

Introduction to Kernel Techniques for Pattern Analysis

[Part of ELEC3002/7005: Computational Methods in Electrical Engineering]

Conrad Sanderson

NICTA
Queensland Laboratory

conrad.sanderson@nicta.com.au

March 2009

Intro to Kernels

- Outline
- Books + Acks

Pattern Analysis

- Classification (Recognition)
- Clustering
- Outlier Detection
- Distance Function
- Problem!

Kernels

- Dist. Function Recasted (1)
- Dot Product
- Dot Product on Steroids
- Mapping Examples
- Mercer's Theorem
- Dist. Function Recasted (2)
- Example Kernels
- Constructing New Kernels

Kernel Machines

Outlier Detection

- Required Parameters
- The Hidden Centroid
- Distance to Centroid
- Standard Deviation
- The Machine

Classification

- Centroid Classifier
- Linear Decision Boundary
- Support Vector Machine

Outline

■ Pattern Analysis

- Classification
- Clustering
- Outlier Detection

■ Kernels

- Recasting the Distance Function
- Dot Product
- Mercer's Theorem
- Example Kernels

■ Kernel Machines

- Types
- Outlier Detection
- Classification (Recognition)
- Support Vector Machines

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Books + Acknowledgements

- S. Theodoridis & K. Koutroumbas, **Pattern Recognition** (3rd edition), Academic Press, 2006.
- Christopher Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006.
- John Shawe-Taylor & Nello Cristianini, **Kernel Methods for Pattern Analysis**, Cambridge, 2004.

My thanks go to:

- Erik Berglund
- Charles Gretton

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Pattern Analysis

- Classification (Recognition)
- Clustering
- Outlier Detection
- (Regression)

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

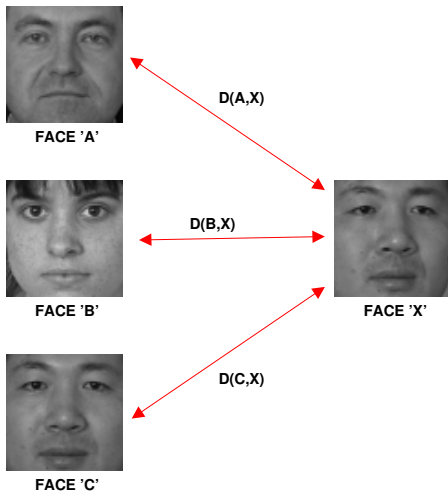
Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Classification (Recognition)

- e.g. recognise faces
- requires a distance function to compare faces



- more sophisticated methods later

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

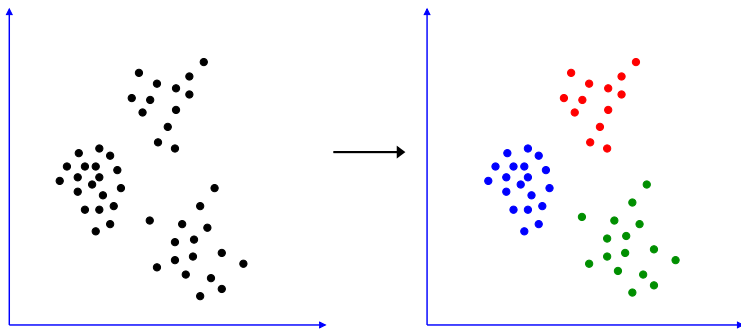
Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Clustering

- find clusters in data (i.e. lump similar points together)
- requires a distance function to compare points
- can be used for compression of images & sounds



Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

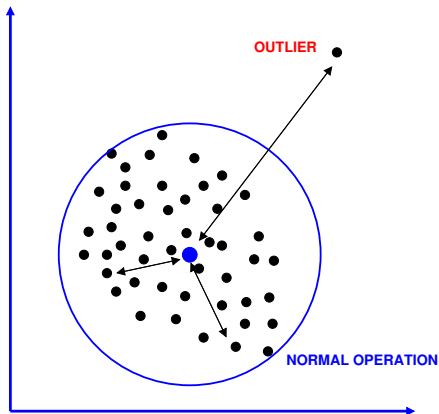
Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Outlier Detection

- find points that don't fit a model
- requires a distance function to compare points
- used to detect anomalies
- e.g. detect malfunctions of an engine subsystem



Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Distance Function

- Assume our points are vectors in N -dimensional Euclidean space:

$$\mathbf{a} = [a_1 \ a_2 \ a_3 \ \cdots \ a_N]^T$$
$$\mathbf{b} = [b_1 \ b_2 \ b_3 \ \cdots \ b_N]^T$$

- Distance between two vectors:

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^N (a_i - b_i)^2}$$
$$= \|\mathbf{a} - \mathbf{b}\|$$

where $\|\mathbf{x}\|$ is the norm of vector \mathbf{x} :

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2}$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Problem!

- Assumption that points are N -dimensional vectors can be very restrictive
- What if the points/objects are:
 - sets of vectors... with varying size (cardinality) ?
 - graphs (used in biology) ?
 - texts ?
 - comprised of two or more different descriptions ? (e.g. identity of person = face + speech)
- Traditional method of attack:
 - Design a highly specific classification/clustering/etc algorithm that is tightly coupled to the type of data
 - Use a black box method and hope it works
 - Devise an ad-hoc technique (hack) and hope it works

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Distance Function Recasted (1)

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^N (a_i - b_i)^2}$$

$$\begin{aligned}\{d(\mathbf{a}, \mathbf{b})\}^2 &= \sum_{i=1}^N (a_i - b_i)^2 \\ &= (a_1 - b_1)^2 + \dots + (a_N - b_N)^2 \\ &= (a_1 a_1 - 2a_1 b_1 + b_1 b_1) + \dots + (a_N a_N - 2a_N b_N + b_N b_N) \\ &= a_1 a_1 + a_2 a_2 + \dots + a_N a_N \\ &\quad - 2a_1 b_1 - 2a_2 b_2 - \dots - 2a_N b_N \\ &\quad + b_1 b_1 + b_2 b_2 + \dots + b_N b_N \\ &= \langle \mathbf{a}, \mathbf{a} \rangle - 2 \langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{b}, \mathbf{b} \rangle\end{aligned}$$

■ $\langle \mathbf{a}, \mathbf{b} \rangle$ is the dot product of \mathbf{a} and \mathbf{b} .

■ ... so what ?

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Dot Product (1)

■ also known as *inner product* and *scalar product*

■ $\langle \mathbf{a}, \mathbf{a} \rangle = \|\mathbf{a}\|^2$

■ $\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle$ (commutative, i.e. symmetric)

■ $\langle \mathbf{a}, (\mathbf{b} + \mathbf{c}) \rangle = \langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{a}, \mathbf{c} \rangle$ (distributive)

■ $\langle (c\mathbf{a}), \mathbf{b} \rangle = c \langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, (c\mathbf{b}) \rangle$ (mul by a constant)

■ $\langle \mathbf{0}, \mathbf{a} \rangle = 0$

■ $\cos(\theta) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|}$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

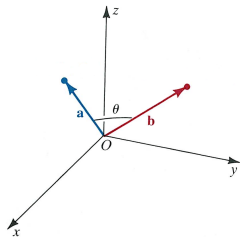
Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Dot Product (2)

- Normalised dot product:



- $$\frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

- \mathbf{a} and \mathbf{b} are 180° apart,

$$\frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} = -1$$

- \mathbf{a} and \mathbf{b} are orthogonal,

$$\frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} = 0$$

- \mathbf{a} and \mathbf{b} are the same,

$$\frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} = 1$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Dot Product on Steroids

- Let's say we have a function $\phi(\mathbf{x})$ which maps \mathbf{x} to a different space, e.g. from \mathcal{R}^2 to \mathcal{R}^3 :

$$\mathbf{x} \in \mathcal{R}^2 \rightarrow \mathbf{y} \in \mathcal{R}^3$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

- Let's do a dot product in the new space:

$$\langle \phi(\mathbf{a}), \phi(\mathbf{b}) \rangle = \underbrace{k(\mathbf{a}, \mathbf{b})}_{\text{kernel}}$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

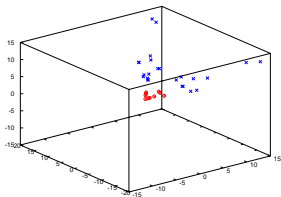
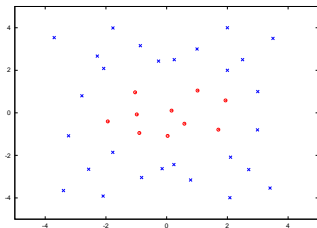
Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Mapping Example (1)

- Specific example:

$$\mathbf{x} \in \mathcal{R}^2 \rightarrow \mathbf{y} \in \mathcal{R}^3$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 & x_1 \\ \sqrt{2} x_1 & x_2 \\ x_2 & x_2 \end{bmatrix}$$



Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Mapping Example (2)

- So ...

$$\begin{aligned}k(\mathbf{a}, \mathbf{b}) &= \left\langle \left[a_1^2 \quad \sqrt{2}a_1a_2 \quad a_2^2 \right]^T, \left[b_1^2 \quad \sqrt{2}b_1b_2 \quad b_2^2 \right]^T \right\rangle \\ &= a_1^2b_1^2 + 2a_1a_2b_1b_2 + a_2^2b_2^2 \\ &= (a_1b_1 + a_2b_2)^2 \\ &= \{ \langle \mathbf{a}, \mathbf{b} \rangle \}^2\end{aligned}$$

- We've just computed the dot product in \mathcal{R}^3 without mapping to \mathcal{R}^3
- Turns out we can compute dot products in infinite dimensional spaces

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Mercer's Theorem (1)

■ $\mathbf{x} \in \mathcal{R}^N \rightarrow \phi(\mathbf{x}) \in \mathcal{H}$ (\mathcal{H} is a Euclidean space, for now...)

■ The dot product has an equivalent representation:

$$k(\mathbf{a}, \mathbf{b}) = \sum_i \phi_i(\mathbf{a})\phi_i(\mathbf{b})$$

- $\phi_i(\mathbf{x})$ is the i -th component of $\phi(\mathbf{x})$
- $k(\mathbf{a}, \mathbf{b})$ is a symmetric function satisfying:

$$\int k(\mathbf{a}, \mathbf{b}) g(\mathbf{a}) g(\mathbf{b}) d\mathbf{a} d\mathbf{b} \geq 0 \quad (1)$$

for any $g(\mathbf{x})$ that satisfies

$$\int g(\mathbf{x})^2 d\mathbf{x} < +\infty \quad (2)$$

- Opposite is also true: for any $k(\mathbf{a}, \mathbf{b})$ which satisfies (1) and (2) there exists a space where $k(\mathbf{a}, \mathbf{b})$ defines a dot product

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Mercer's Theorem (2)

- In general, \mathcal{H} is a Hilbert space – a complete linear space equipped with a dot product operation
- $\phi(\mathbf{x})$ is a mapping into a Hilbert space
- $k(\mathbf{a}, \mathbf{b}) = \langle \phi(\mathbf{a}), \phi(\mathbf{b}) \rangle$ (dot product in this Hilbert space)
- **x doesn't have to be a vector**
– as long as $\phi(\mathbf{x})$ exists we're fine!
- Don't need to explicitly know $\phi(\mathbf{x})$
– as long as $k(\mathbf{a}, \mathbf{b})$ exists we're fine!
- The theorem doesn't say how to find a Hilbert space
– it doesn't tell us how to automatically find kernel functions
- Design of kernels is a manual process
– you have to prove that your kernel satisfies Mercer's conditions

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Mercer's Theorem (3)

- Need a simple way to test whether a kernel is valid

- Construct a kernel matrix: (aka Gram matrix)

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & k(\mathbf{x}_N, \mathbf{x}_2) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

- $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_1, \dots, \mathbf{x}_N)$ contains all possible points

- Can approximate \mathbf{X} with training data (careful!)

- Kernel is valid when \mathbf{K} is positive semi-definite

- Positive semi-definite matrix: all eigenvalues are ≥ 0

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Distance Function Recasted (2)

- Replace dot products in original space with dot products in a Hilbert space: (kernel trick)

$$\{d(\mathbf{a}, \mathbf{b})\}^2 = \langle \mathbf{a}, \mathbf{a} \rangle - 2\langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{b}, \mathbf{b} \rangle$$

$$\begin{aligned}\{d_{\mathcal{H}}(\mathbf{a}, \mathbf{b})\}^2 &= \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle - 2\langle \phi(\mathbf{a}), \phi(\mathbf{b}) \rangle + \langle \phi(\mathbf{b}), \phi(\mathbf{b}) \rangle \\ &= k(\mathbf{a}, \mathbf{a}) - 2k(\mathbf{a}, \mathbf{b}) + k(\mathbf{b}, \mathbf{b})\end{aligned}$$

- “Backwards compatible” for $\phi(\mathbf{x}) = \mathbf{x}$
- Overall classification/clustering/etc algorithm can **stay the same** – only the kernel function changes
- Other algorithms which rely on dot products can also use kernels – e.g. Support Vector Machines (SVMs)

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Example Kernels (Vectorial)

■ Polynomials

$$k(\mathbf{a}, \mathbf{b}) = (\langle \mathbf{a}, \mathbf{b} \rangle + 1)^q, \quad q > 0$$

■ Radial Basis Functions (infinite dimensional space)

$$k(\mathbf{a}, \mathbf{b}) = \exp\left(-\frac{\|\mathbf{a} - \mathbf{b}\|^2}{\sigma^2}\right)$$

■ Hyperbolic Tangent (used in Artificial Neural Networks)

$$k(\mathbf{a}, \mathbf{b}) = \tanh(\beta \langle \mathbf{a}, \mathbf{b} \rangle + \gamma)$$

- for values of β and γ so that Mercer's conditions are satisfied (e.g. $\beta = 2$ and $\gamma = 1$)

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Constructing New Kernels

- Given valid kernels $k_1(\mathbf{a}, \mathbf{b})$ and $k_2(\mathbf{a}, \mathbf{b})$, these kernels will also be valid:

$$k(\mathbf{a}, \mathbf{b}) = c \cdot k_1(\mathbf{a}, \mathbf{b}) \quad c > 0$$

$$k(\mathbf{a}, \mathbf{b}) = k_1(\mathbf{a}, \mathbf{b}) + k_2(\mathbf{a}, \mathbf{b})$$

$$k(\mathbf{a}, \mathbf{b}) = k_1(\mathbf{a}, \mathbf{b}) \cdot k_2(\mathbf{a}, \mathbf{b})$$

$$k(\mathbf{a}, \mathbf{b}) = f(\mathbf{a}) \cdot k_1(\mathbf{a}, \mathbf{b}) \cdot f(\mathbf{b}) \quad f(\cdot) = \text{any function}$$

$$k(\mathbf{a}, \mathbf{b}) = q\{ k_1(\mathbf{a}, \mathbf{b}) \} \quad q(\cdot) = \text{polynomial w/ +ve coefficients}$$

$$k(\mathbf{a}, \mathbf{b}) = \exp\{ k_1(\mathbf{a}, \mathbf{b}) \}$$

$$k(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{A} \mathbf{b} \quad \left\{ \begin{array}{l} \mathbf{a} \text{ and } \mathbf{b} \text{ are vectors} \\ \mathbf{A} = \text{symmetric positive semi-definite matrix} \end{array} \right.$$

$$k(\mathbf{a}, \mathbf{b}) = k_3 (\psi(\mathbf{a}), \psi(\mathbf{b})) \quad \left\{ \begin{array}{l} \psi(\mathbf{x}) \text{ maps } \mathbf{x} \text{ to } \mathcal{R}^M \\ k_3(\cdot, \cdot) \text{ is a valid kernel in } \mathcal{R}^M \end{array} \right.$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Kernel Machines

- Traditional algorithms where the distance function has been modified to use kernels
 - nearest neighbour classification
 - clustering
 - ...
- Algorithms relying on dot products
 - Kernel PCA (Principal Component Analysis)
 - SVMs (Support Vector Machines)
 - ...
- Dense vs Sparse Kernel Machines
 - Outlier Detection (dense)
 - Centroid Classifier (dense)
 - Support Vector Machine (sparse)

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

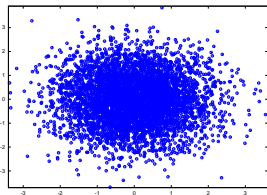
Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

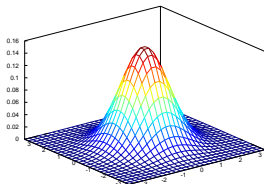
Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Outlier Detection

- Task: machine that automatically detects outliers
 - i.e. points that don't fit our model of data
- Simplifying assumption: model the data as an isotropic Gaussian distribution... in Hilbert space
 - variance is the same in all dimensions



- Distribution of 2D data (training data)



- Model of distribution (2D Gaussian)
- The closer to the center, the higher the probability

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

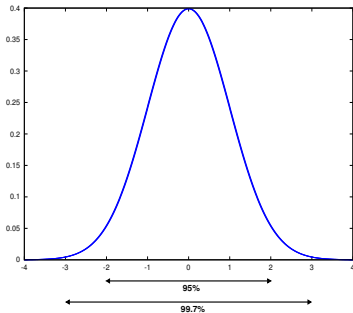
Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Required Parameters



- Need approximations of:
 - centroid (mean)
 - standard deviation
- 95% of data is within 2 SDs of the centroid
- 99.7% of data is within 3 SDs of the centroid

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

The Hidden Centroid

- All we have is the user-supplied kernel $k(\mathbf{a}, \mathbf{b})$
 - we don't know what $\phi(\mathbf{x})$ is

- Training data: $\mathbf{X} = \{ \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \}$

- Centroid in Hilbert space:

$$\phi_C = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i)$$

- We don't know where ϕ_C is ...

- ... but we can calculate the distance to ϕ_C

$$\begin{aligned} \{d_H(\mathbf{a}, \mathbf{b})\}^2 &= \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle - 2 \langle \phi(\mathbf{a}), \phi(\mathbf{b}) \rangle + \langle \phi(\mathbf{b}), \phi(\mathbf{b}) \rangle \\ &= \|\phi(\mathbf{a}) - \phi(\mathbf{b})\|^2 \end{aligned}$$

$$\therefore \|\phi(\mathbf{a}) - \phi_C\|^2 = \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle - 2 \langle \phi(\mathbf{a}), \phi_C \rangle + \langle \phi_C, \phi_C \rangle$$

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Distance to Centroid

$$\begin{aligned}\|\phi(\mathbf{a}) - \phi_C\|^2 &= \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle - 2 \langle \phi(\mathbf{a}), \phi_C \rangle + \langle \phi_C, \phi_C \rangle \\ &= \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle \\ &\quad - 2 \left\langle \phi(\mathbf{a}), \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i) \right\rangle \\ &\quad + \left\langle \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i), \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x}_i) \right\rangle \\ &= \langle \phi(\mathbf{a}), \phi(\mathbf{a}) \rangle \\ &\quad - 2 \frac{1}{N} \sum_{i=1}^N \langle \phi(\mathbf{a}), \phi(\mathbf{x}_i) \rangle \\ &\quad + \frac{1}{N} \sum_{i=1}^N \frac{1}{N} \sum_{j=1}^N \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\ &= k(\mathbf{a}, \mathbf{a}) \\ &\quad - \frac{2}{N} \sum_{i=1}^N k(\mathbf{a}, \mathbf{x}_i) \\ &\quad + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Standard Deviation

- Traditional approximation of standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

- Our analog:

$$\begin{aligned}\sigma_{\mathcal{H}} &= \sqrt{\frac{1}{N} \sum_{p=1}^N \|\phi(\mathbf{x}_p) - \phi_C\|^2} \\ &= \sqrt{\frac{1}{N} \sum_{p=1}^N \left(k(\mathbf{x}_p, \mathbf{x}_p) - \frac{2}{N} \sum_{i=1}^N k(\mathbf{x}_p, \mathbf{x}_i) + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j) \right)} \\ &= \sqrt{\frac{1}{N} \sum_{p=1}^N k(\mathbf{x}_p, \mathbf{x}_p) - \frac{2}{N^2} \sum_{p=1}^N \sum_{i=1}^N k(\mathbf{x}_p, \mathbf{x}_i) + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j)} \\ &= \sqrt{\underbrace{\frac{1}{N} \sum_{p=1}^N k(\mathbf{x}_p, \mathbf{x}_p)}_{\text{average of diagonal entries in K}} - \underbrace{\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(\mathbf{x}_i, \mathbf{x}_j)}_{\text{average value of K}}}\end{aligned}$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

The Machine

- Given training data:
 - calculate kernel matrix \mathbf{K} (needed to approximate ϕ_C)
 - calculate $\sigma_{\mathcal{H}}$
- Checking whether a new point, \mathbf{z} , is an outlier:

$$\| \phi(\mathbf{z}) - \phi_C \| \overset{\text{further than 3 SDs}}{> 3 \cdot \sigma_{\mathcal{H}}} \rightarrow \text{outlier}$$

$$\| \phi(\mathbf{z}) - \phi_C \| \leq 3 \cdot \sigma_{\mathcal{H}} \rightarrow \text{inlier}$$

- Upsides:
 - Works with any data for which we have a kernel
- Downsides:
 - Need to keep all training data
 - Can be slow if training set is large and/or the kernel function is slow
 - Relying on the implicit mapping to make the data have a Gaussian distribution

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

Classification

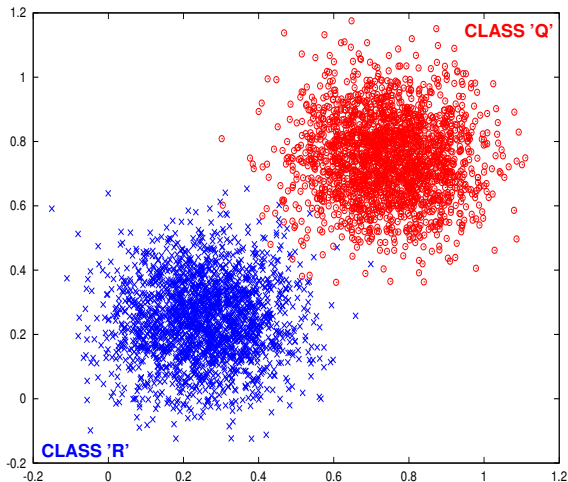
Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Classification

- Task is to make a machine that automatically classifies a given point as belonging to one of two classes



Intro to Kernels

- Outline
- Books + Acks

Pattern Analysis

- Classification (Recognition)
- Clustering
- Outlier Detection
- Distance Function
- Problem!

Kernels

- Dist. Function Recasted (1)
- Dot Product
- Dot Product on Steroids
- Mapping Examples
- Mercer's Theorem
- Dist. Function Recasted (2)
- Example Kernels
- Constructing New Kernels

Kernel Machines

Outlier Detection

- Required Parameters
- The Hidden Centroid
- Distance to Centroid
- Standard Deviation
- The Machine

Classification

- Centroid Classifier
- Linear Decision Boundary
- Support Vector Machine

Centroid Classifier

- Simplifying assumption: we can represent classes **Q** and **R** as the centers (centroids) of their respective training points
- Classification rule:

$$h(\mathbf{a}) = \begin{cases} \text{class } \mathbf{Q} & \text{if } \|\phi(\mathbf{a}) - \phi_C^{\mathbf{Q}}\| < \|\phi(\mathbf{a}) - \phi_C^{\mathbf{R}}\| \\ \text{class } \mathbf{R} & \text{otherwise} \end{cases}$$

- Upsides:
 - Works with any data for which we have a kernel
- Downsides:
 - Need to keep all training data
 - Can be slow if training set is large and/or the kernel function is slow
 - Relying on the implicit mapping to make the data easy to separate

Intro to Kernels

Outline

Books + Acks

Pattern Analysis

Classification (Recognition)

Clustering

Outlier Detection

Distance Function

Problem!

Kernels

Dist. Function Recasted (1)

Dot Product

Dot Product on Steroids

Mapping Examples

Mercer's Theorem

Dist. Function Recasted (2)

Example Kernels

Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters

The Hidden Centroid

Distance to Centroid

Standard Deviation

The Machine

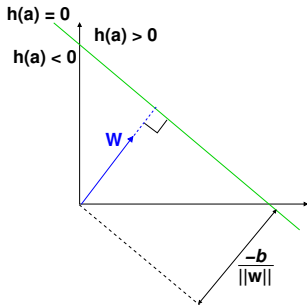
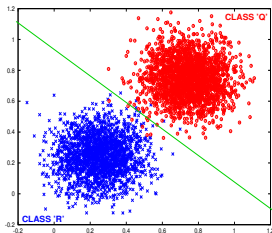
Classification

Centroid Classifier

Linear Decision Boundary

Support Vector Machine

Linear Decision Boundary



- Rather than finding centroids, find a linear separator function

- $h(\mathbf{a}) = \langle \mathbf{w}, \mathbf{a} \rangle + b$

- if $h(\mathbf{a}) > 0$ assign \mathbf{a} to **Q**

- if $h(\mathbf{a}) < 0$ assign \mathbf{a} to **R**

- Many ways to find \mathbf{w} and b

- This form works only with vectors

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Support Vector Machine (1)

- Linear case:

$$h(\mathbf{a}) = \langle \mathbf{w}, \mathbf{a} \rangle + b$$

- Using a mapping to Hilbert space:

$$\mathbf{w}_{\mathcal{H}} = \sum_{i=1}^N \beta_i \phi(\mathbf{x}_i)$$

- $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ are training points
- β_i 's are found by an optimiser

- In Hilbert space:

$$h(\mathbf{a}) = \langle \mathbf{w}_{\mathcal{H}}, \phi(\mathbf{a}) \rangle + b$$

$$= \left\langle \sum_{i=1}^N \beta_i \phi(\mathbf{x}_i), \phi(\mathbf{a}) \right\rangle + b$$

$$= \sum_{i=1}^N \beta_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{a}) \rangle + b$$

$$= \sum_{i=1}^N \beta_i k(\mathbf{x}_i, \mathbf{a}) + b$$

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

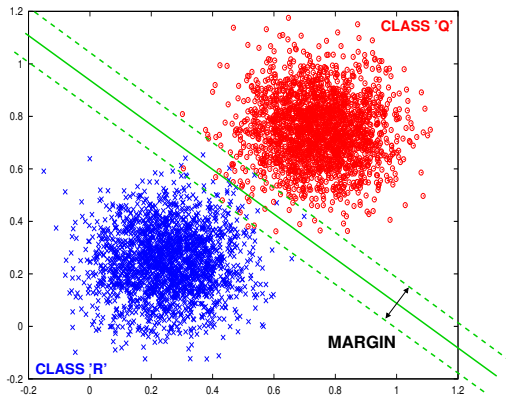
Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Support Vector Machine (2)



- $h(\mathbf{a}) = \sum_{i=1}^N \beta_i k(\mathbf{x}_i, \mathbf{a}) + b$

- β_i 's and b are selected so that the margin is maximised \rightarrow makes the classifier more robust

- Exact procedure is beyond the scope of this intro

Intro to Kernels

- Outline
- Books + Acks

Pattern Analysis

- Classification (Recognition)
- Clustering
- Outlier Detection
- Distance Function
- Problem!

Kernels

- Dist. Function Recasted (1)
- Dot Product
- Dot Product on Steroids
- Mapping Examples
- Mercer's Theorem
- Dist. Function Recasted (2)
- Example Kernels
- Constructing New Kernels

Kernel Machines

Outlier Detection

- Required Parameters
- The Hidden Centroid
- Distance to Centroid
- Standard Deviation
- The Machine

Classification

- Centroid Classifier
- Linear Decision Boundary
- Support Vector Machine

Support Vector Machine (3)

$$h(\mathbf{a}) = \sum_{i=1}^N \beta_i k(\mathbf{x}_i, \mathbf{a}) + b$$

- Many β_i 's are zero \therefore many \mathbf{x}_i are not used
- Training points with non-zero β_i 's: “support vectors”
- Typically the number of support vectors is much less than the number of training points (sparse)
- Upsides:
 - Works with any data for which we have a kernel
 - Usually much faster than the centroid classifier
 - Often better performance than an Artificial Neural Network

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

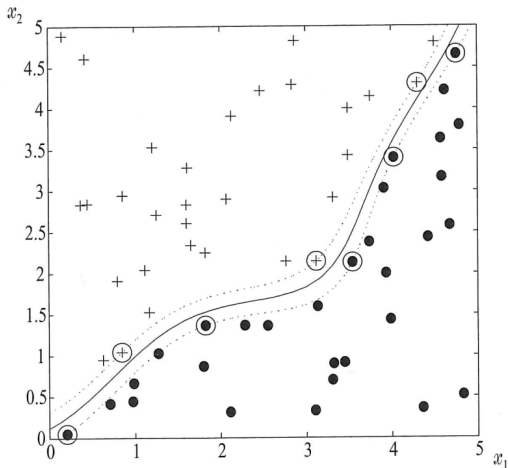
Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine

Support Vector Machine (4)



- SVM classifier using an RBF kernel
- Training data not linearly separable in original space
- Training data linearly separable in Hilbert space
- Circled points are the support vectors

Intro to Kernels

Outline
Books + Acks

Pattern Analysis

Classification (Recognition)
Clustering
Outlier Detection
Distance Function
Problem!

Kernels

Dist. Function Recasted (1)
Dot Product
Dot Product on Steroids
Mapping Examples
Mercer's Theorem
Dist. Function Recasted (2)
Example Kernels
Constructing New Kernels

Kernel Machines

Outlier Detection

Required Parameters
The Hidden Centroid
Distance to Centroid
Standard Deviation
The Machine

Classification

Centroid Classifier
Linear Decision Boundary
Support Vector Machine