Learning with Kernels

Advanced Signal Processing Seminar

Christina Leitner

Signal Processing and Speech Communication Laboratory
Graz University of Technology

11 November 2009

Outline

Intro – A Simple Learning Problem

Classification

- Sets $\mathcal{X}$ and $\mathcal{Y}$
- Each data point $\mathbf{x}_i$ is assigned a label y_i.
- Training data: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})$
- Goal: classify an unseen data point $\mathbf{x}$, i.e. predict its label y
- Find $(\mathbf{x}, y)$ that is *similar* to the training examples.
- How to measure similarity?
  Use the standard inner product, for a vector space $\mathbb{R}^n$:

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{n} a_i b_i$$

Intro – Example

Prototype classifier

- Separate patterns from class A and B
- Classification criterion: distance to means of class samples (blackboard example)

Intro – Example

Prototype classifier

- Separate patterns from class A and B
- Classification criterion: distance to means of class samples
- Euclidean distance between two points: norm of difference vector $\mathbf{d}$

$$\begin{aligned} \|\mathbf{d}\| &= \|\mathbf{x}-\mathbf{y}\| \\ &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2} \\ &= \sqrt{\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle} \end{aligned}$$

Intro – Example

Prototype classifier - Summary

- If $\langle \bar{a} - \bar{b}, x \rangle > \theta$ then $\mathbf{x} \in A$ else $\in B$

  with $\theta = \frac{1}{2}(\langle \bar{a}, \bar{a} \rangle - \langle \bar{b}, \bar{b} \rangle)$

- Decision boundary along $\langle \bar{a} - \bar{b}, x \rangle = \theta$, orthogonal to distance vector of means

Intro – Inner Product Space

Definition

A vector space $\mathcal{X}$ over the reals $\mathbb{R}$ is an *inner product space* if there exists a real-valued symmetric bilinear map $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$.

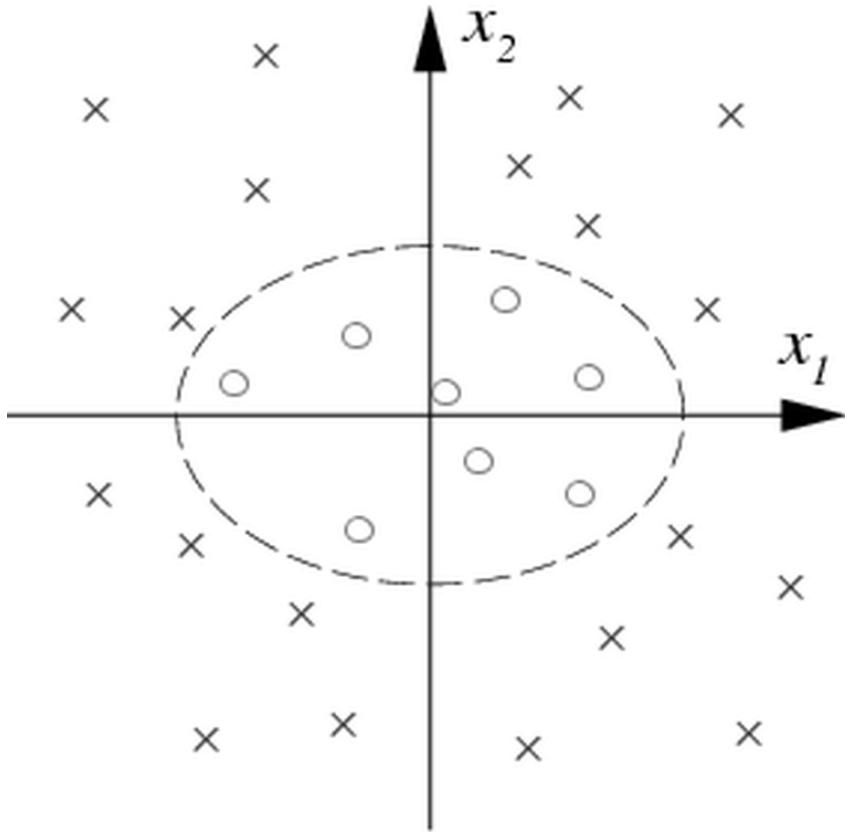
Angles and lengths

$$\cos\phi = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

Mapping – Non-linear Problems

- How to separate these data sets?

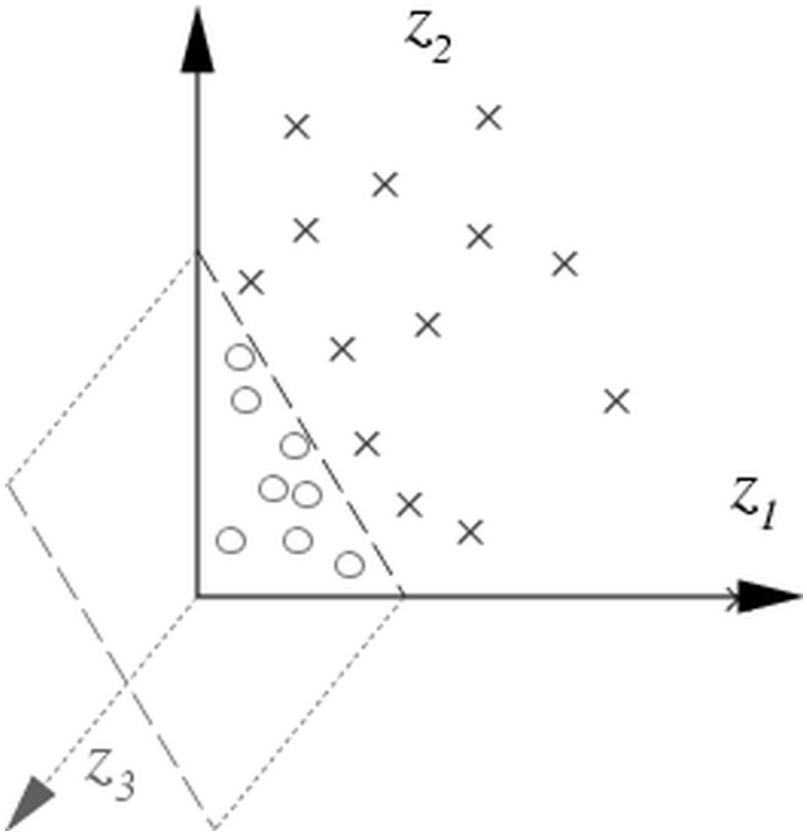


- Use a map to a space were the sets are separable!

Maps – Non-linear Problems

$$\Phi : \mathcal{X} \rightarrow \mathcal{Z}$$
$$\mathbf{x} \mapsto \mathbf{z}$$

Data can be separated by a *linear* hyperplane!



Maps – Non-linear Problems

- Example:

$$\begin{aligned} \Phi : \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (x_1, x_2) &\mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{aligned}$$

- For the similarity measure the feature space has to be an inner product space!

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = x_1^2 y_1^2 + 2x_1x_2y_1y_2 + y_1^2 y_2^2 = \langle \mathbf{x}, \mathbf{y} \rangle^2$$

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 := k(\mathbf{x}, \mathbf{y}) \quad \text{is a kernel function}$$

Kernels – Definition

Definition

A kernel is a function k that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ satisfies

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$$

where Φ is a mapping from $\mathcal{X}$ to an (inner product) feature space $\mathcal{Z}$

$$\Phi : \mathbf{x} \mapsto \Phi(\mathbf{x}) \in \mathcal{Z}$$

Kernels – The Kernel Trick

- Instead of $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$ compute $\langle \mathbf{x}, \mathbf{y} \rangle^2$.
- Compute inner product directly in the input space.
- No mapping to the feature space is necessary.
- Computationally more efficient!
- This is called the kernel trick.

Kernels – The Kernel Trick

Summary

- Any algorithm that only depends on inner products can benefit from the kernel trick.
- Kernels are a generalization of inner products.
- They can be seen as a nonlinear similarity measure.

Kernels – Examples for Kernel Functions

- Polynomial

$$k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^d$$

- Sigmoid

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\kappa \langle \mathbf{x}, \mathbf{y} \rangle + \Omega)$$

- Gaussian

$$k(\mathbf{x}, \mathbf{y}) = \exp(-(\mathbf{x} - \mathbf{y})^2 / (2\sigma^2))$$

Kernels – Positive Definite Kernels

- A Kernel can always be constructed by

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$$

- How to find out if $k(.,.)$ defines an inner product in feature space without actually computing Φ?

- The kernel has to be *positive definite*.
  - Symmetric: $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})$.
  - Its *kernel matrix* or *Gram matrix* K is *positive semi-definite*.

$$K_{ij} := k(x_i, x_j) \quad \text{and} \quad v^T K v \geq 0 \quad \text{for all vectors v}$$

RKHS – Feature space

Construction of the feature space

- Define a feature map.
- Turn $\Phi(\mathcal{X})$ into a linear space.
- Endow it with a dot product satisfying

$$\langle k(.,x_i), k(.,x_j)\rangle = k(x_i,x_j)$$

- Complete the space to get a *reproducing kernel Hilbert space* (RKHS).

RKHS – The Reproducing Kernel Hilbert Space

Definition

A *Hilbert Space* $\mathcal{H}$ is an inner product space with the additional properties that it is *separable* and *complete*.

- Completeness refers to the property that every Cauchy sequence of elements of $\mathcal{H}$ converges to an element h of $\mathcal{H}$
- A space $\mathcal{H}$ is separable if for any $\epsilon > 0$ there is a finite set of elements $h_1, ..., h_N$ of $\mathcal{H}$ such that for all h of $\mathcal{H}$

$$\min_i \|h_i - h\| < \epsilon$$

RKHS – Turn It into a Linear Space

Form all linear combinations

$$
\begin{aligned}
f(.) &= \sum_{i=1}^{m} \alpha_i k(., x_i), \\
g(.) &= \sum_{j=1}^{m'} \beta_i k(., x'_j),
\end{aligned}
$$

$(m, m' \in \mathbb{N}, \alpha_i, \beta_i \in \mathbb{R}, x_i, x'_j \in \mathcal{X})$.

RKHS – Endow It with an Inner Product

$$\begin{aligned}\langle f, g\rangle &:= \sum_{i=1}^{m}\sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \\ &= \sum_{i=1}^{m} \alpha_i g(x_i) = \sum_{j=1}^{m'} \beta_j f(x'_j)\end{aligned}$$

- This is well-defined, symmetric and bilinear.

RKHS – The Reproducing Kernel Property

Two special cases:

- Assume

$$f(.) = k(., x)$$

  In this case we have

$$\langle k(., x), g \rangle = g(x)$$

- If moreover

$$g(.) = k(., x')$$

  we have

$$\langle k(., x), k(., x') \rangle = k(x, x')$$

- k is called a *reproducing kernel*.

RKHS – The Reproducing Kernel Property

- Because of the reproducing property of the kernel, the feature space is called *reproducing kernel Hilbert space* (RKHS).

Discussion

Thank you for your attention!

References I


F. Jäkel, B. Schölkopf, F.A. Wichmann,
"A tutorial on kernel on kernel methods for categorization",
Journal of Mathematical Psychology, 2007.


B. Schölkopf,
Statistical learning and kernel methods,
Technical Report, Microsoft Research Cambridge, 2000.


J. Shawe-Taylor and N. Cristianini,
"Kernel Methods for Pattern Analysis",
Cambridge University Press, 2004.


B. Schölkopf and A. Smola,
Learning with kernels,
MIT Press, Cambridge, MA, 2002.


B. Schölkopf and A. Smola,
"Learning with kernels"
[Online; accessed 11-November-2009]. Available:
`http://www.learning-with-kernels.org/`

References II

B. Schölkopf and A. Smola,
"Introduction to kernel methods"
Videolecture and slides [Online; accessed 10-November-2009]. Available:
http://demo.viidea.com/mlss04_scholkopf_ikm/

C. M. Bishop,
Pattern Recognition and Machine Learning,
Springer, 2006.