UNIVERSITY OF LEEDS

# Class: Machine Learning

## Support Vector Machines

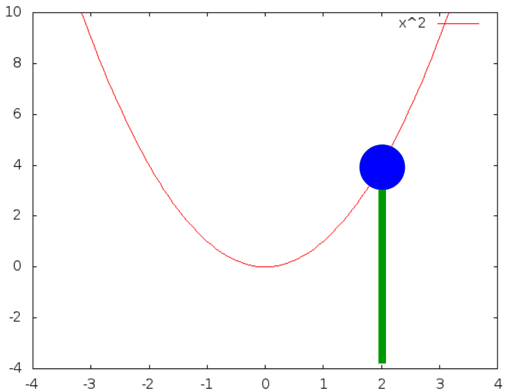
Instructor: Matteo Leonetti

# Learning outcomes

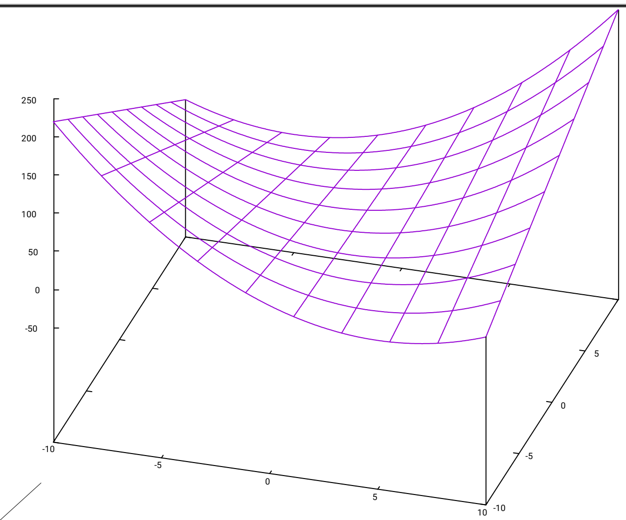
- Derive the dual formulation of Support Vector Machine
- Explain the kernel trick
- Apply dual SVMs and the kernel trick to datasets.

# The Dual Problem

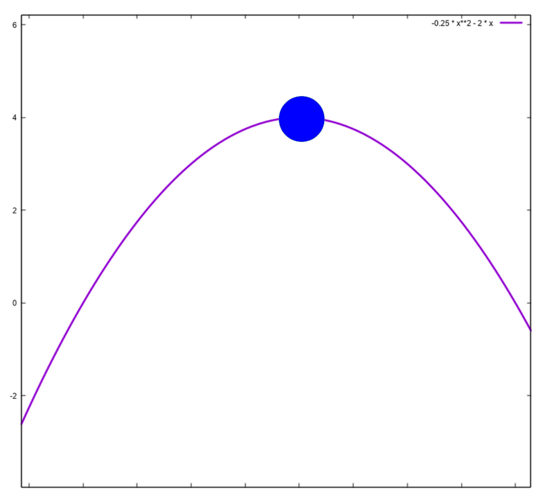
UNIVERSITY OF LEEDS




$$\min \; f(x) = x^2$$
$$\text{s.t.} \quad x = 2$$



$$L(x, \lambda) = x^2 + \lambda(x - 2)$$




$$\max \; q(\lambda) = -\frac{1}{4}\lambda^2 - 2\lambda$$

$$\boxed{f(2) = q(-4) = 4}$$

# KKT Conditions

UNIVERSITY OF LEEDS

min $f(\boldsymbol{x})$

min $f(\boldsymbol{x})$

Subject to

$$h_i(\boldsymbol{x})=0 \quad \forall\, i=1,\ldots,m$$

min $f(\boldsymbol{x})$

Subject to

$$h_i(\boldsymbol{x})\leq 0 \quad \forall\, i=1,\ldots,m$$

Corresponding system of equations

$$\nabla_x f(\boldsymbol{x})=0$$

$$\nabla_x L(\boldsymbol{x},\boldsymbol{\lambda})=0$$

$$\nabla_\lambda L(\boldsymbol{x},\boldsymbol{\lambda})=0$$

$$\nabla_x L(\boldsymbol{x},\boldsymbol{\lambda})=0$$

$$\lambda_i g_i(\boldsymbol{x})=0 \quad \forall\, i=1,\ldots,n$$

$$\lambda_i\geq 0 \quad \forall\, i=1,\ldots,n$$

# Duality and SVM

UNIVERSITY OF LEEDS

What is the dual formulation of this?

minimise: $$\frac{1}{2}\|\boldsymbol{w}\|^2$$

Subject to the constraints: $$t_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + w_0) \geq 1$$

# Follow the Duality Recipe

UNIVERSITY OF LEEDS

1. compile constraints into the Lagrangian

$$\min\ f(x) = x^2$$

s.t. $-x-3 \leq 0$

$$x+2 \leq 0$$

$$L(x, \boldsymbol{\lambda}) = x^2 + \lambda_1(-x-3) + \lambda_2(x+2)$$

min $\frac{1}{2}\|\boldsymbol{w}\|^2$

s.t.: $t_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} + w_0) \geq 1$

min $\frac{1}{2}\|\boldsymbol{w}\|^2$

s.t.: $1 - t_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_i} + w_0) \leq 0$

$$L(\boldsymbol{w}, w_{0,} \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \lambda_n(1 - t_n(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_n + w_0))$$

# Follow the Duality Recipe

UNIVERSITY OF LEEDS

2. solve for the optimal primal variables

$$L(\boldsymbol{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \lambda_n \left(1 - t_n\left(\boldsymbol{w}^T \boldsymbol{x_n} + w_0\right)\right)$$

$$\nabla_x L(x, \boldsymbol{\lambda}) = 2x - \lambda_1 + \lambda_2 = 0$$

$$\downarrow$$

$$x = \frac{\lambda_1 - \lambda_2}{2}$$

$$\downarrow$$

$$\nabla_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{x_n} = 0 \qquad \boldsymbol{w}^* = \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{x_n}$$

$$\frac{\partial L}{\partial w_0} = -\sum_{n=1}^{N} \lambda_n t_n = 0$$

# Follow the Duality Recipe

3. substitute the solution for x

$$L(x,\boldsymbol{\lambda}) = x^2 + \lambda_1(-x-3) + \lambda_2(x+2)$$

$$x = \frac{\lambda_1 - \lambda_2}{2}$$

$$q(\boldsymbol{\lambda}) = -\frac{1}{4}\lambda_1^2 - \frac{1}{4}\lambda_2^2 - 3\lambda_1 + 2\lambda_2 + \frac{1}{2}\lambda_1\lambda_2$$

# Follow the Duality Recipe

UNIVERSITY OF LEEDS

3. substitute the solution for w and $w_0$

$$L(\boldsymbol{w}, w_{0,} \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \lambda_n (1 - t_n(\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{n}} + w_0))$$

$$\boldsymbol{w}^* = \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{x}_n \qquad\qquad \sum_{n=1}^{N} \lambda_n t_n = 0$$

$$L(\boldsymbol{\lambda}) = \frac{1}{2}\|\sum_n \lambda_n t_n x_n\|^2 + \sum_n \lambda_n (1 - t_n((\sum_k \lambda_k t_k x_k) x_n + w_0))$$

$$= \frac{1}{2}\|\sum_n \lambda_n t_n x_n\|^2 + \sum_n \lambda_n - \sum_n \lambda_n t_n w_0 - \sum_n \lambda_n t_n (\sum_k \lambda_k t_k x_k) x_n$$

$$= \frac{1}{2}\|\sum_n \lambda_n t_n x_n\|^2 + \sum_n \lambda_n - \underbrace{\sum_n \lambda_n t_n w_0}_{=0} - (\sum_n \lambda_n t_n x_n)(\sum_k \lambda_k t_k x_k)$$

# Formulations

UNIVERSITY OF LEEDS

Min

$$\frac{1}{2}\|\boldsymbol{w}\|^2$$

Subject to

$$t_i(\boldsymbol{w}^{\boldsymbol{T}} x_i + w_0) \geq 1$$

Max

$$\sum_{n=1}^{N} \lambda_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m \boldsymbol{x}_n^T \boldsymbol{x}_{\boldsymbol{m}}$$

Subject to

$$\lambda_n \geq 0$$

$$\sum_{n=1}^{N} \lambda_n t_n = 0$$

$$\boldsymbol{w}^* = \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{x}_n$$

$$w_0 = \frac{1}{N_s} \sum_{j \in \text{support vectors}} \left(t_j - \boldsymbol{w}^t \boldsymbol{x}_j\right)$$

$$w_0 = \frac{1}{N_s} \sum_{j \in \text{support vectors}} \left(t_j - \sum_{i=1}^{N} \lambda_i t_i \boldsymbol{x}_i^T \boldsymbol{x}_j\right)$$

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{x}^T \boldsymbol{x}_{\boldsymbol{n}} + w_0$$

# Dual problem

UNIVERSITY OF LEEDS

Max $$L(\boldsymbol{\lambda})=\sum_{n=1}^{N}\lambda_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m \boldsymbol{x}_n^T\boldsymbol{x}_m$$

$$\lambda_n \geq 0$$

$$\sum_{n=1}^{N}\lambda_n t_n = 0$$

The input vectors only appear **multiplied**

To classify:

$$y(\boldsymbol{x})=\sum_{n=1}^{N}\lambda_n t_n \boldsymbol{x}^T\boldsymbol{x}_n + w_0$$

# Remember how we expanded the dimensions...

The original dataset has 1 variable:

$$\langle x_{1,} t_1 \rangle , \langle x_{2,} t_2 \rangle , \ldots , \langle x_N , t_N \rangle$$

But we want a higher dimensional space...

Let's use polynomial features: $\Phi_i(x) = x^i$

Our points become:

$$\langle 1, x_{1,} x_{1,}^2 x_{1,}^3 \ldots , x_{1,}^d t_1 \rangle , \langle 1, x_{2,} x_{2,}^2 x_{2,}^3 \ldots , x_{2,}^d t_2 \rangle , \ldots , \langle 1, x_N , x_N^2 , x_N^3 , \ldots , x_N^d , t_N \rangle$$

# Features

Substitute with features

$$L(\boldsymbol{\lambda}) = \sum_{n=1}^{N} \lambda_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m \boldsymbol{\Phi}(x_n)^T \boldsymbol{\Phi}(x_m)$$

$$y(\boldsymbol{x}) = \sum_{n=1}^{N} \lambda_n t_n \boldsymbol{\Phi}(x)^T \boldsymbol{\Phi}(x_n) + w_0$$

# Experience the power of kernels

Given the two 2 dimensional points:

$$x = \langle 1, -1 \rangle, \ y = \langle -1, 2 \rangle$$

Compute the order 2 features:

$$\Phi(x) = \langle 1, \sqrt{2}\, x_1, \sqrt{2}\, x_2, \sqrt{2}\, x_1 x_2, x_1^2, x_2^2 \rangle$$

Compute the dot product:

$$\Phi(x)^T \Phi(y) = ?$$

Evaluate:

$$(1 + \boldsymbol{x}^T \boldsymbol{y})^2 = ?$$

# Example: Polynomial features

$$\begin{bmatrix} 1 & x_1 & x_1^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_2 \\ x_2^2 \end{bmatrix} = 1 + x_1 x_2 + x_1^2 x_2^2$$

Note that: $(1 + x_1 x_2)^2 = 1 + 2 x_1 x_2 + x_1^2 x_2^2$

Which is close!
If it wasn't for that factor of 2...

But wait, the features can be whatever we want...

$$\begin{bmatrix} 1 & \sqrt{2} x_1 & x_1^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ \sqrt{2} x_2 \\ x_2^2 \end{bmatrix} = 1 + 2 x_1 x_2 + x_1^2 x_2^2 = (1 + x_1 x_2)^2$$

# Kernels

UNIVERSITY OF LEEDS

$$k(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{\Phi}(\boldsymbol{x})^T\boldsymbol{\Phi}(\boldsymbol{y})$$

$$k(\boldsymbol{x},\boldsymbol{x})=(1+\boldsymbol{x}^T\boldsymbol{y})^s$$

Polynomials

$$k(\boldsymbol{x},\boldsymbol{y})=\exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma}\right)=\exp\left(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2\right)$$

“Gaussian” (RBF)

$$k(\boldsymbol{x},\boldsymbol{y})=\tanh\left(\kappa\boldsymbol{x}^T\boldsymbol{y}-\delta\right)$$

Sigmoid

# Constructing kernels

UNIVERSITY OF LEEDS

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= ck_1(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b)
\end{aligned}
$$

# Kernels

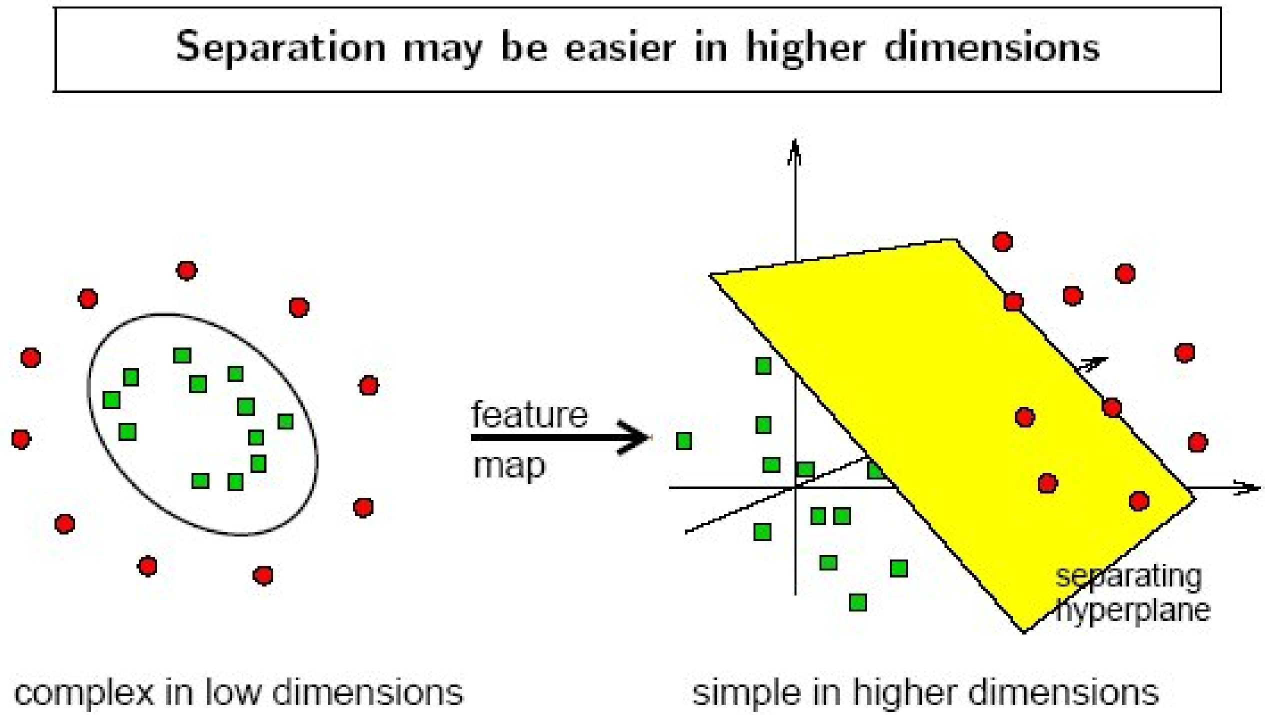
Substitute with kernels!

$$L(\boldsymbol{\lambda})=\sum_{n=1}^{N}\lambda_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\lambda_n\lambda_m t_n t_m k(x_n,x_m)$$

$$y(\boldsymbol{x})=\sum_{n=1}^{N}\lambda_n t_n k(x,x_n)+w_0$$

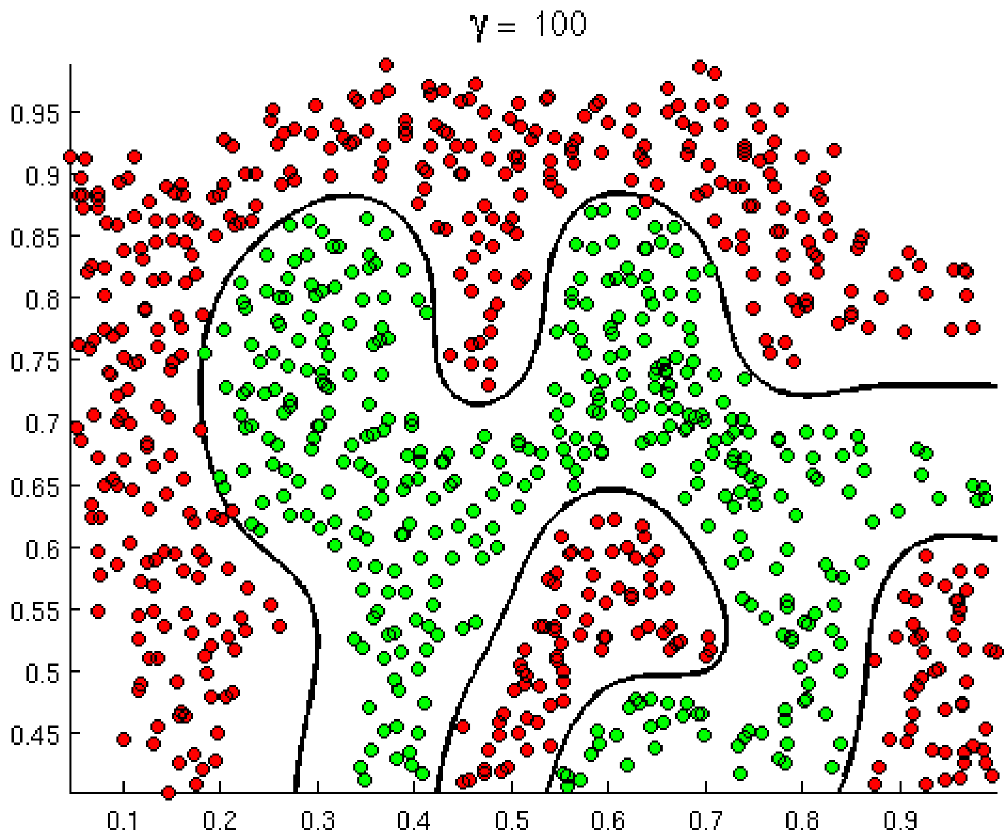
# Spaces




Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

# Example

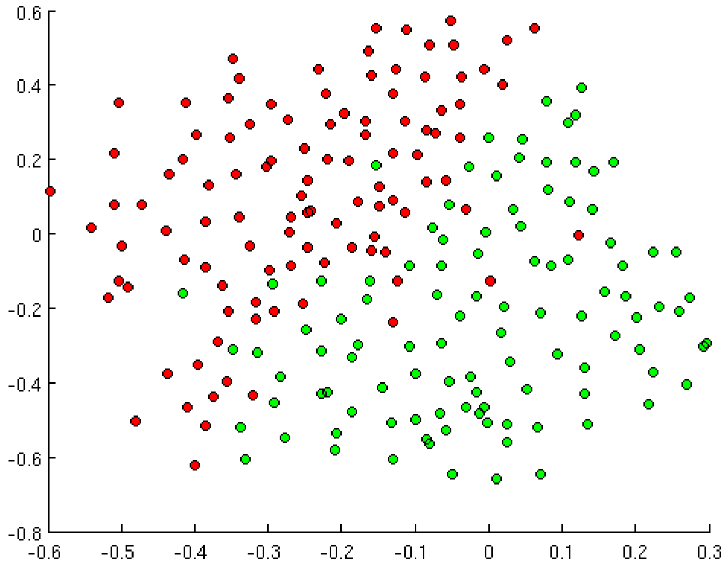


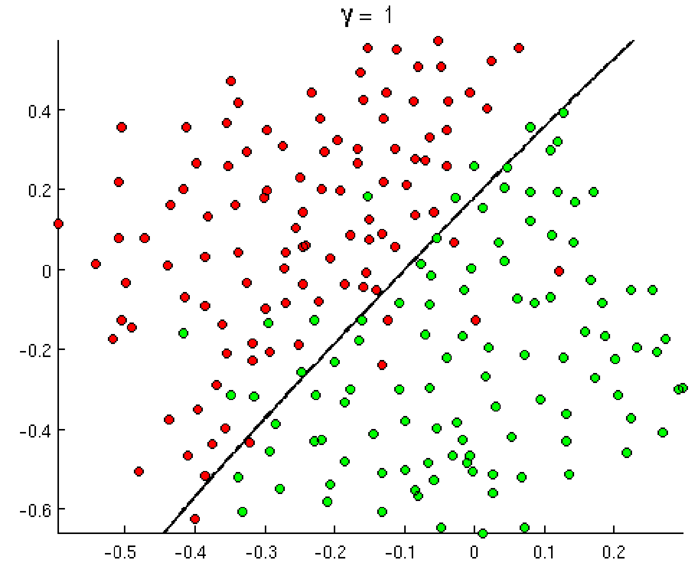



Gaussian kernel

[from Andrew Ng's ML class]

# Example 2

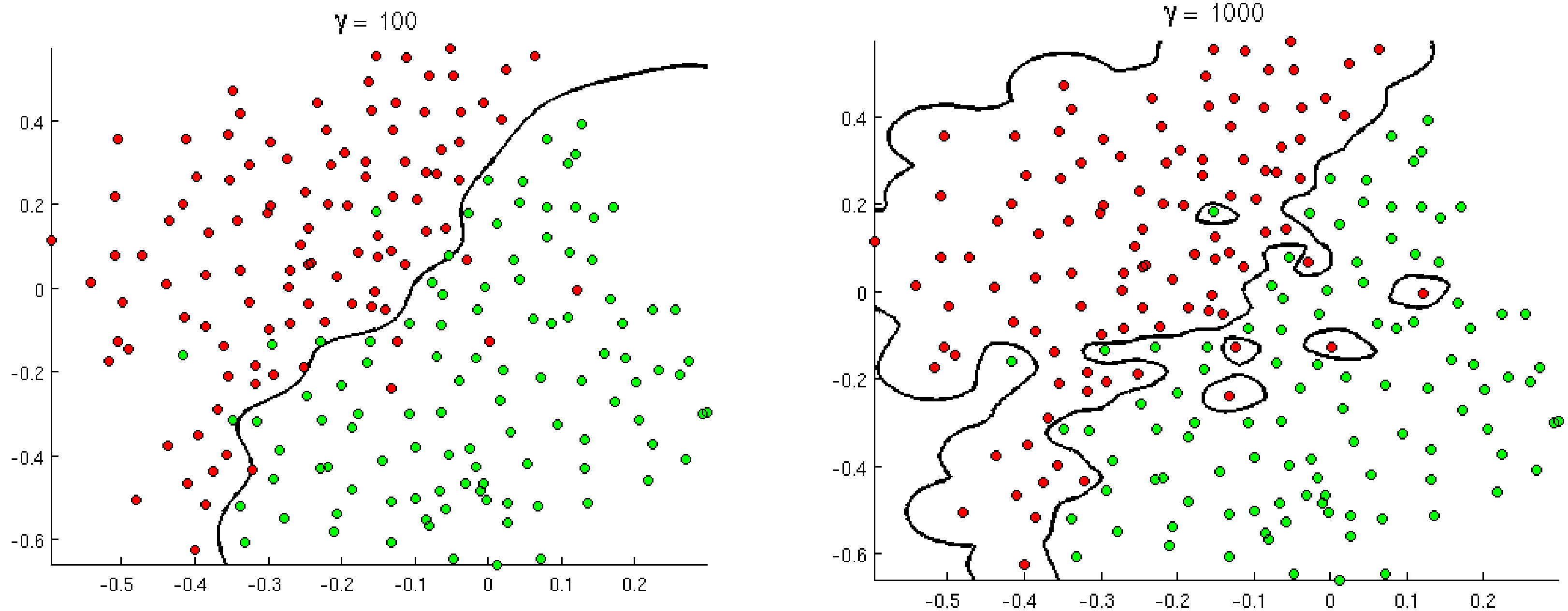



Gaussian kernel

[from Andrew Ng's ML class]

# Example 2




Gaussian kernel

[from Andrew Ng's ML class]

# History

UNIVERSITY OF LEEDS

- 1963 - Vladimir Vapnik, Alexey Chervonenkis

- 1992 - Isabelle Guyon

  *Proposed the dual formulation with the kernel trick*

- 1995 – Corinna Cortes (now head of Google Research)

  *Proposed the soft-margin SVM*

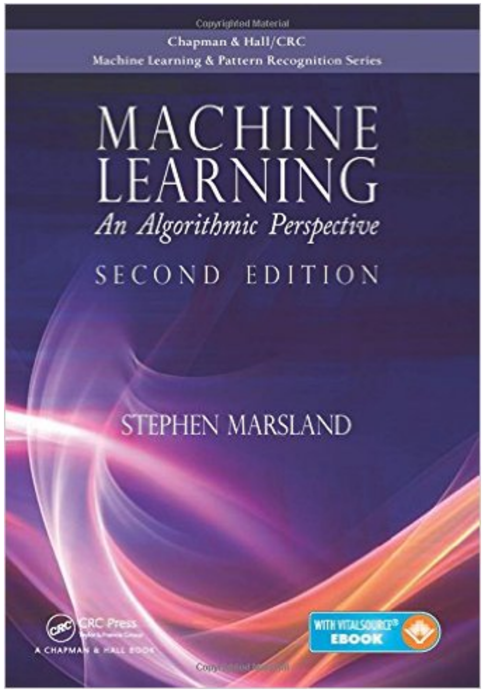
*(They all worked together in the 90s at Bell Labs)*

# Conclusion

# Learning outcomes

UNIVERSITY OF LEEDS

- Derive the dual formulation of Support Vector Machine
- Explain the kernel trick
- Apply dual SVMs and the kernel trick to datasets.


Copyrighted Material
Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series
MACHINE LEARNING
An Algorithmic Perspective
SECOND EDITION
STEPHEN MARSLAND
CRC Press
A CHAPMAN & HALL BOOK
WITH VITALSOURCE® EBOOK
Copyrighted Material

Chapter 8.2