SVM classifiers (the basics)

Johan Suykens

KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be
http://www.esat.kuleuven.be/stadius

Lecture 2

Contents

- Goals, applications
- Linear SVM classifier: separable and non-separable case
- Basics of optimization theory, QP problem, properties of solution
- Primal and dual form of SVM classifier model
- Nonlinear SVM classifier
- Mercer theorem, positive definite kernels, feature map and kernel trick
- Automatic determination of number of hidden units
- Geometrical interpretation of support vectors

Goals of SVMs

Support Vector Machines ...

- aim at solving general **nonlinear** classification and function estimation problems
- fundamentally recast the problem in terms of a **convex optimization problem**.
- can automatically find the **number of hidden units** from the convex optimization problem.
- can learn and generalize in **huge dimensional input spaces**.
- make use of **kernels** in the models.

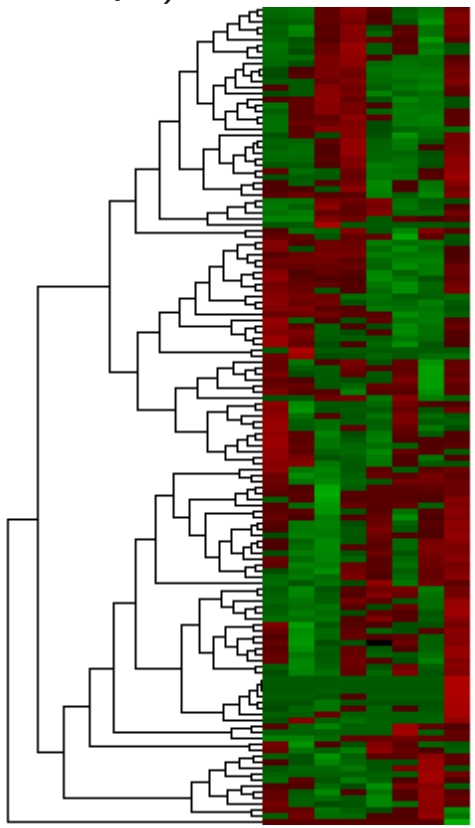
... a few examples

Examples of applications with classification of data in **high dimensional input spaces**:

- microarray experiments in bioinformatics
- classification of brain tumours
- text categorization
- handwritten digit recognition

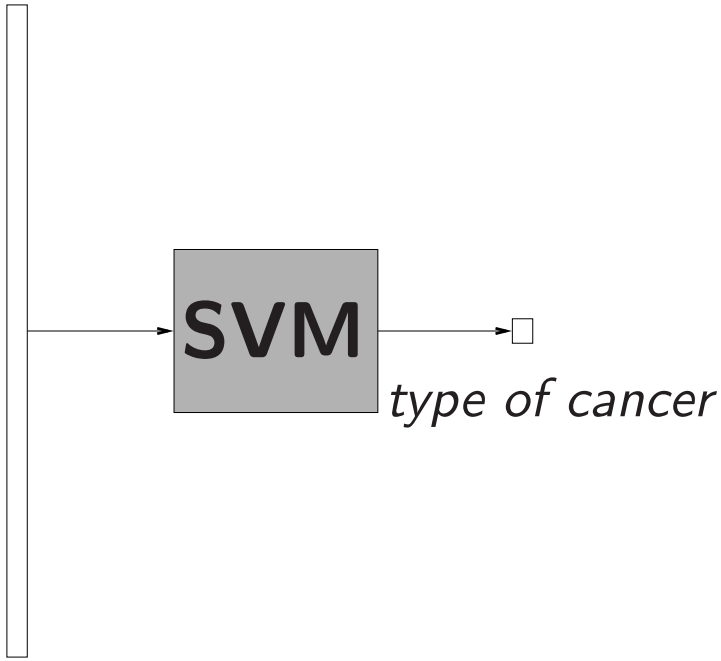
Microarray data (1)

Microarray technology (DNA chips) in order to measure activities of genes:



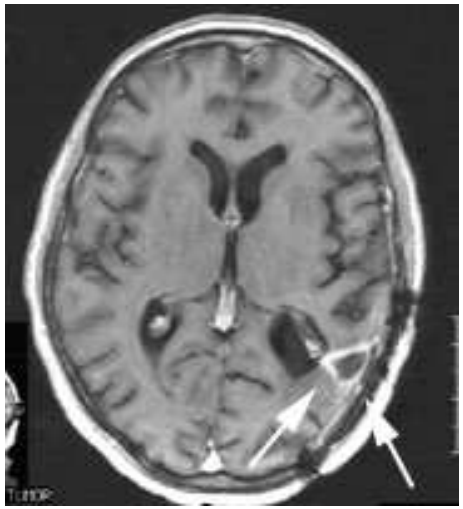
Rows: expression levels of genes; columns: different experiments
Typically: thousands of genes, but only 50-100 experiments

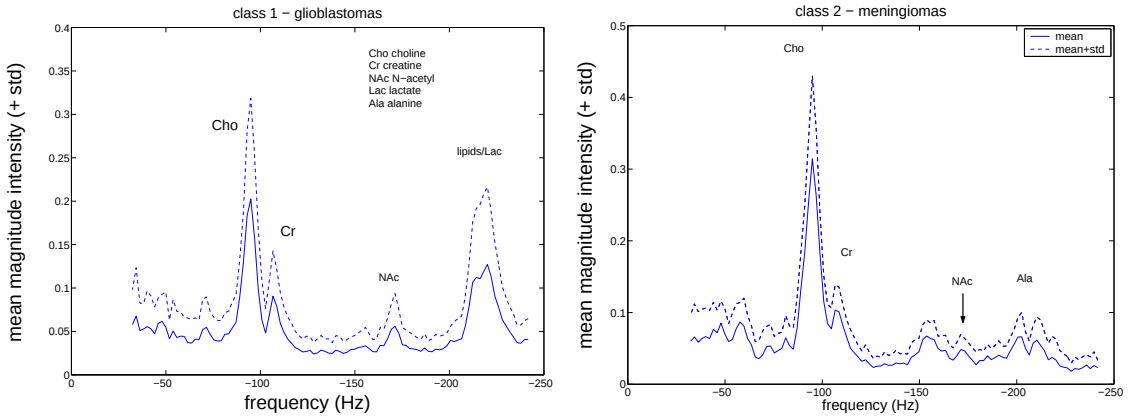
Microarray data (2)



MIT Leukemia dataset (7,129 gene expressions; 38 training and 34 test samples): for classical neural networks such as MLPs one first has to do a dimensionality reduction of the input space. SVMs on the other hand are able to learn and generalize on input vectors with 7000 genes!

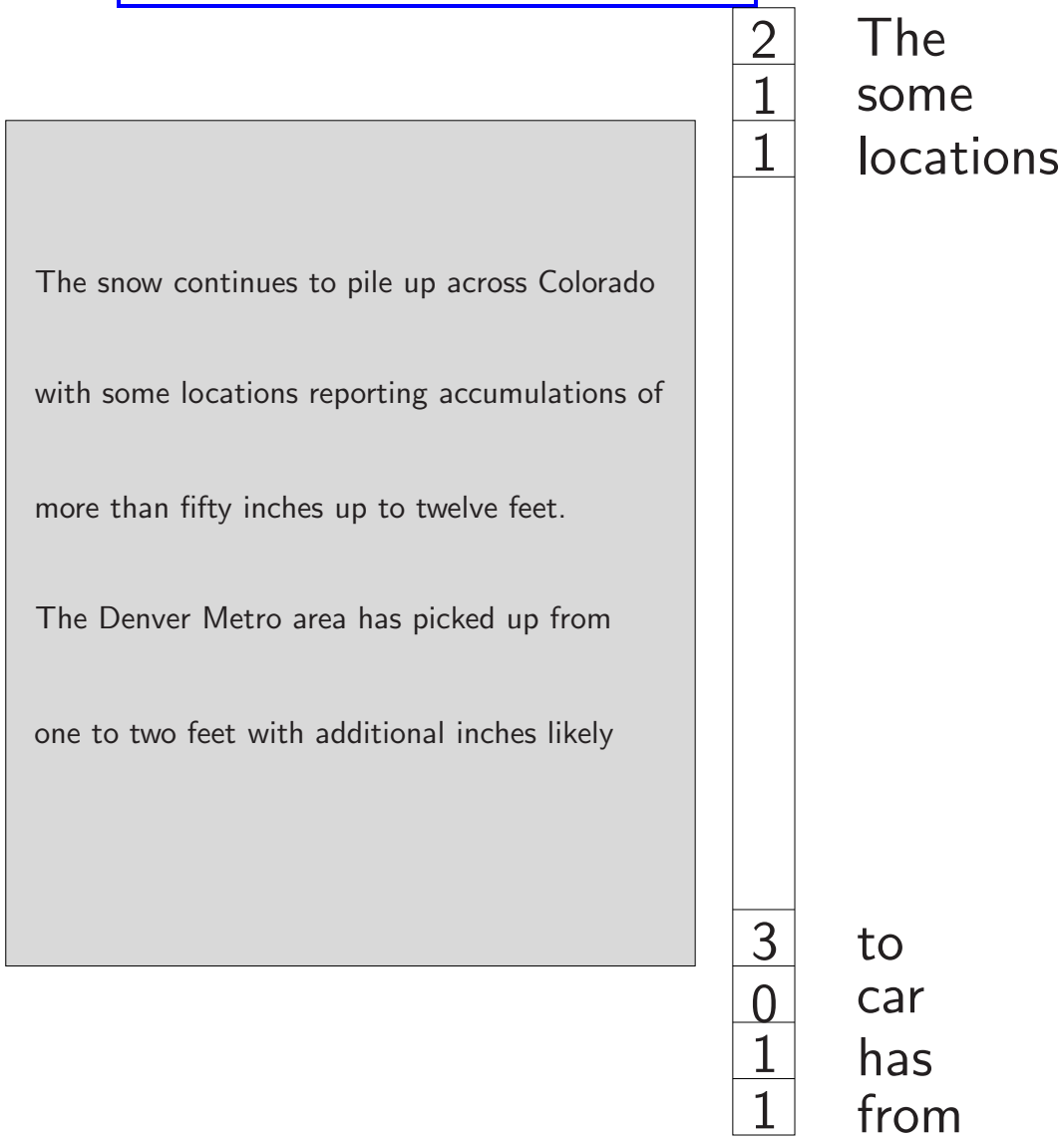
Classification of brain tumors



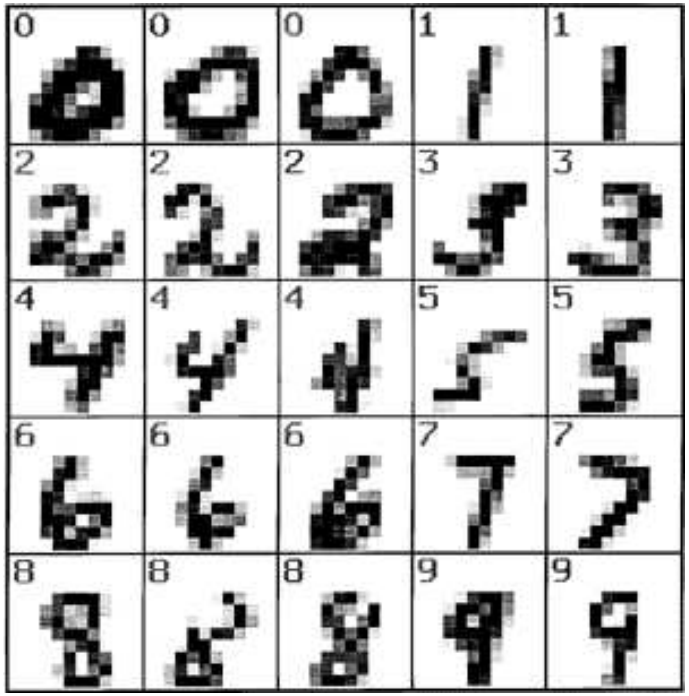


Classification of tumors based on measured NMR spectra.

Text categorization

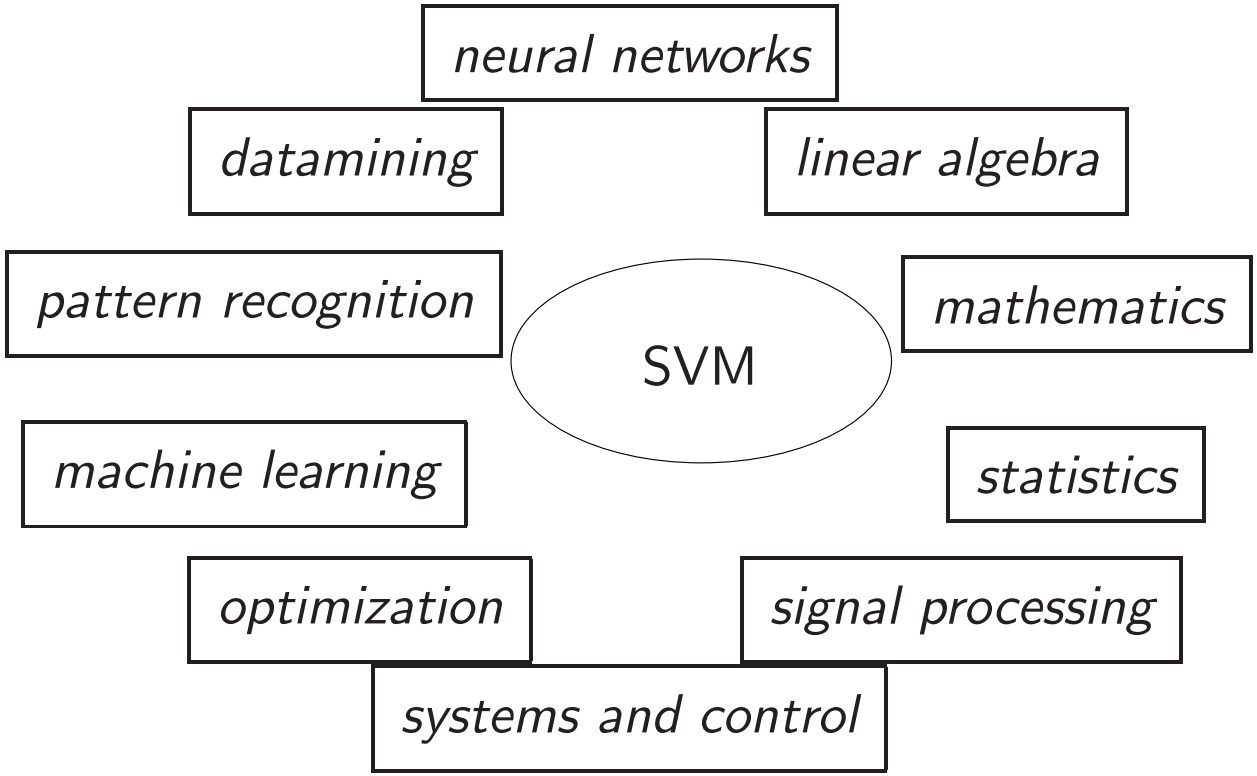


Handwritten digit recognition



Classify handwritten digits based on the given images
(Typical benchmark: US Postal database)

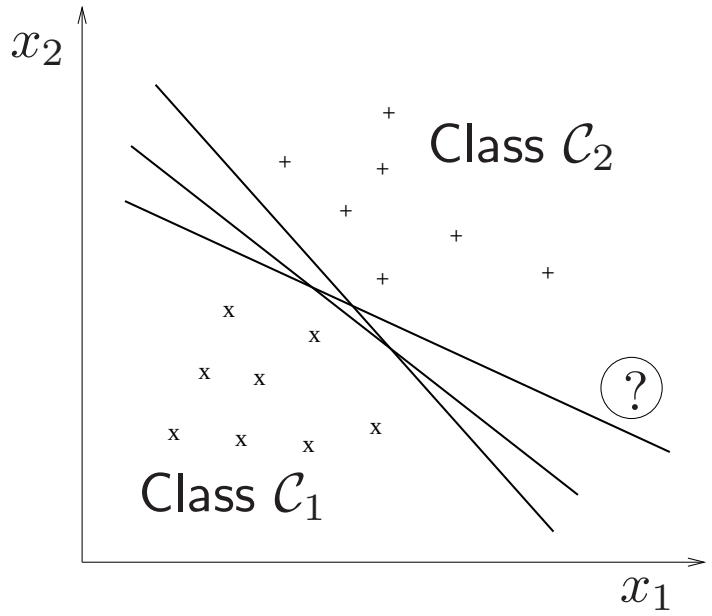
Interdisciplinary challenges



Link between SVM theory and its applications in many different areas.

Classification - Linear separating hyperplane (1)

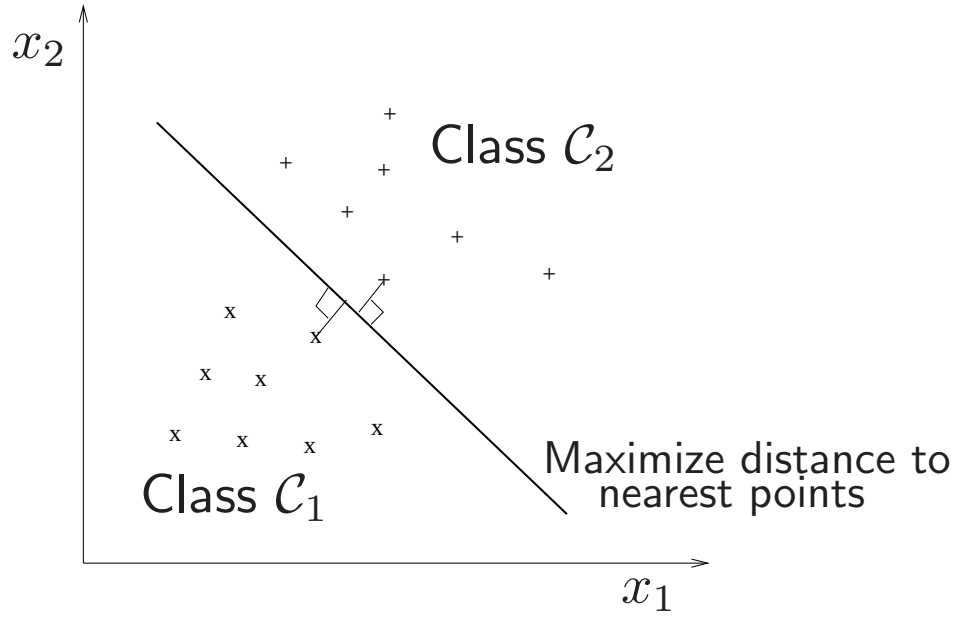
- Consider two linearly separable classes:



- Clearly, different separating hyperplanes are possible.

Classification - Linear separating hyperplane (2)

- Consider instead the hyperplane that maximizes the distance to the nearest points:

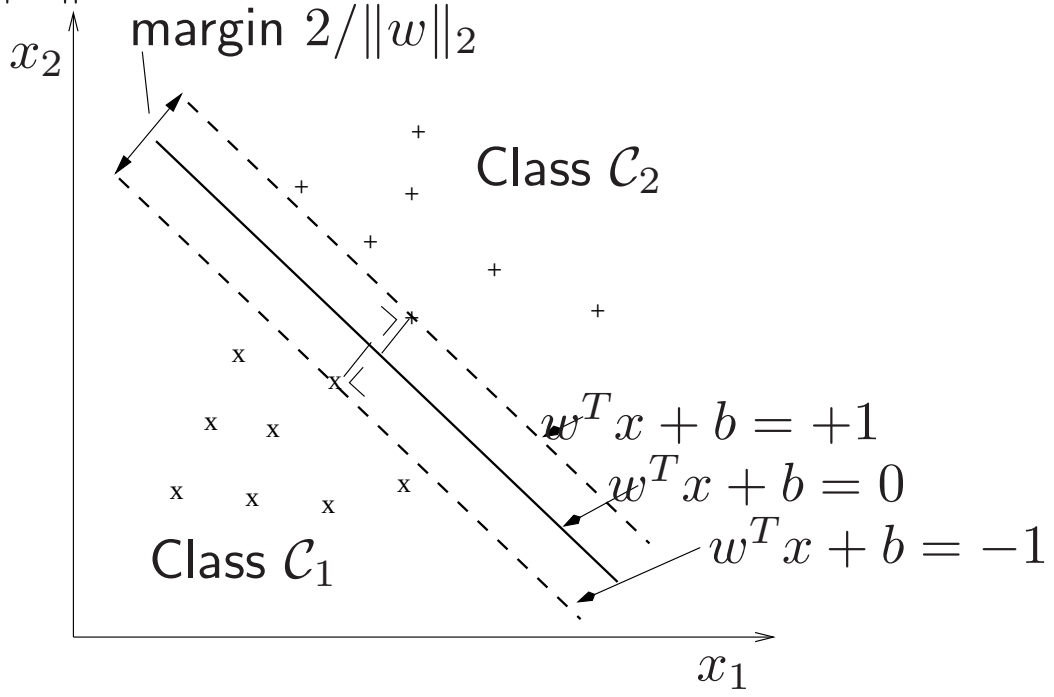


Classification - Linear separating hyperplane (3)

- A rescaling of the problem is done such that

$$\min_i |w^T x_i + b| = 1$$

i.e. the scaling is done such that the point closest to the hyperplane has a distance $1/\|w\|_2$.



Classification - Linear separating hyperplane (4)

- The **margin** between the classes is then equal to $2/\|w\|_2$.
- Minimizing $\|w\|_2$ corresponds to maximizing the margin.
- Note that:

$$w^T x_1^* + b = 1 \text{ and } w^T x_2^* + b = -1$$
$$\Rightarrow w^T(x_1^* - x_2^*) = 2 \Rightarrow \frac{w^T}{\|w\|_2}(x_1^* - x_2^*) = \frac{2}{\|w\|_2}$$

where x_1^* and x_2^* are the nearest points within the hyperplanes of the two different classes.

SVM Classifier - Linear separable case (1)

- Large margin separating hyperplane [Vapnik, 1964].

- Given a training set $\{x_k, y_k\}_{k=1}^N$, input patterns $x_k \in \mathbb{R}^n$, output patterns $y_k \in \mathbb{R}$ where $y_k \in \{-1, +1\}$. Assume

$$\begin{cases} w^T x_k + b \geq +1 & , \quad \text{if} \quad y_k = +1 \\ w^T x_k + b \leq -1 & , \quad \text{if} \quad y_k = -1 \end{cases}$$

  This is equivalent to

$$y_k[w^T x_k + b] \geq 1, \quad k = 1, ..., N$$

  (i.e. require that **all** training data are **correctly** classified)

SVM Classifier - Linear separable case (2)

- Optimization problem (primal problem):

$$\min_{w,b} \frac{1}{2} w^T w \quad \text{s.t.} \quad y_k[w^T x_k + b] \geq 1, \quad k = 1, ..., N$$

  i.e. **maximize margin** and classify all training data correctly.

- Lagrangian

$$\mathcal{L}(w, b; \alpha) = \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \{y_k[w^T x_k + b] - 1\}$$

  with Lagrange multipliers $\alpha_k \geq 0$ for $k = 1, ..., N$.

SVM Classifier - Linear separable case (3)

- Solution given by **saddle point** of Lagrangian:

$$\max_{\alpha} \min_{w,b} \mathcal{L}(w, b; \alpha)$$

One obtains

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k y_k x_k \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \end{cases}$$

Resulting classifier

$$y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k x_k^T x + b]$$

SVM Classifier - Linear separable case (4)

- Quadratic Programming (QP) problem (**dual problem**): solve the problem in the Lagrange multipliers

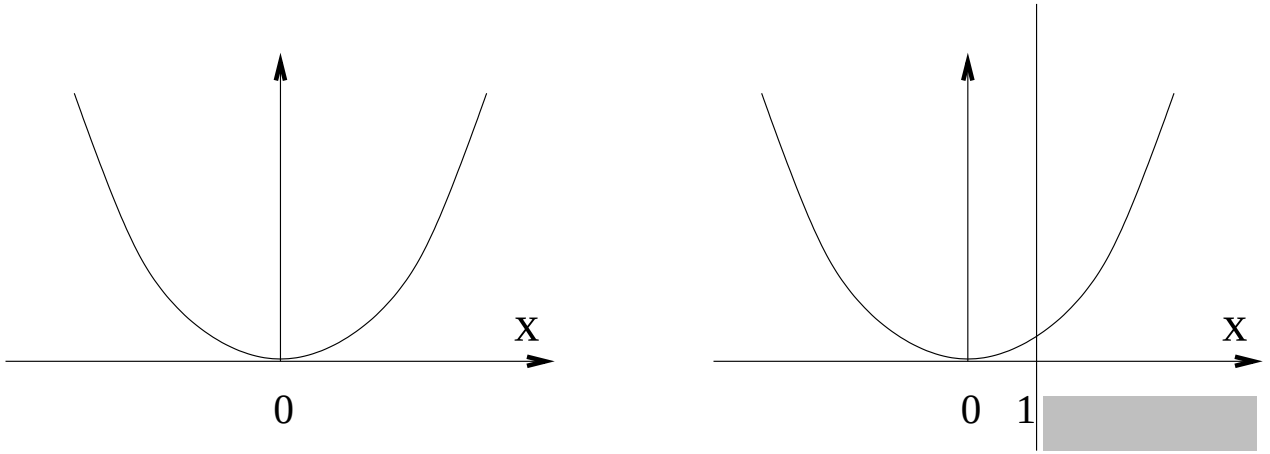
$$\max_{\alpha} \mathcal{Q}(\alpha) = -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l x_k^T x_l \ \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k$$

such that

$$\begin{cases} \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \alpha_k \geq 0, \ \ k = 1, ..., N \end{cases}$$

Intermezzo: constrained optimization (1)

A simple quadratic function ...



- Example of **unconstrained** optimization problem (solution: $x = 0$)

$$\min_x x^2$$

- Example of **constrained** optimization problem (active constraint)

$$\min_x x^2 \text{ such that } x \geq 1$$

Intermezzo: constrained optimization (2)

- **First possible way** to solve the constrained problem: Consider the Lagrangian

$$\mathcal{L}(x;\lambda) = x^2 - \lambda(x-1)$$

with Lagrange multiplier $\lambda \geq 0$. The optimal solution is characterized by the **saddle point**

$$\max_{\lambda} \min_{x} \mathcal{L}(x;\lambda)$$

First compute

$$\frac{\partial \mathcal{L}}{\partial x} = 2x - \lambda = 0 \Rightarrow x = \frac{\lambda}{2}.$$

Then substitute this expression back into the Lagrangian and solve the following dual problem (in the Lagrange multiplier)

$$\max_{\lambda} (\frac{\lambda}{2})^2 - \lambda(\frac{\lambda}{2} - 1)$$

which gives the solution $\lambda = 2$. Hence $x = 1$.

Intermezzo: constrained optimization (3)

- **Second possible way** to solve the constrained problem: Make use of a **slack variable** by writing the constraint $x \geq 1$ as $x - \delta^2 = 1$ with δ an additional unknown. The problem becomes

$$\min_{x,\delta} x^2 \text{ such that } x - \delta^2 = 1$$

with Lagrangian

$$\mathcal{L}(x, \delta; \lambda) = x^2 - \lambda(x - \delta^2 - 1)$$

and

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0 & \rightarrow \quad 2x - \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \delta} = 0 & \rightarrow \quad -2\lambda\delta = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 0 & \rightarrow \quad x - \delta^2 - 1 = 0. \end{cases}$$

From the second condition it follows that when $\lambda = 0$ one would have $x = 0$ which violates the constraint. Hence, one should have $\delta = 0$ which gives then as solution $x = 1$.

Solution by Quadratic programming

QP problem: Note that

- the size of the matrix grows with the number of training data points (if $N = 10^6$ then the size of the matrix is $10^6 \times 10^6$!)

- the size of α solution vector is only determined by the number of training data N and not by the dimension of the input space.

Properties of the solution

- **Positive definite** matrix in QP problem: unique α solution;
  Positive semidefinite matrix in QP problem: possibly many solutions to QP problem, but w, b solution is unique.

- The solution is **sparse** i.e. many elements in the solution vector α are zero. Hence, instead of taking the sum over all training data in

$$y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k x_k^T x + b]$$

  one has

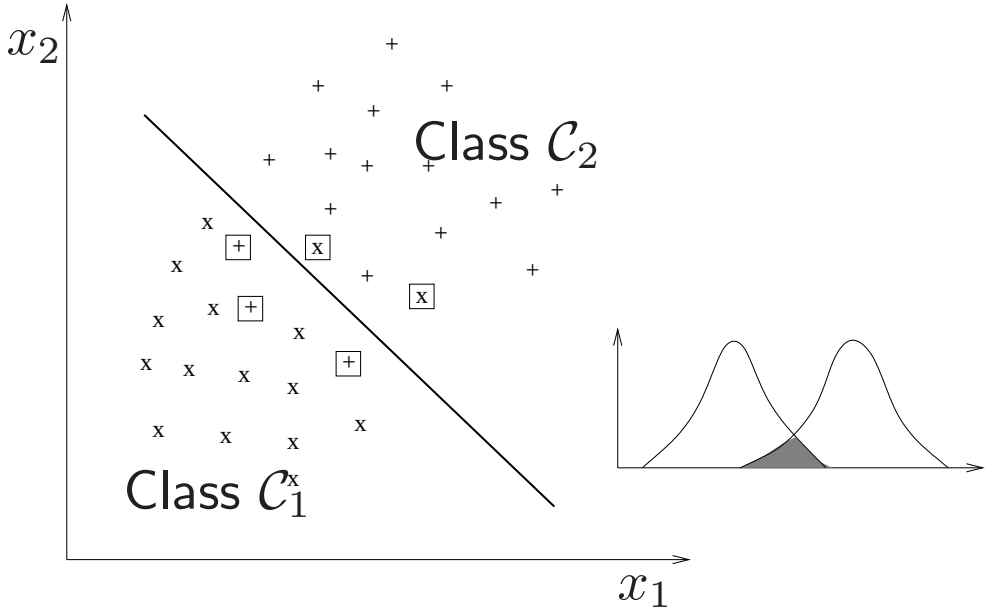
$$y(x) = \text{sign}[\sum_{k \in \mathcal{S}_{\text{SV}}} \alpha_k y_k x_k^T x + b]$$

  ($\mathcal{S}_{\text{SV}}$ denotes the set of all **nonzero** α_k's)

- **support values:** all non-zero α_k;
  support vectors: training data corresponding to $\alpha_k \neq 0$.

SVM Classifier - non-separable data (1)

- Non-separable data (e.g. due to overlapping distributions):



- Modify the inequalities into

$$y_k[w^T x_k + b] \geq 1 - \xi_k, \quad k = 1, ..., N$$

with **slack variables** $\xi_k \geq 0$ (the original inequalities can be violated for certain points if needed). Misclassified points correspond to $\xi_k > 1$.

SVM Classifier - non-separable data (2)

- Optimization problem (primal problem):

$$\min_{w,b,\xi} \mathcal{J}(w,\xi) = \frac{1}{2} w^T w + c \sum_{k=1}^{N} \xi_k$$

subject to

$$\begin{cases} y_k[w^T x_k + b] \geq 1 - \xi_k, & k = 1, ..., N \\ \xi_k \geq 0, & k = 1, ..., N. \end{cases}$$

- Lagrangian

$$\mathcal{L}(w, b, \xi; \alpha, \nu) = \mathcal{J}(w,\xi) - \sum_{k=1}^{N} \alpha_k \{y_k[w^T x_k + b] - 1 + \xi_k\} - \sum_{k=1}^{N} \nu_k \xi_k$$

with Lagrange multipliers $\alpha_k \geq 0$, $\nu_k \geq 0$ for $k = 1, ..., N$.

SVM Classifier - non-separable data (3)

- Solution given by saddle point of Lagrangian:

$$\max_{\alpha,\nu} \min_{w,b,\xi} \mathcal{L}(w, b, \xi; \alpha, \nu)$$

One obtains

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k y_k x_k \\ \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \quad \rightarrow \quad 0 \leq \alpha_k \leq c, \ k = 1, ..., N \end{cases}$$

SVM Classifier - non-separable data (4)

- Quadratic programming problem (dual problem):

$$\max_{\alpha_k} \mathcal{Q}(\alpha) = -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l \, x_k^T x_l \, \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k$$

  such that

$$\begin{cases} \sum_{k=1}^{N} \alpha_k y_k = 0 \\ 0 \leq \alpha_k \leq c, \;\; k = 1, ..., N. \end{cases}$$

  This leads to so-called **box constraints** on α_k.

- Computation of b:
  product of dual variables and constraints should be zero at the optimum
  (from KKT (Karush-Kuhn-Tucker) conditions)

Intermezzo: KKT conditions (1)

- **Constrained optimization** problem

$$\begin{array}{lll} \text{minimize} & f(x) & x \in \mathbb{R}^n \\ \text{subject to} & c_i(x) = 0, & i \in \mathcal{C}_E \\ & c_i(x) \geq 0, & i \in \mathcal{C}_I \end{array}$$

  where $f(x)$ objective function, c_i $(i = 1, 2, ..., p)$ constraint functions, $\mathcal{C}_E$ set of equality constraints, $\mathcal{C}_I$ set of inequality constraints.

- Any point that satisfies all the constraints is called a **feasible point**.

- The **Lagrangian** function is $\mathcal{L}(x, \lambda) = f(x) - \sum_i \lambda_i c_i(x)$.

Intermezzo: KKT conditions (2)

- First order (i.e. first derivative) necessary conditions:
  If x^* is a local minimizer of the above constrained optimization problem and a certain regularity assumption holds at x^*, then there exist Lagrange multipliers λ^* such that x^*, λ^* satisfy the following system:

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda) &= 0 & \\ c_i(x) &= 0 & i \in \mathcal{C}_E \\ c_i(x) &\geq 0 & i \in \mathcal{C}_I \\ \lambda_i &\geq 0 & i \in \mathcal{C}_I \\ \lambda_i c_i(x) &= 0 & \forall i \end{aligned}$$

  called the **Karush-Kuhn-Tucker (KKT) conditions**.

- The final condition $\lambda_i^* c_i^* = 0$ is referred to as the **complementarity condition** and states that λ_i^* and c_i^* cannot be both non-zero.

Intermezzo: KKT conditions (3)

- **Example:**

$$\min_x x^2 \text{ such that } x \geq 1$$

Consider the Lagrangian

$$\mathcal{L}(x;\lambda) = x^2 - \lambda(x-1)$$

with Lagrange multiplier $\lambda \geq 0$.

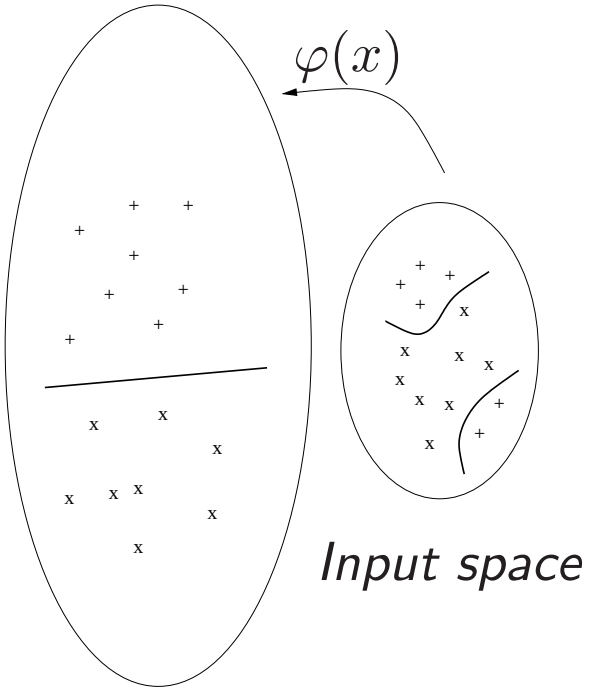
- **KKT conditions:**

$$\begin{aligned}
&\nabla_x \mathcal{L} = 2x - \lambda = 0 \\
&x - 1 \geq 0 \\
&\lambda \geq 0 \\
&\lambda(x-1) = 0
\end{aligned}$$

From the last condition one either has $\lambda = 0$ or $x = 1$ but $\lambda = 0$ leads to $x = 0$ which violates the constraints. Hence, the solution $x = 1$.

Nonlinear SVMs: feature map

Mapping of input data into a high dimensional feature space [Vapnik, 1995]:



Construction of linear separating hyperplane in the feature space, after nonlinear mapping $\varphi(x)$ of the input data into the feature space.
However, no explicit construction of nonlinear mapping $\varphi(x)$ is needed! This will become clear soon...

Nonlinear SVM classifier (1)

- Given a training set $\{x_k, y_k\}_{k=1}^N$, input patterns $x_k \in \mathbb{R}^n$, class labels $y_k \in \mathbb{R}$ where $y_k \in \{-1, +1\}$

- Classifier:

$$y(x) = \text{sign}[w^T \varphi(x) + b]$$

with $\varphi(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_h}$ mapping to high dimensional feature space (can be infinite dimensional)

- For separable data, assume

$$\begin{cases} w^T \varphi(x_k) + b \geq +1 \quad , & \text{if} \quad y_k = +1 \\ w^T \varphi(x_k) + b \leq -1 \quad , & \text{if} \quad y_k = -1 \end{cases}$$

which is equivalent to

$$y_k[w^T \varphi(x_k) + b] \geq 1, \quad k = 1, ..., N$$

Nonlinear SVM classifier (2)

- Optimization problem (**non-separable case**):

$$\min_{w,b,\xi} \mathcal{J}(w,\xi) = \frac{1}{2} w^T w + c \sum_{k=1}^{N} \xi_k$$

subject to

$$\begin{cases} y_k[w^T \varphi(x_k) + b] \geq 1 - \xi_k, \quad k = 1, ..., N \\ \xi_k \geq 0, \quad k = 1, ..., N. \end{cases}$$

- Construct **Lagrangian**:

$$\mathcal{L}(w, b, \xi; \alpha, \nu) = \mathcal{J}(w, \xi_k) - \sum_{k=1}^{N} \alpha_k \{y_k[w^T \varphi(x_k) + b] - 1 + \xi_k\} - \sum_{k=1}^{N} \nu_k \xi_k$$

with Lagrange multipliers $\alpha_k \geq 0$, $\nu_k \geq 0$ $(k = 1, ..., N)$.

Nonlinear SVM classifier (3)

- Solution given by **saddle point** of Lagrangian:

$$\max_{\alpha,\nu} \min_{w,b,\xi} \mathcal{L}(w, b, \xi; \alpha, \nu)$$

One obtains

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_k} = 0 \quad \rightarrow \quad 0 \leq \alpha_k \leq c, \ k = 1, ..., N \end{cases}$$

Nonlinear SVM classifier (4)

- Quadratic programming problem (**dual problem**):

$$\max_{\alpha_k} \mathcal{Q}(\alpha) = -\frac{1}{2} \sum_{k,l=1}^{N} y_k y_l \, K(x_k, x_l) \, \alpha_k \alpha_l + \sum_{k=1}^{N} \alpha_k$$

such that

$$\begin{cases} \sum_{k=1}^{N} \alpha_k y_k = 0 \\ \\ 0 \leq \alpha_k \leq c, \;\; k = 1, ..., N. \end{cases}$$

Note: w and $\varphi(x_k)$ are **not calculated**!

- **Mercer theorem:** choose positive definite kernel such that

$$K(x_k, x_l) = \varphi(x_k)^T \, \varphi(x_l)$$

Nonlinear SVM classifier (5)

- Obtained **classifier**:

$$y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k \, y_k \, K(x, x_k) + b]$$

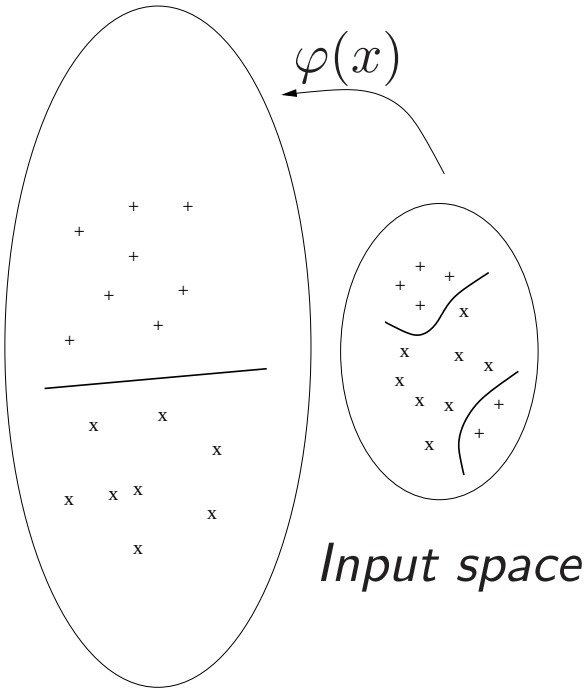
with α_k positive real constants, b real constant, that follow as solution to the QP problem. Non-zero α_k are called support values and the corresponding data points are called support vectors. The bias term b follows from KKT conditions.

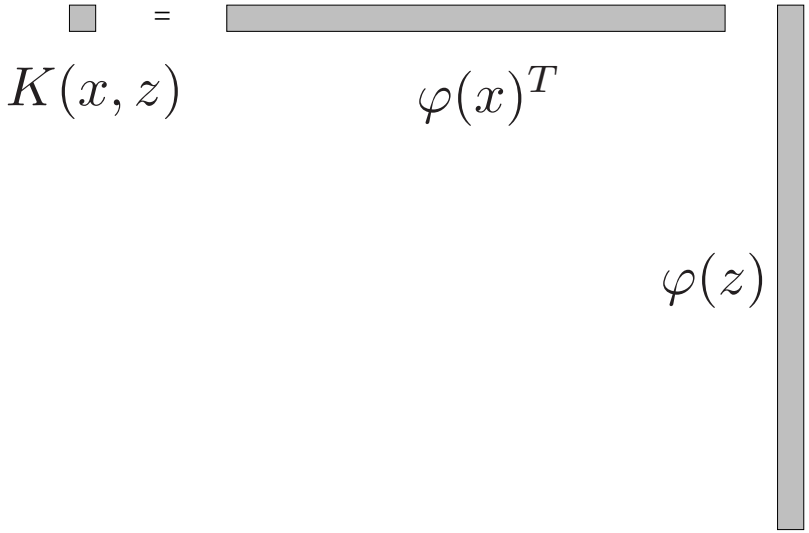
- Some possible **kernels** $K(\cdot, \cdot)$:

$K(x, x_k) = x_k^T x$ (linear SVM)
$K(x, x_k) = (x_k^T x + \tau)^d$ (polynomial SVM of degree d and $\tau \geq 0$)
$K(x, x_k) = \exp(-\|x - x_k\|_2^2 / \sigma^2)$ (RBF SVM)
$K(x, x_k) = \tanh(\kappa \, x_k^T x + \theta)$ (MLP SVM)

In the case of RBF and MLP kernel, the number of hidden units corresponds to the number of support vectors.

Feature space and kernel trick



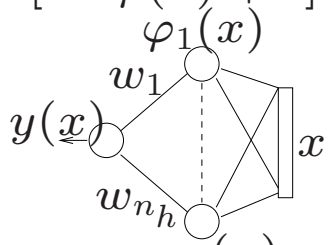


Primal-dual interpretations of SVMs

Primal problem

Parametric: $w \in \mathbb{R}^{n_h}$

$y(x) = \text{sign}[w^T \varphi(x) + b]$



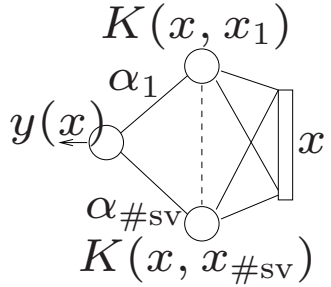
Kernel trick

$K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$

Dual problem

Kernel-based: $\alpha \in \mathbb{R}^N$

$y(x) = \text{sign}[\sum_{k=1}^{\#\text{sv}} \alpha_k y_k K(x, x_k) + b]$



Mercer Theorem

There exists a mapping φ and an expansion

$$K(x,z) = \sum_i \varphi_i(x)\varphi_i(z), \qquad x, z \in \mathbb{R}^n$$

if and only if, for any $g(x)$ such that

$$\int g(x)^2 dx \quad \text{is finite}$$

one has

$$\int K(x,z)g(x)g(z)dxdz \geq 0.$$

(hence a **positive definite kernel**)

Feature space to kernel representation

Primal weight space (feature space):

$$y(x) = \text{sign}[w^T \varphi(x) + b]$$

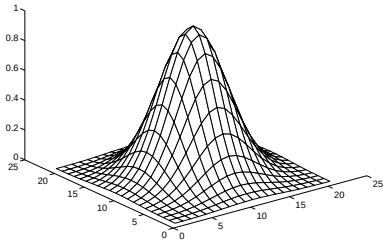
Use $w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k)$:

$$y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k \varphi(x_k)^T \varphi(x) + b]$$

Dual space (application of Mercer theorem):

$$y(x) = \text{sign}[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b]$$

RBF kernel

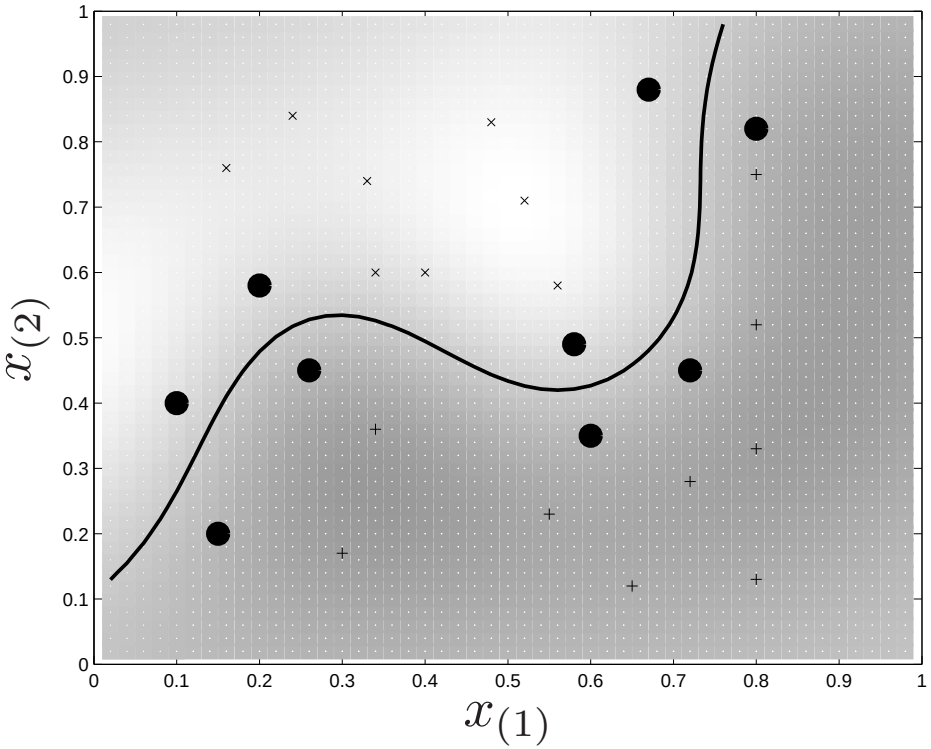


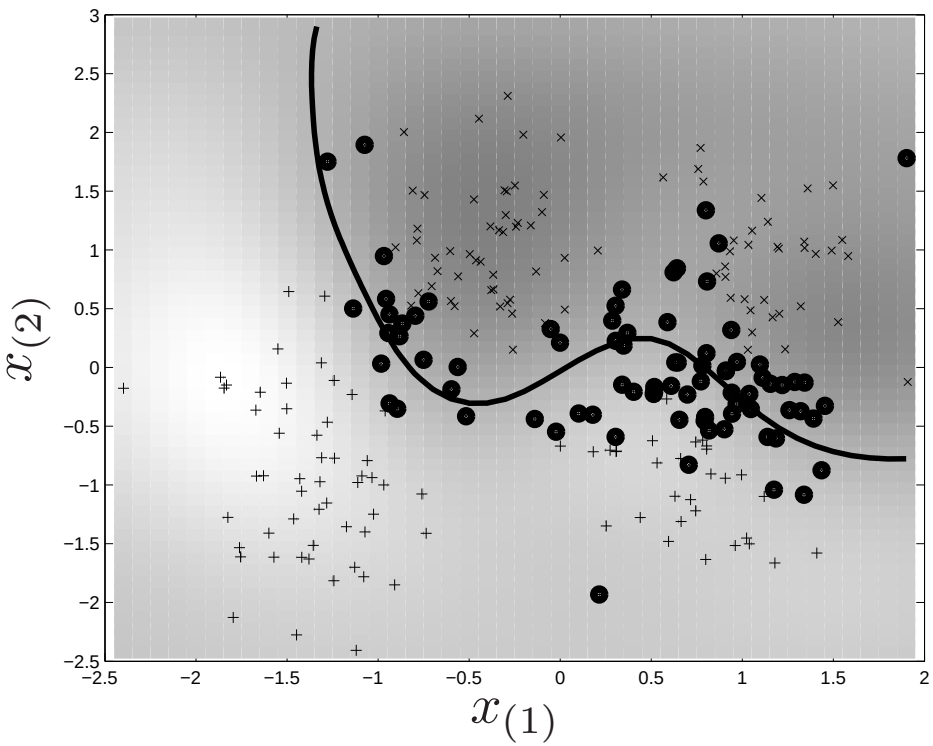
SVM classifier in dual space:

$$
\begin{aligned}
y(x) &= \operatorname{sign}\Big[\sum_{k=1}^{N} \alpha_k y_k \exp\left(-\frac{\|x - x_k\|_2^2}{\sigma^2}\right) + b\Big] \\
&= \operatorname{sign}\Big[\sum_{k \in \mathcal{S}_{\mathrm{SV}}} \alpha_k y_k \exp\left(-\frac{\|x - x_k\|_2^2}{\sigma^2}\right) + b\Big]
\end{aligned}
$$

Each hidden unit corresponds to a support vector (corresponding to non-zero support values α_k). Hence $\boxed{\#\text{ hidden units} = \#\text{ support vectors}}$

Geometrical interpretation of SV





- Decision boundary can be expressed in terms of a limited number of support vectors (subset of given training data); sparseness property

- Classifier follows from the solution to a convex QP problem.