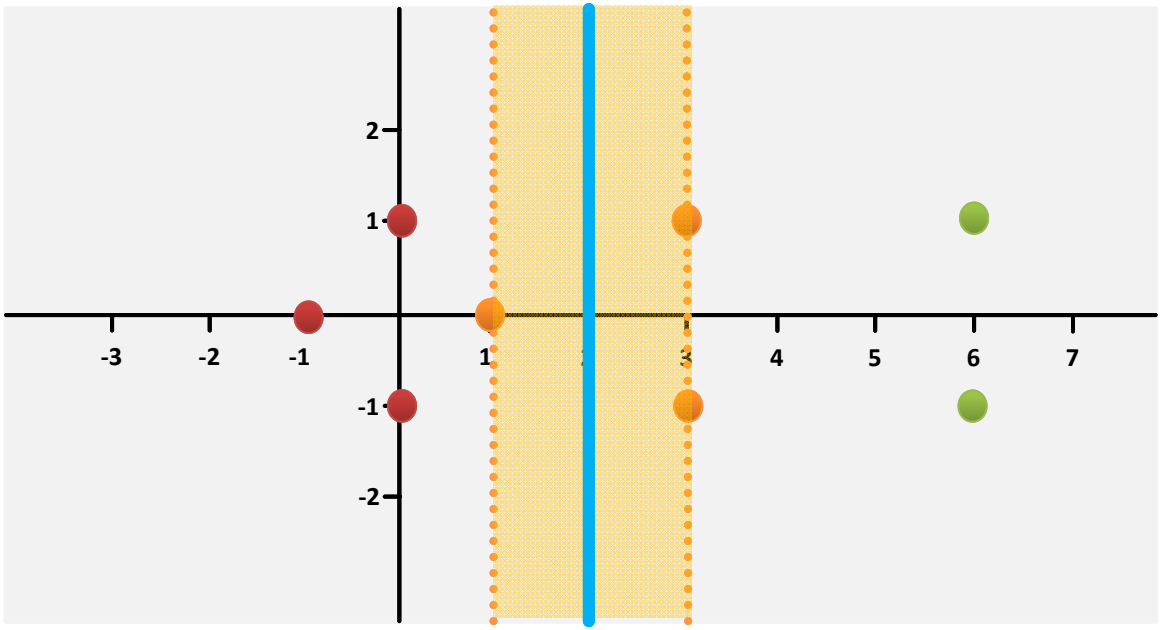
[3] SVM Example

Linear SVM Example (8)

- Context of linear hyperplane equation (linear function):

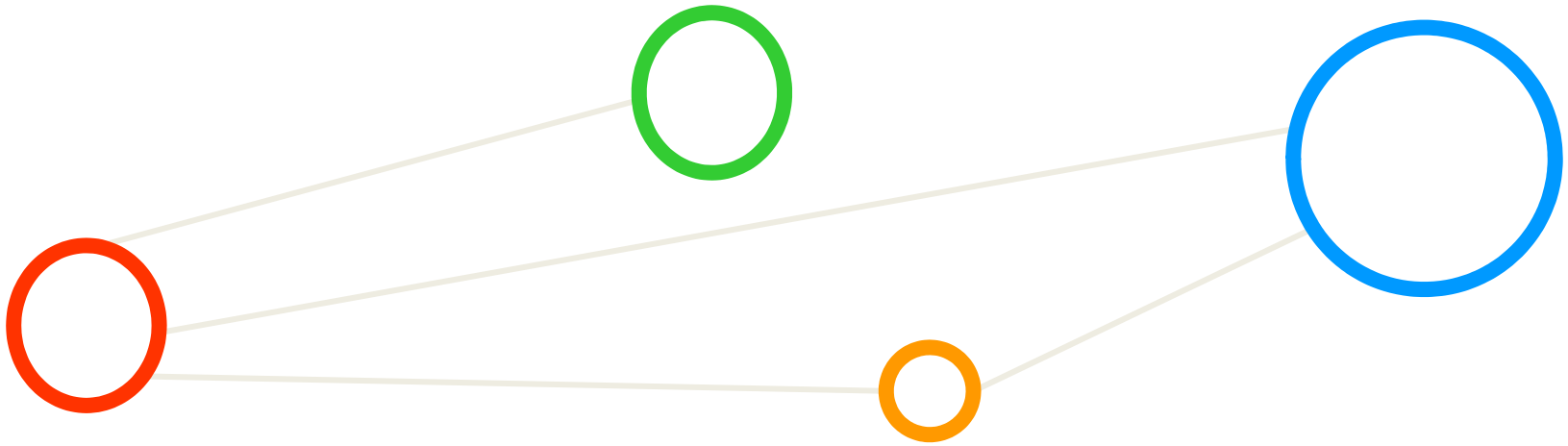
$$y = w\,x + b$$

$\tilde{w} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$ *bias b from vector* → $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, b = - 2



[3] SVM Example

Step-Wise Approach



Step-Wise Approach

- **SVM beginners often get unsatisfactory results, because they miss some easy but significant steps**

1. Transform data to SVM package format **Preprocessing**
2. Conduct simple ‘scaling‘ on the data **Preprocessing**
3. Consider the ‘Radial Basis Function (RBF)‘ Kernel K

$$K(x, y) = e^{-\Upsilon \, || \, x - y \, ||^2}$$

Model Selection

4. Use cross-validation to find best parameter C and ϒ **Model Selection**
5. Use C and ϒ to train the whole training set **Training**
6. Test **Testing**

[4] SVM Guide

Step 1: Categorical Feature

Preprocessing

- **SVM requires that each data instance is represented as a vector of real numbers**

- Problematic: Categorical attributes
  - Need to convert into a numeric data representation
- Example
  - Use m numbers to represent an m-category attribute
  - Only one of the m numbers is 1, others are 0

three-category attribute: {red, green, blue}

represented as

(0,0,1), (0,1,0), (1,0,0)

- Experience: if number of values in an attribute is not too large, this coding might be more stable than using a single number

[4] SVM Guide

Step 2: Simple Scaling (1) Preprocessing

- **Scaling before applying SVM is a very important process**
- **Scaling avoids attributes in greater numeric ranges dominating those in smaller numeric ranges**
- **It further avoids numerical dificulties during the calculation**

- Input data is often represented as vector
  - Apply ‘rescaling or normalization‘ of them (used term depend on field)
- Rescaling of vectors
  - Add/subtract a constant – then multiply/divide by a constant
- Normalizing a vector
  - Divide by a norm of the vector
    (e.g. make the Euclidean length of the vector equal to one)
  - Often refers to rescaling by the minimum and range of the vector
    (e.g. to make all the elements lie between 0 and 1)

[4] SVM Guide

[5] AI FAQ

Step 2: Simple Scaling (2) Preprocessing

- Scaling avoids numerical difficulties
  - Kernel values depend on inner products of feature vectors (e.g. the linear kernel and the polynomial kernel)
  - Especially large attribute values might cause numerical problems
- Ranges
  - ‘Linearly scaling‘ each attribute to certain ranges: [-1 ; +1] or [0 ; 1]
- Data Scaling
  - Use the same method to scale both training and testing data
- Example
  - Scale first attribute of training data from [-10; +10] to [-1; +1]
  - Then scale first attribute of testing data from [-11; +8] to [-1,1; +0,8]

[4] SVM Guide

Step 3: Radial Basis Function

Model Selection

- **Model selection for SVMs means to choose one of four possible ones: linear, polynomial, radial basis function, sigmoid**
- **Often recommended Kernel is Radial basis function**

- Linear

$$K(x_i, yj) = xiTxj$$

ϒ, r, d are kernel parameters

- Polynomial

$$K(x_i, yj) = (\Upsilon\, xiTxj + r)^d, \Upsilon > 0$$

- Sigmoid

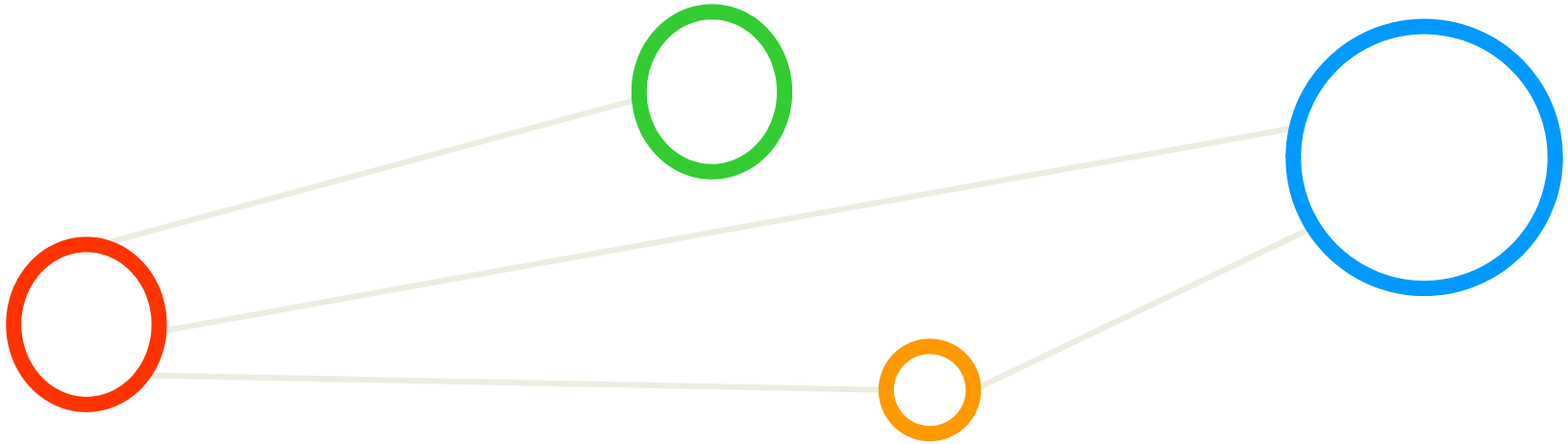
$$K(x_i, yj) = \tanh(\Upsilon\, xiTxj + r)$$

- Radial Basis Function (RBF)

$$K(x_i, yj) = e^{-\Upsilon\ ||\ xi - yj\ ||^2} = \exp(-\Upsilon\ ||\ x_i - y_j\ ||^2), \Upsilon > 0$$

[4] SVM Guide

Application Areas



Application Areas

- Only a few samples are involved in the determination of the classication (or regression functions)
  - This facilitates the application of SVMs to problems that involve a large amount of data (aka ‘big data’)
- Typical Application Areas
  - Text processing applications, e.g. text categorization
  - Bioinformatics tasks
  - Image recognition
- Application communities
  - Often involved in the solution of consulting
  - ‘Industrial data analysis’ problems

[1] SVMs with Applications

Text Categorization (1)

- Goal: Consists of the classication of documents into a predefined number of given categories
  - Used by many Internet search engines, e.g. to select Web pages related to user queries.
- Example: News Messages
  - Document collection made up of Usenet News messages
  - Organized in ‘predefined classes’: computation, religion, statistics, etc.
- Key Benefit: Automation of Classification Process
  - Input: new document
  - Task: conduct the ‘category assignment’ in an automatic way

[1] SVMs with Applications

Text Categorization (2)

- Mathematical Representation
  - Documents are represented in a vector space of 'dimension equal to the number of different words' in the vocabulary
  - Text categorization problems involve 'high-dimensional inputs' and the data set consists of a 'sparse document by term matrix'

[1] SVMs with Applications

Text Categorization (3)

- Data Set: Reuters data base
  - Text collection composed of 21,578 documents and 118 categories
  - The data space has dimension 9947
    (the number of different words that describe the documents)
- Example of SVM Performance
  - Results obtained using a SVM with a linear kernel
- First Measure: Average Rate of Success
  - SVM is 87%, Naive Bayes (72%), Bayesian Networks (80%), Classification Trees (79%), and *k*-nearest neighbors (82%)

- **Results for SVM with a linear kernel in text categorization are consistently better along categories than those obtained with four widely used classification methods (Naive Bayes, Bayesian networks, Classication trees, *k*-nearest neighbors)**

[1] SVMs with Applications

Text Categorization (4)

- **SVM text classifiers training time is most impressive benefit**
- **Much faster than other classification methods such as Naive Bayes or Classification Trees**

- Second Measure: Training Time
  - 4x faster than Naive Bayes
  - 35x faster than Classification Trees
- Reasons for extraordinary performance
  - Algorithms take advantage of 'sparsity in the document by term matrix'
- Other Methods Not Useful
  - Methods that involve the 'diagonalization of large and dense matrices'
  - E.g. like the 'criterion matrix in Fisher flexible discriminant analysis'
  - Out of consideration for text classication, because of their expensive computational requirements

[1] SVMs with Applications

Bioinformatics Tasks

- Goal: Analyzing microarray data
  - Analyzing ‘biological samples’ using their genetic expression profiles

- SVMs have been applied to:
  - E.g. tissue classication
  - E.g. gene function prediction
  - E.g. protein subcellular location prediction
  - E.g. protein secondary structure prediction
  - E.g. protein fold prediction

- **SVM outperform other classification methods (or in worst case are at least similiar to the best non-SVM method)**

[1] SVMs with Applications

Bioinformatics: Protein Subcellular Location Prediction

- Goal
  - Predict protein subcellular positions from 'prokaryotic sequences'
- Categories
  - Three possible location categories
  - [cytoplasmic, periplasmic, extracellular]
- Approach
  - Classification: the problem reduces to classifying 20-dimensional vectors into three (highly unbalanced) classes
- Measure: Prediction Accuracy
  - SVMs (with Gaussian Kernel) is 91.4%
  - Neural Networks is 81%
  - First-order Markov Chain is 89.1%

Image Processing (1)

- Example: Handwritten Digit Identification
  - U.S. Postal Service data base contains 9298 samples of digits
  - Divided into 7291 training samples and 2007 samples for testing
  - Obtained from real-life zip codes
  - Digit is represented by a 16 × 16 gray level matrix
  - Each data point is represented by a Vector in R256
  - Manual: Human classication error for this problem is known to be 2.5%
  - Automated: Error rate for a standard SVM is 4%
    (with a third degree polynomial kernel)

[1] SVMs with Applications

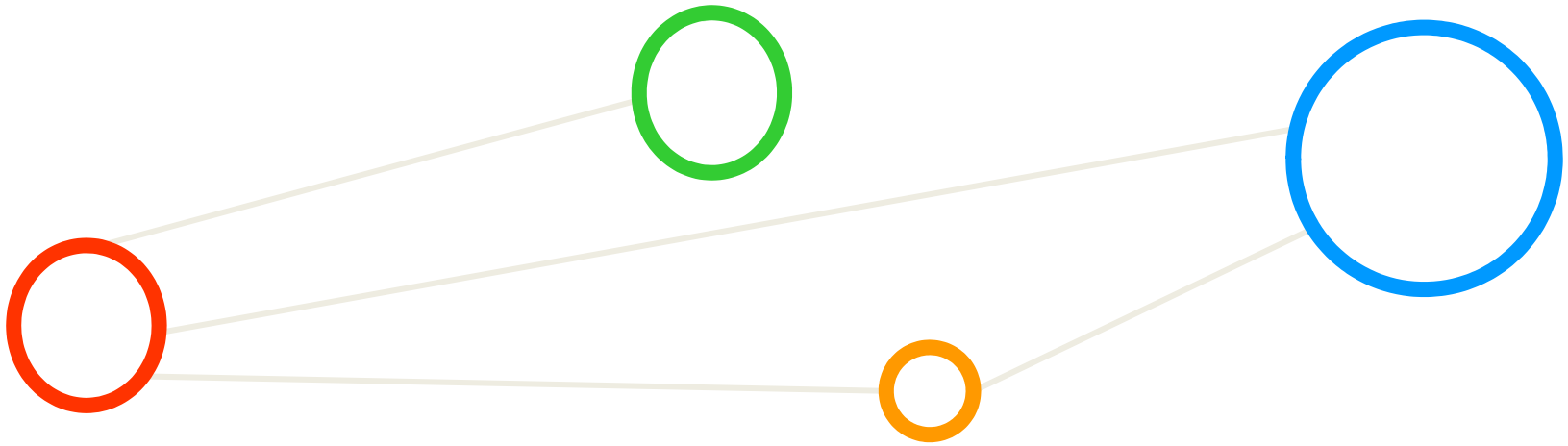
Image Processing (2)

- Example: Face Recognition
  - E.g. Used for ‘gender detection’
  - The data contain 1755 face images (1044 males and 711 females),
  - Overall error rate for a SVM with a Gaussian kernel is 3.2%

  - E.g. another application of SVMs is the ‘detection of human faces’ in gray-level images
  - Determine in an image the location of human faces
  - If there are any, return an encoding of their position
  - Detection rate for a SVM with a second degree polynomial kernel is 97.1%

[1] SVMs with Applications

Lecture Bibliography



Lecture Bibliography

- [1] Javier M. Moguerza and Alberto Muñoz, ‘Support Vector Machines with Applications’, Statistical Science, Vol.21, No.3, pp. 322.336, DOI: 10.1214/088342306000000493, Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.951&rep=rep1&type=pdf
- [2] N. Cristianini and J. Shawe-Taylor, ’An Introduction to Support Vector Machines’ Cambridge Univ. Press
- [3] D. Ventura, ‘SVM Example‘, Online: http://axon.cs.byu.edu/Dan/678/miscellaneous/SVM.example.pdf
- [4] Chih-Wei Hsu, Chih-Chung Chang, und Chih-Jen Lin, 'A Practical Guide to Support Vector Classification’, Department of Computer Science, National Taiwan University, 2010
- [5] AI FAQ/neural Nets Index, 2014 Advameg, Inc, Online: http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html

services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
Cross-Disciplinary
PB
community
Environment
Modelling
different
EU
methods
High
big
nodes
centers
Key
scientific
important
Science
etc
DKRZ
Use
storage
basis
Infrastructure
files
model
international
SMAQ
engineering
References
formats
simulations
ENES
technologies
brain
increasing
via
EUDAT
physical
IPCC
Big
Project
Data
MapReduce
create
processing
Common
TB
Methods
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
use
DLCL
Level
Research
Matters
Concrete
external
performance
JSC
computational
datasets
research
e.g
science
HPC
Juelich
system
Landscape
acquisition
large-scale
compute
management
System
Climate
modelling
AR5
Hardware
distributed
cores
access
hundreds
Nutshell
Services
Technologies
Earth
Structure
current
network
DLCLs
Understanding
structures
directory
project
General
Germany
Simulation
Processing
Provide
NASA
Energy
systems
day
Health