Support Vector Machines: Maximum Margin Classifiers

Machine Learning and Pattern Recognition:
September 16, 2008

Piotr Mirowski
Based on slides by Sumit Chopra and Fu-Jie Huang

Outline

- What is behind Support Vector Machines?
  - Constrained Optimization
  - Lagrange Duality
- Support Vector Machines in Detail
  - Kernel Trick
  - LibSVM demo

Binary Classification Problem

- Given: Training data generated according to the distribution D

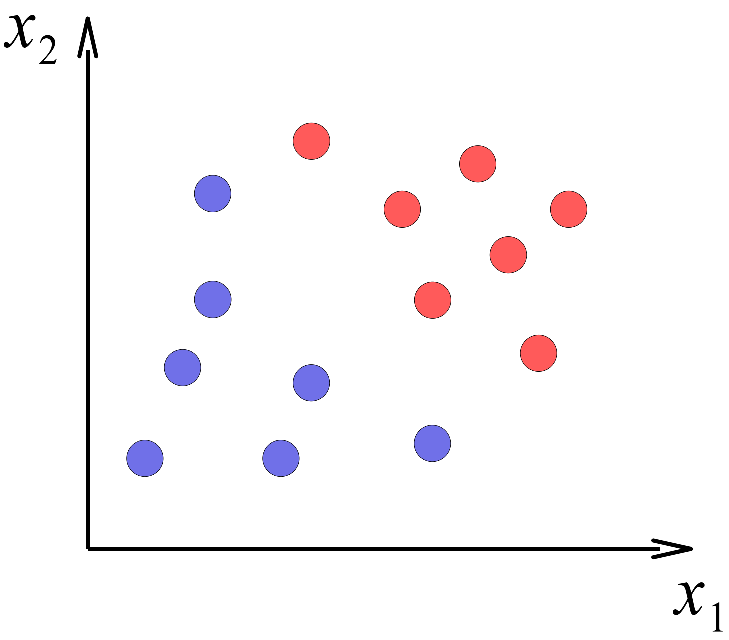
$$\underbrace{x_1}_{\text{input}},\ \underbrace{y_1}_{\text{label}}),\ldots,(x_p,y_p)\in\underbrace{\Re^n}_{\text{input space}}\times\underbrace{\{-1,1\}}_{\text{label space}}$$

- Problem: Find a classifier (a function) $h(x):\Re^n\rightarrow\{-1,1\}$ such that it generalizes well on the test set obtained from the same distribution D

- Solution:
  - Linear Approach: linear classifiers (e.g. Perceptron)
  - Non Linear Approach: non-linear classifiers (e.g. Neural Networks, SVM)

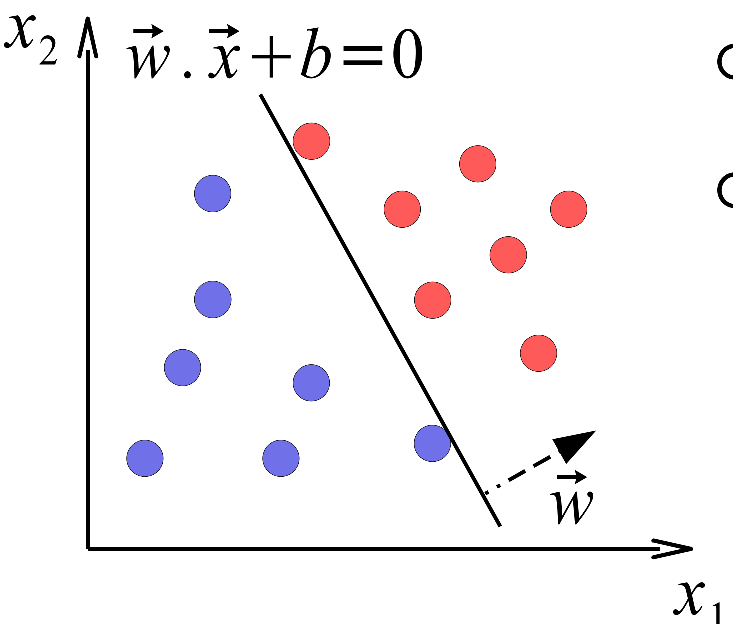
Linearly Separable Data

- Assume that the training data is linearly separable



Linearly Separable Data

- Assume that the training data is linearly separable



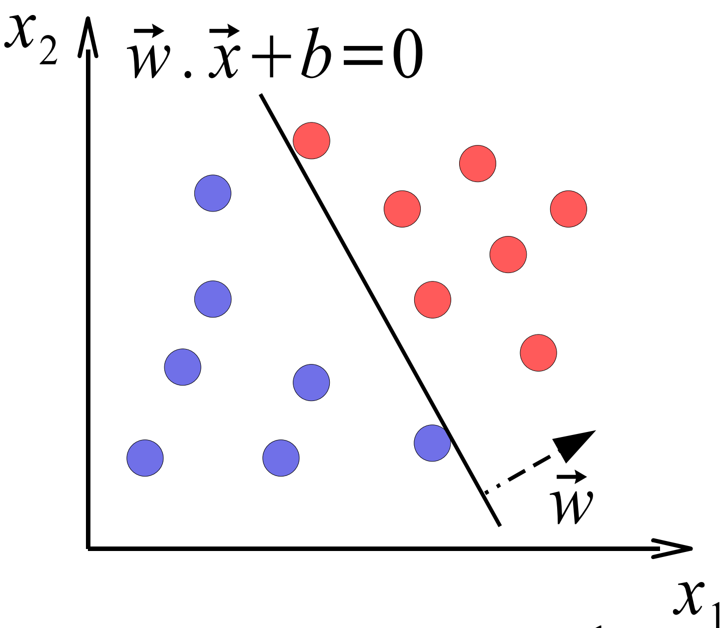
abscissa on axis parallel to $\vec{w}$

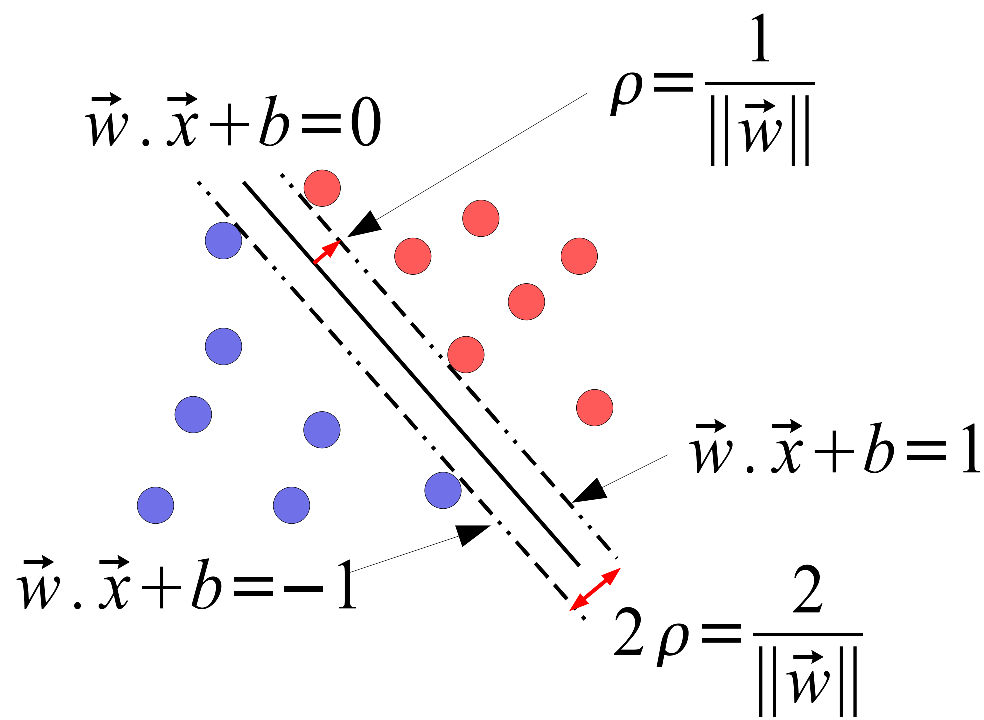
abscissa of origin *0* is b

- Then the classifier is: $h(x) = \vec{w}.\vec{x}+b \quad where\ w\in\Re^n, b\in\Re$
- Inference: $sign(h(x)) \in \{-1,1\}$

Linearly Separable Data

- Assume that the training data is linearly separable





- Margin $\rho = \frac{1}{\|\vec{w}\|}$
- Maximize margin ρ (or 2ρ) so that:

  For the closest points: $h(x) = \vec{w}.\vec{x}+b \in \{-1,1\}$

Optimization Problem

- A Constrained Optimization Problem

$$\min_{w} \frac{1}{2} \|\vec{w}\|^2$$
$$s.t.:$$
$$y_i(\vec{w}\,.\,\vec{x}_i + b) \geqslant 1, \quad i = 1, \ldots, m$$

label input

- Equivalent to maximizing the margin $\rho = \frac{1}{\|\vec{w}\|}$
- A convex optimization problem:
  - Objective is convex
  - Constraints are affine hence convex

Optimization Problem

- Compare:

$$\min_{w} \underbrace{\frac{1}{2}\|\vec{w}\|^2}_{\text{objective}}$$

$$s.t.:$$

$$\underbrace{y_i(\vec{w}\,.\,\vec{x}_i + b) \geqslant 1}_{\text{constraints}}, \quad i = 1, \ldots, m$$

- With:

$$\min_{w} \left(\sum_{i=1}^{p} \underbrace{\left(-y_i(\vec{w}\,.\,\vec{x}_i + b)\right)}_{\text{energy/errors}} + \underbrace{\frac{\lambda}{2}\|\vec{w}\|^2}_{\text{regularization}} \right)$$

Optimization: Some Theory

- The problem:

$$
\begin{aligned}
&\min_x f_0(x) && \longleftarrow \text{objective function} \\
&s.t.: && \\
&f_i(x) \leqslant 0, \quad i=1,\dots,m && \longleftarrow \text{inequality constraints} \\
&h_i(x) = 0, \quad i=1,\dots,p && \longleftarrow \text{equality constraints}
\end{aligned}
$$

- Solution of problem: x^o
  - Global (unique) optimum – if the problem is convex
  - Local optimum – if the problem is not convex

Optimization: Some Theory

- Example: Standard Linear Program (LP)

$$\min_{x} \ c^T x$$
$$s.t.:$$
$$Ax = b$$
$$x \geqslant 0$$

- Example: Least Squares Solution of Linear Equations (with L_2 norm regularization of the solution x) i.e. Ridge Regression

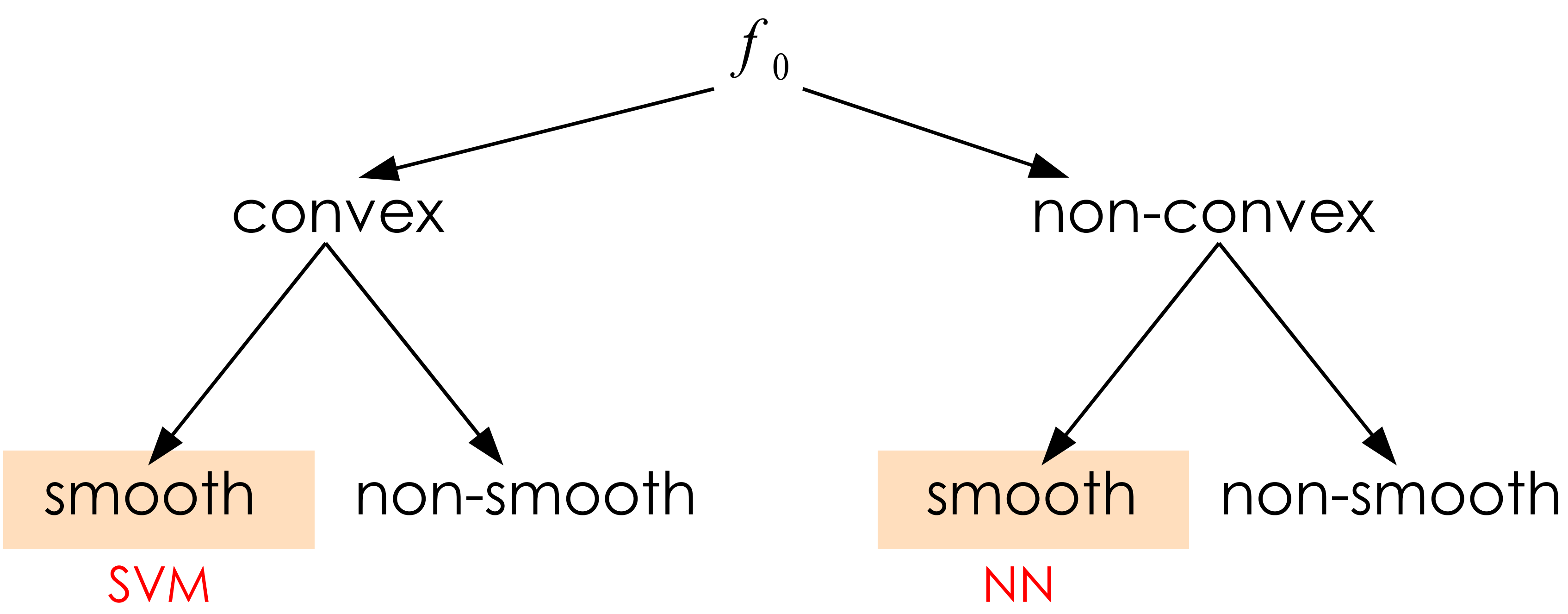
$$\min_{x} \ x^T x$$
$$s.t.:$$
$$Ax = b$$

Big Picture

- Constrained / unconstrained optimization
- Hierarchy of objective function:

  smooth = infinitely derivable

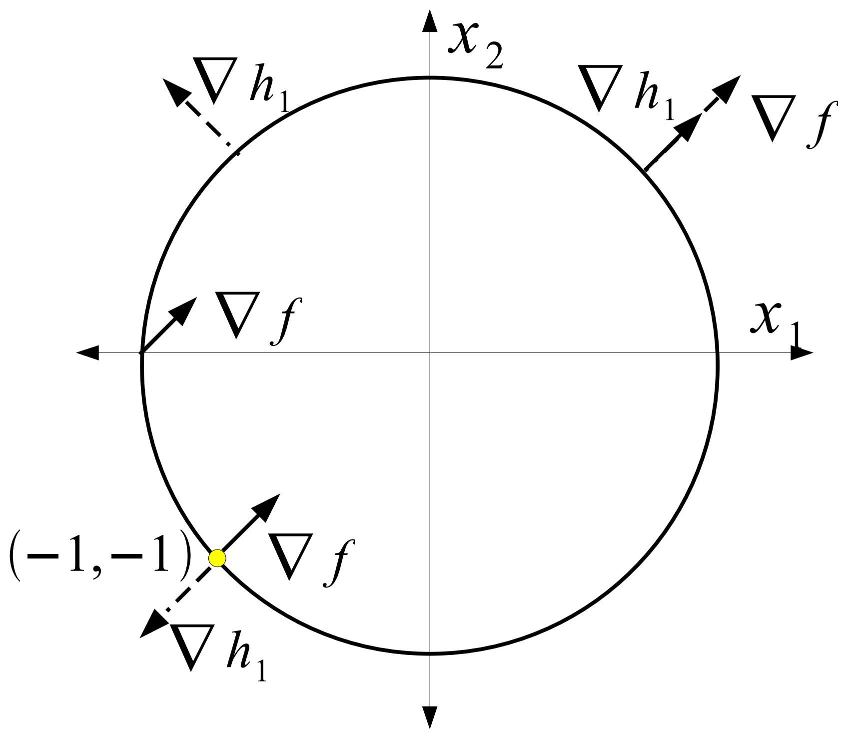
  convex = has a global optimum



Toy Example: Equality Constraint

- Example 1: $\min \quad x_1 + x_2 \qquad \equiv f$

  $s.t.: \quad x_1^2 + x_2^2 - 2 = 0 \qquad \equiv h_1$



$$\nabla f = \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{pmatrix}$$

$$\nabla h_1 = \begin{pmatrix} \dfrac{\partial h_1}{\partial x_1} \\ \dfrac{\partial h_1}{\partial x_2} \end{pmatrix}$$

- At Optimal Solution: $\boxed{\nabla f(x^o) = \lambda_1^o \nabla h_1(x^o)}$

Toy Example: Equality Constraint

- x is not an optimal solution, if there exists $s \neq 0$ such that

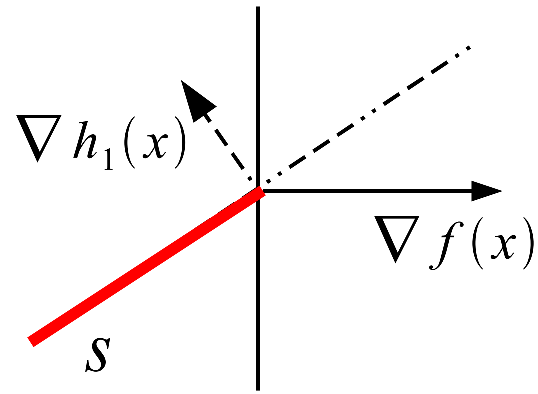
$$h_1(x+s) = 0$$
$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

$$\cancel{h_1(x+s)} = \cancel{h_1(x)} + \nabla h_1(x)^T s = \nabla h_1(x)^T s = 0 \qquad (1)$$

$$f(x+s) - f(x) = \nabla f(x)^T s < 0 \qquad (2)$$

- Such an s can exist only when $\nabla h_1(x)$ and $\nabla f(x)$ are not parallel



Toy Example: Equality Constraint

- Thus we have

$$\boxed{\nabla f(x^o) = \lambda_1^o \nabla h_1(x^o)}$$

- The Lagrangian

Lagrange multiplier or dual variable for h_1

$$\boxed{L(x, \lambda_1) = f(x) - \lambda_1 h_1(x)}$$

- Thus at the solution

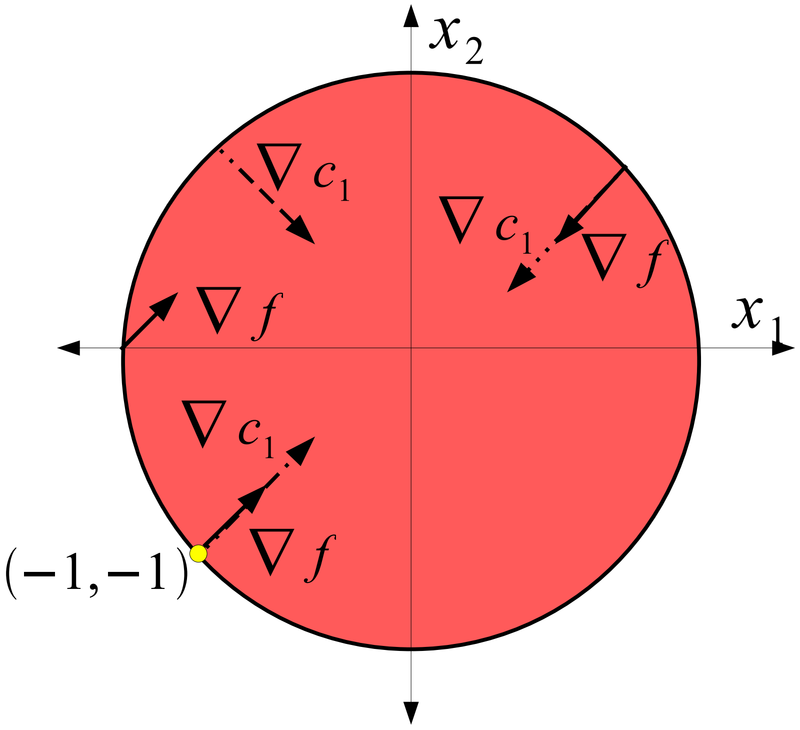
$$\boxed{\nabla_x L(x^o, \lambda_1^o) = \nabla f(x^o) - \lambda_1^o \nabla h_1(x^o) = 0}$$

- This is just a necessary (not a sufficient) condition"
  x solution implies $\nabla h_1(x) \parallel \nabla f(x)$

Toy Example: Inequality Constraint

- Example 2: $\quad min \quad x_1 + x_2 \qquad \equiv f$

$$s.t.: \quad 2 - x_1^{\,2} - x_2^{\,2} \geqslant 0 \qquad \equiv c_1$$



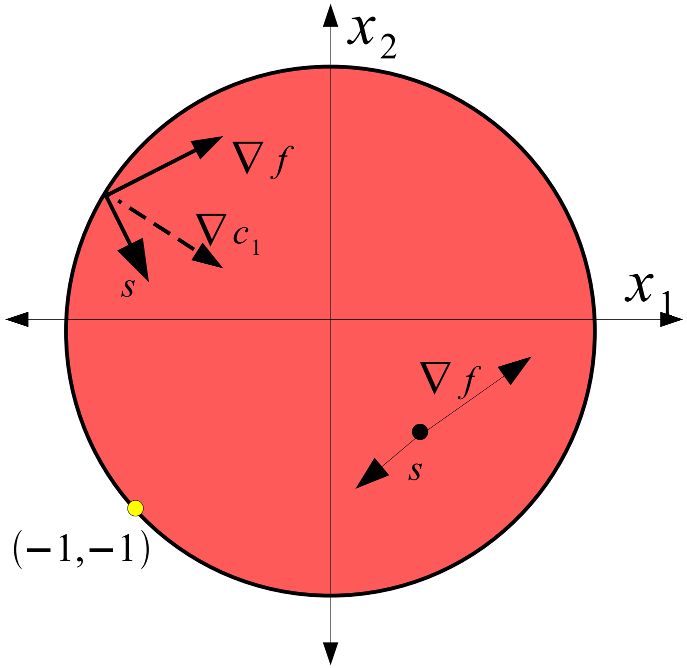
$$\nabla f = \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{pmatrix}$$

$$\nabla c_1 = \begin{pmatrix} \dfrac{\partial c_1}{\partial x_1} \\ \dfrac{\partial c_1}{\partial x_2} \end{pmatrix}$$

Toy Example: Inequality Constraint

- x is not an optimal solution, if there exists $s \neq 0$ such that

$$c_1(x+s) \geqslant 0$$
$$f(x+s) < f(x)$$



- Using first order Taylor's expansion

$$c_1(x+s) = c_1(x) + \nabla c_1(x)^T s \geqslant 0 \qquad (1)$$

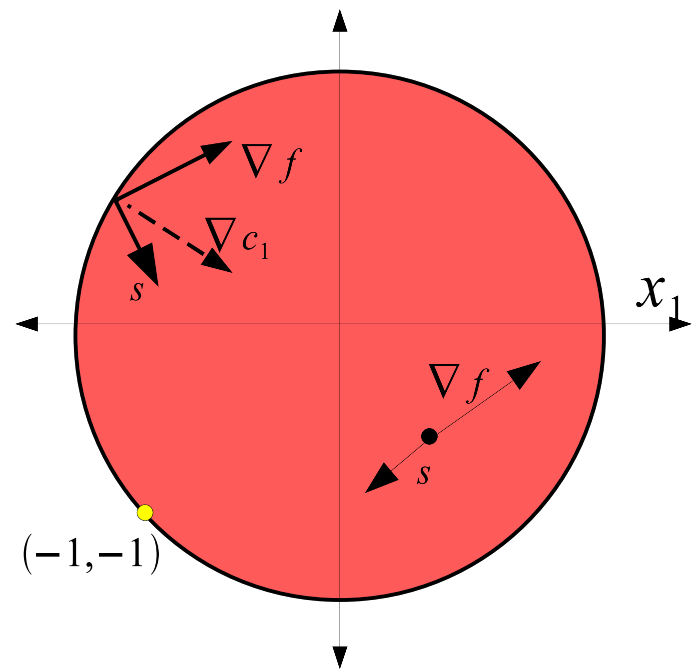
$$f(x+s) - f(x) = \nabla f(x)^T s < 0 \qquad (2)$$

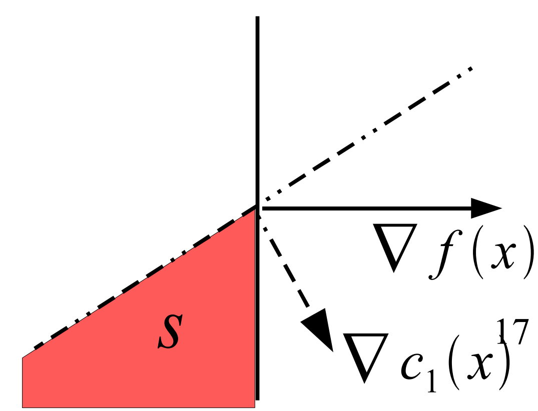
Toy Example: Inequality Constraint

- Case 1: Inactive constraint $c_1(x) > 0$
  - Any sufficiently small s as long as $\nabla f_1(x) \neq 0$
  - Thus $s = -\alpha \nabla f(x) \quad where\ \alpha > 0$
- Case 2: Active constraint $c_1(x) = 0$

$$\nabla c_1(x)^T s \geqslant 0 \qquad (1)$$
$$\nabla f(x)^T s < 0 \qquad (2)$$

$$\boxed{\nabla f(x) = \lambda_1 \nabla c_1(x), \qquad where\ \lambda_1 \geqslant 0}$$





Toy Example: Inequality Constraint

- Thus we have the Lagrangian (as before)

$$\boxed{L(x,\lambda_1) = f(x) - \lambda_1 c_1(x)}$$

Lagrange multiplier or dual variable for c_1

- The optimality conditions

$$\boxed{\nabla_x L(x^o,\lambda_1^o) = \nabla f(x^o) - \lambda_1^o \nabla c_1(x^o) = 0 \quad \textit{for some} \quad \lambda_1 \geqslant 0}$$

and

$$\boxed{\lambda_1^o c_1(x^o) = 0}$$

Complementarity condition

either $c_1(x^o) = 0$ or $\lambda_1^o = 0$

(active) (inactive)

Same Concepts in a More General Setting

Lagrange Function

- The Problem

$$\min_x f_0(x)$$ objective function

$$s.t.:$$

$$f_i(x) \leqslant 0, \quad i=1,\ldots,m$$ m inequality constraints

$$h_i(x) = 0, \quad i=1,\ldots,p$$ p equality constraints

- Standard tool for constrained optimization: the Lagrange Function

$$\boxed{L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)}$$

dual variables or Lagrange multipliers

Lagrange Dual Function

- Defined, for λ, ν as the minimum value of the Lagrange function over x

m inequality constraints
p equality constraints

$$g : \Re^m \times \Re^p \to \Re$$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_1 f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

Lagrange Dual Function

- Interpretation of Lagrange dual function:
  - Writing the original problem as unconstrained problem but with hard indicators (penalties)

$$\underset{x}{minimize} \quad \left(f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right)$$

where

$$I_0(u) = \begin{cases} \overbrace{0 \quad u \leqslant 0}^{\text{satisfied}} \\ \underbrace{\infty \quad u > 0}_{\text{unsatisfied}} \end{cases} \qquad I_1(u) = \begin{cases} \overbrace{0 \quad u = 0}^{\text{satisfied}} \\ \underbrace{\infty \quad u \neq 0}_{\text{unsatisfied}} \end{cases}$$

indicator functions

Lagrange Dual Function

- Interpretation of Lagrange dual function:
  - The Lagrange multipliers in Lagrange dual function can be seen as "softer" version of indicator (penalty) functions.

$$\underset{x}{\text{minimize}} \left(f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right)$$

$$\inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem:

$$\boxed{g(\lambda, \nu) \leqslant p^o}$$

- Proof: Let $\hat{x}$ be a feasible optimal point and let $\lambda \geqslant 0$. Then we have:

$$\begin{aligned} f_i(\hat{x}) &\leqslant 0 \qquad i=1,\ldots,m \\ h_i(\hat{x}) &= 0 \qquad i=1,\ldots,p \end{aligned}$$

Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem:

$$\boxed{g(\lambda, \nu) \leqslant p^o}$$

- Proof: Let $\hat{x}$ be a feasible optimal point and let $\lambda \geqslant 0$. Then we have:

$$\begin{aligned} f_i(\hat{x}) &\leqslant 0 \qquad i=1,\ldots,m \\ h_i(\hat{x}) &= 0 \qquad i=1,\ldots,p \end{aligned}$$

- Thus

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \sum_{i=1}^{m} \lambda_i f_i(\hat{x}) + \sum_{i=1}^{p} \nu_i h_i(\hat{x}) \leqslant f_0(\hat{x})$$

Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem:

$$\boxed{g(\lambda,\nu)\leqslant p^o}$$

- Proof: Let $\hat{x}$ be a feasible optimal point and let $\lambda\geqslant 0$. Then we have:

$$\begin{aligned} f_i(\hat{x}) &\leqslant 0 \qquad i=1,\ldots,m \\ h_i(\hat{x}) &= 0 \qquad i=1,\ldots,p \end{aligned}$$

- Thus

$$L(\hat{x},\lambda,\nu) = f_0(\hat{x})+\boxed{\overbrace{\sum_{i=1}^{m}\lambda_i f_i(\hat{x})+\sum_{i=1}^{p}\nu_i h_i(\hat{x})}^{.\ \leqslant\ 0}}\leqslant f_0(\hat{x})$$

Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem:

$$\boxed{g(\lambda, \nu) \leqslant p^o}$$

- Proof: Let $\hat{x}$ be a feasible optimal point and let $\lambda \geqslant 0$. Then we have:

$$\begin{aligned} f_i(\hat{x}) &\leqslant 0 \qquad i=1,\ldots,m \\ h_i(\hat{x}) &= 0 \qquad i=1,\ldots,p \end{aligned}$$

- Thus

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \overbrace{\boxed{\sum_{i=1}^{m} \lambda_i f_i(\hat{x}) + \sum_{i=1}^{p} \nu_i h_i(\hat{x})}}^{.\ \leqslant\ 0} \leqslant f_0(\hat{x})$$

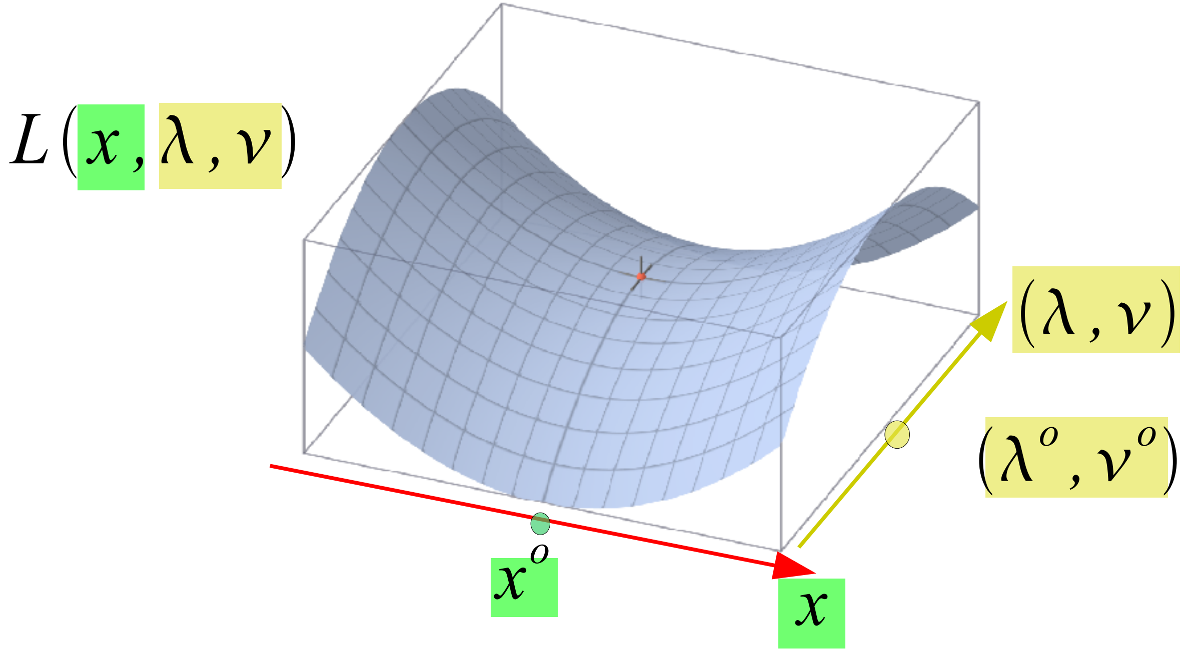
- Hence

$$\boxed{g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leqslant L(\hat{x}, \lambda, \nu) \leqslant f_0(\hat{x})}$$

Sufficient Condition

- If (x^o, λ^o, ν^o) is a saddle point, i.e. if

$$\forall\, x \in \Re^n, \quad \forall\, \lambda \geqslant 0, \quad L(x^o, \lambda, \nu) \leqslant L(x^o, \lambda^o, \nu^o) \leqslant L(x, \lambda^o, \nu^o)$$



- ... then (x^o, λ^o, ν^o) is a solution of p^o

Lagrange Dual Problem

- Lagrange dual function gives a lower bound on the optimal value of the problem.
- We seek the "best" lower bound to minimize the objective:

$$\begin{aligned} &maximize \quad g(\lambda, \nu) \\ &s.t.: \quad \lambda \geqslant 0 \end{aligned}$$

- The dual optimal value and solution:

$$d^o = g(\lambda^o, \nu^o)$$

- The Lagrange dual problem is convex even if the original problem is not.

Primal / Dual Problems

- Primal problem:

$$p^o \qquad \begin{aligned} &\min_{x\in D} f_0(x) \\ &s.t.: \\ &f_i(x) \leqslant 0, \qquad i=1,\ldots,m \\ &h_i(x)=0, \qquad i=1,\ldots,p \end{aligned}$$

- Dual problem:

$$d^o \qquad \begin{aligned} &\max_{\lambda,\nu} \quad g(\lambda,\nu) \\ &s.t.: \quad \lambda \geqslant 0 \\ &g(\lambda,\nu)=\inf_{x\in D}\left(f_0(x)+\sum_{i=1}^{m}\lambda_1 f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x)\right) \end{aligned}$$

Weak Duality

- Weak duality theorem:

$$\boxed{d^o \leqslant p^o}$$

- Optimal duality gap:

$$\boxed{p^o - d^o \geqslant 0}$$

- This bound is sometimes used to get an estimate on the optimal value of the original problem that is difficult to solve.

Strong Duality

- Strong Duality:

$$d^o = p^o$$

- Strong duality does not hold in general.

- Slater's Condition: If $x \in D$ and it is strictly feasible:

$$f_i(x) < 0 \quad for \quad i = 1, \ldots m$$
$$h_i(x) = 0 \quad for \quad i = 1, \ldots p$$

- Strong Duality theorem:
  if Slater's condition holds and the problem is convex, then strong duality is attained:

$$\exists (\lambda^o, \nu^o) \quad with \quad d^o = g(\lambda^o, \nu^o) = \max_{\lambda, \nu} g(\lambda, \nu) = \inf_x L(x, \lambda^o, \nu^o) = p^o$$

Optimality Conditions: First Order

- Complementary slackness:
  if strong duality holds, then at optimality (x^o, λ^o, ν^o)

$$\boxed{\lambda_i^o f_i(x^o) = 0 \qquad i=1,\dots m}$$

- Proof:

$$f_0(x^o) = g(\lambda^o, \nu^o) \qquad \text{(strong duality)}$$

$$= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^o f_i(x) + \sum_{i=1}^{p} \nu_i^o h_i(x) \right)$$

$$\leqslant f_0(x^o) + \sum_{i=1}^{m} \lambda_i^o f_i(x^o) + \sum_{i=1}^{p} \nu_i^o h_i(x^o)$$

$$\leqslant f_0(x^o)$$

$$\forall i, h_i(x) = 0$$

$$\forall i, f_i(x) \leqslant 0, \lambda_i \geq 0$$

Optimality Conditions: First Order

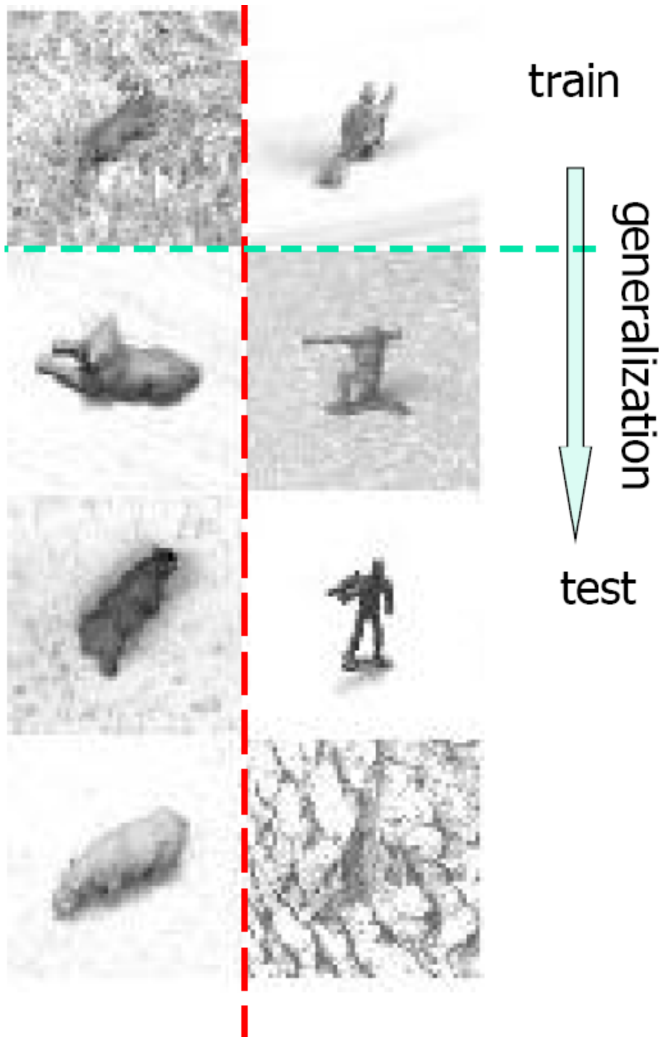
- Karush-Kuhn-Tucker (KKT) conditions
  If the strong duality holds, then at optimality:

$$
\begin{aligned}
f_i(x^o) &\leqslant 0, \quad i=1,\dots,m \\
h_i(x^o) &= 0, \quad i=1,\dots,p \\
\lambda_i^o &\geqslant 0, \quad i=1,\dots,m \\
\lambda_i^o f_i(x^o) &= 0, \quad i=1,\dots,m \\
\nabla f_0(x^o)+\sum_{i=1}^{m} \lambda_i^o \nabla f_i(x^o)+\sum_{i=1}^{p} \nu_i^o \nabla h_i(x^o) &= 0
\end{aligned}
$$

- KKT conditions are
  - necessary in general (local optimum)
  - necessary and sufficient in case of convex problems (global optimum)

What are Support Vector Machines?

- Linear classifiers
- (Mostly) binary classifiers
- Supervised training
- Good generalization with explicit bounds

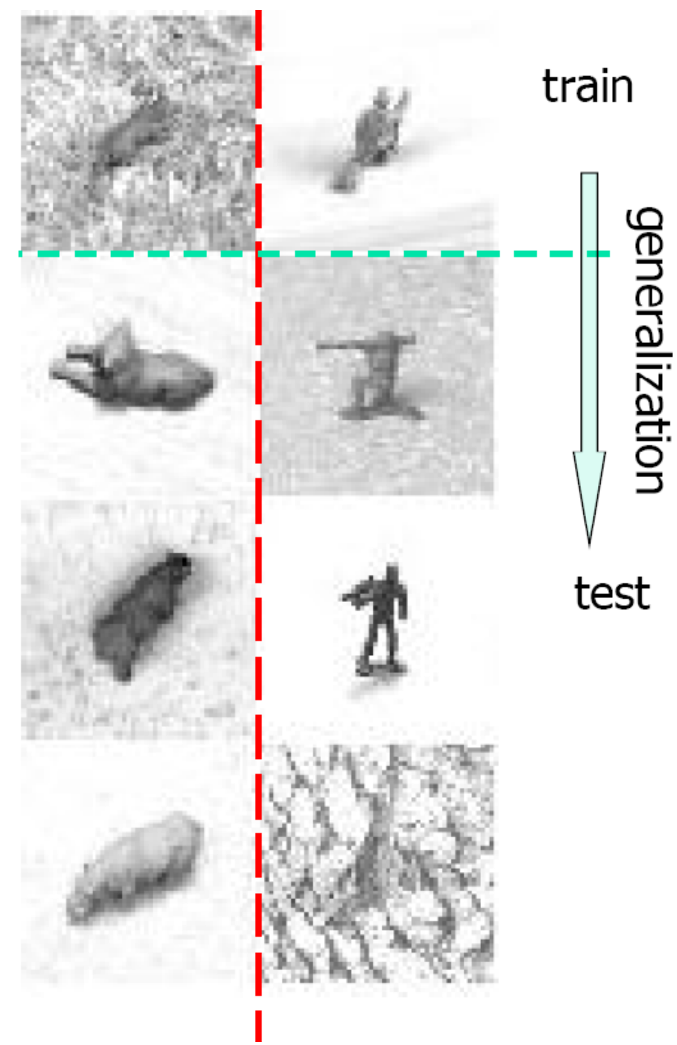


Main Ideas Behind Support Vector Machines

- Maximal margin
- Dual space
- Linear classifiers

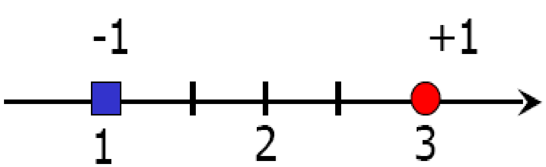
  in high-dimensional space

  using non-linear mapping
- Kernel trick



Quadratic Programming

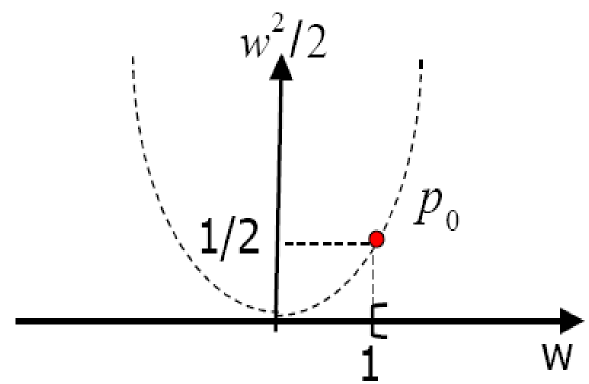
$$\max_{\boldsymbol{w},b}\ \min_{i} \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} \xrightarrow{\min_i |\boldsymbol{w}^T \boldsymbol{x}_i + b| = 1} \begin{array}{c} \min_{w,b} \frac{1}{2}\langle \boldsymbol{w}^T \cdot \boldsymbol{w} \rangle \\ y_i(\langle \boldsymbol{w}^T \cdot \boldsymbol{x}_i \rangle + b) \geq 1 \end{array}$$



$$\min_{w} \frac{w^2}{2}$$

$$(+1)(w \cdot 3 + b) \geqslant 1$$

$$(-1)(w \cdot 1 + b) \geqslant 1$$



Using the Lagrangian

- Combine target and constraints
- Minimize over primal
- Maximize over dual

$$L(x,\boldsymbol{\lambda}) = f_0(x) - \sum \lambda_i f_i(x)$$

$$Q(\boldsymbol{\lambda}) = \min_x L(x,\boldsymbol{\lambda})$$

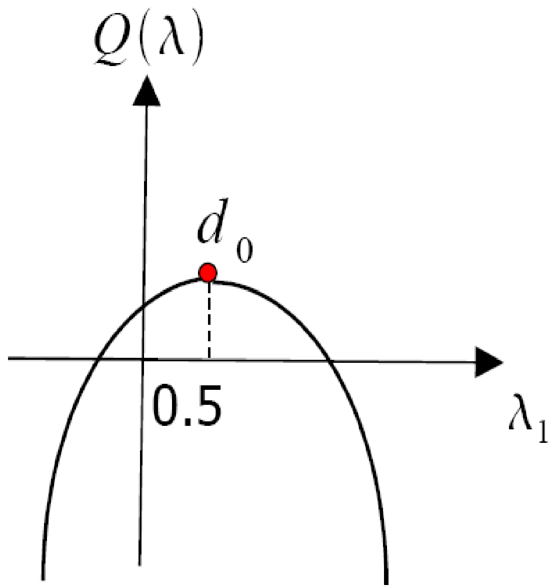
$$\max_\lambda Q(\boldsymbol{\lambda}), \lambda > 0$$

Dual Space

$$\min_{w} \frac{w^2}{2}$$

$$(+1)(w \cdot 3 + b) \geqslant 1$$

$$(-1)(w \cdot 1 + b) \geqslant 1$$



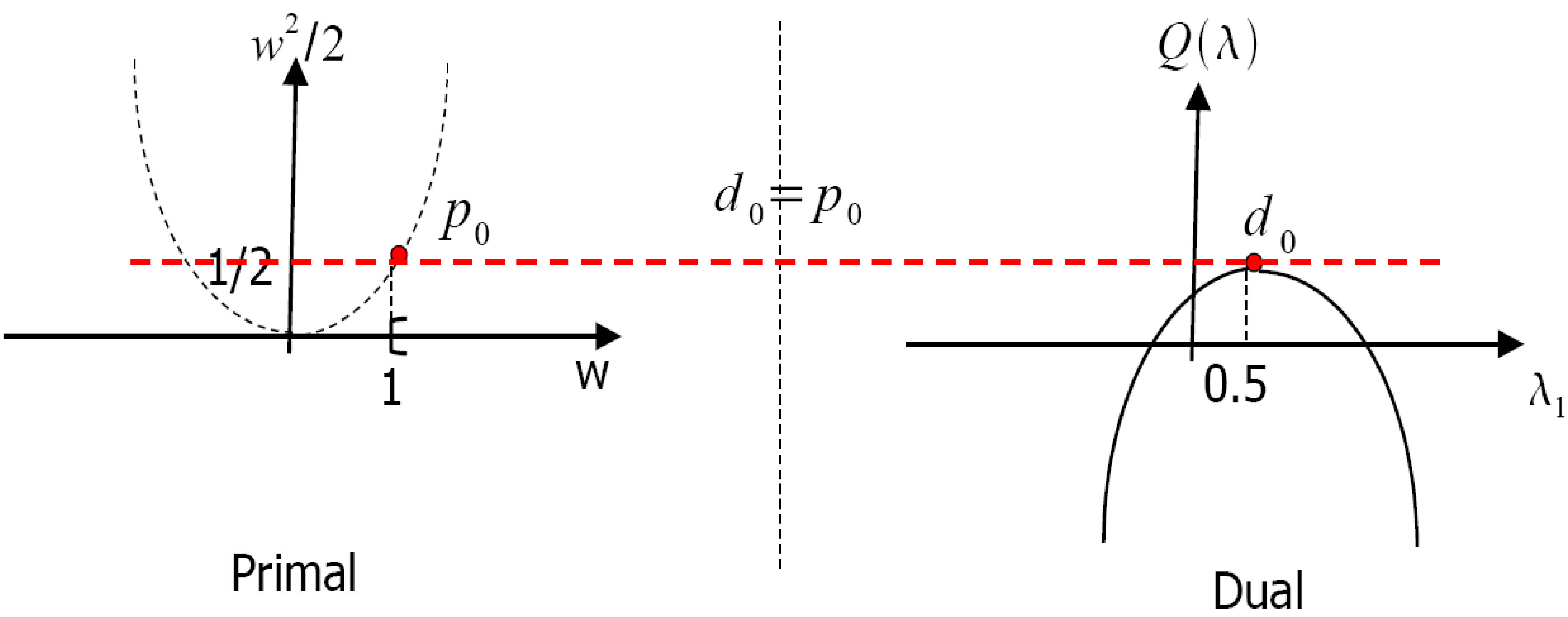
$$L(w, b, \boldsymbol{\lambda}) = w^2/2 - \lambda_1(3\text{w} + b - 1) - \lambda_2(-w - b - 1)$$

$$\min_{w,b} L(w, b, \lambda) \Rightarrow \begin{cases} \lambda_1 = \lambda_2 \\ w = 3\lambda_1 - \lambda_2 = 2\lambda_1 \\ Q(\boldsymbol{\lambda}) = Q(\lambda_1) = -2\lambda_1^2 + 2\lambda_1 \end{cases}$$

$$\max_{\lambda} Q(\lambda) \Rightarrow \lambda_1 = \lambda_2 = 1/2, w = 1, b = 2$$

Strong Duality

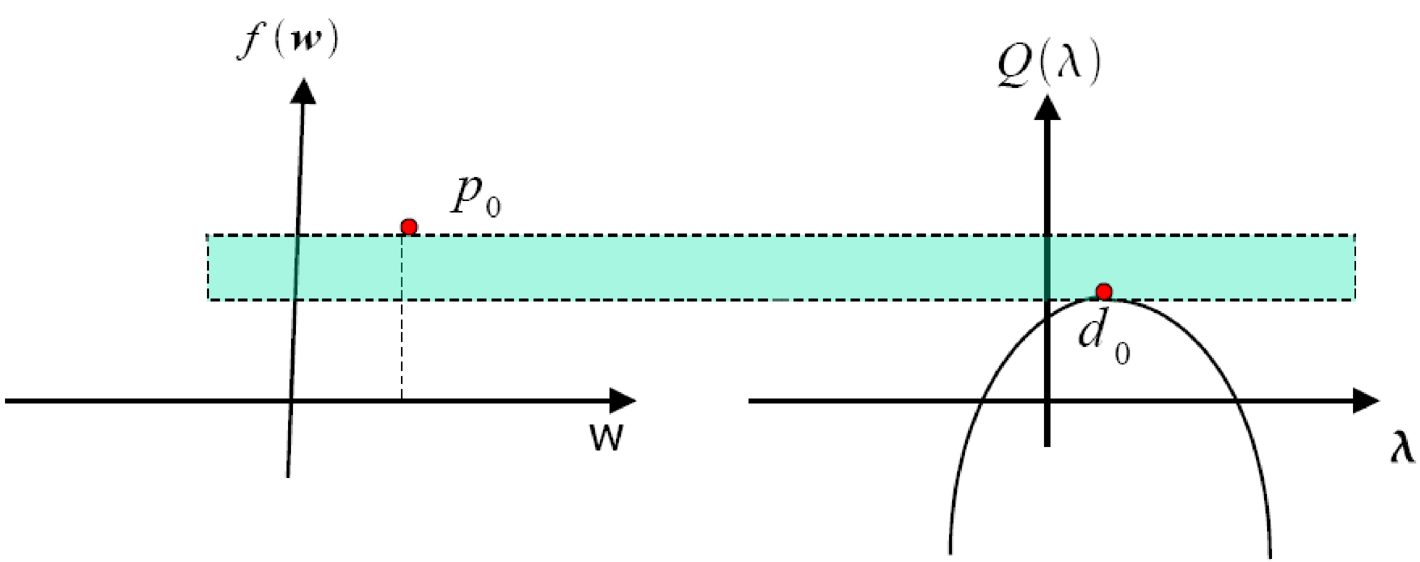
- Primal and dual space optimization:
  - Same result!



Duality Gap

$d_0 < p_0$

- In a general case
  - Strong duality is not true
  - "Step 1-2-3" a lower bound, not a solution



No Duality Gap Thanks to Convexity

- Convex function
  - Quadratic programming
- Convex set
  - Linear constraints
- No duality gap

$$\min_{w,b} \frac{1}{2}\langle \boldsymbol{w}^T \cdot \boldsymbol{w} \rangle$$

$$y_i(\langle \boldsymbol{w}^T \cdot \boldsymbol{x}_i \rangle + b) \geq 1$$

$$d_0 = p_0$$

Dual Form

- H
  - Hessian matrix
  - Gram matrix
- Lambda
  - Support vector
  - Sparse

$$\max_{\lambda} Q(\boldsymbol{\lambda}) = -0.5\boldsymbol{\lambda}^T \boldsymbol{H}\boldsymbol{\lambda} + \boldsymbol{f}^T\boldsymbol{\lambda}$$

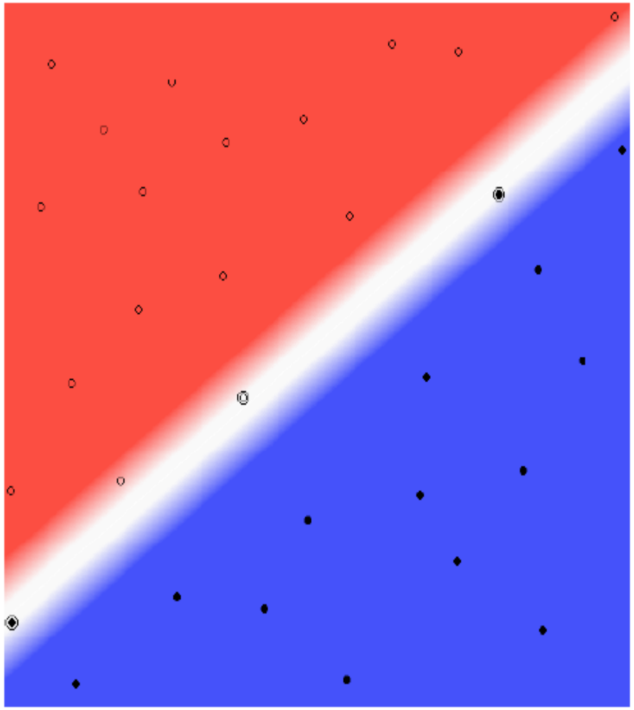
$$\boldsymbol{y}^T\boldsymbol{\lambda} = 0$$

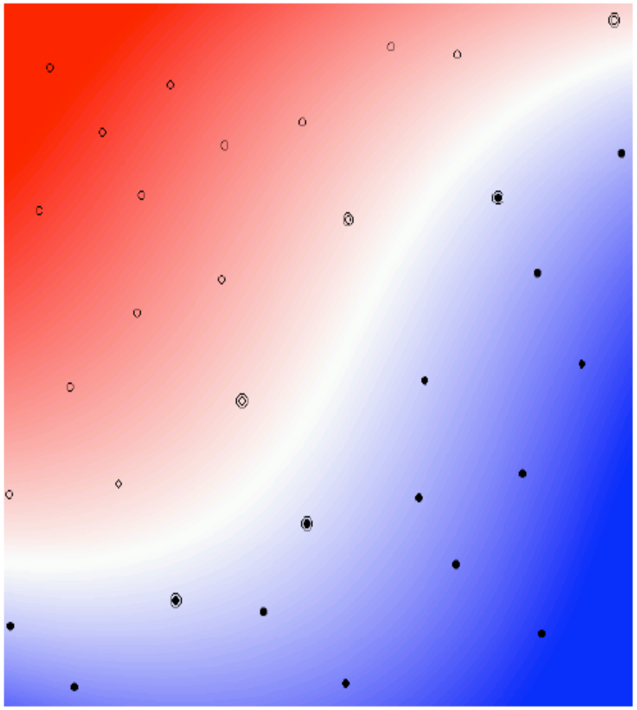
$$\boldsymbol{\lambda} \geq 0$$

$$where,\ H_{ij} = y_i y_j \langle \boldsymbol{x}_i^T \cdot \boldsymbol{x}_j \rangle$$

$$\boldsymbol{f}\ is\ a\ unit\ vector$$

Non-linear separation of datasets

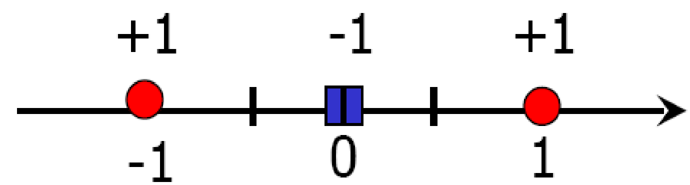




- Non-linear separation is **impossible** in **most problems**

Illustration from Prof. Mohri's lecture notes

Non-separable datasets

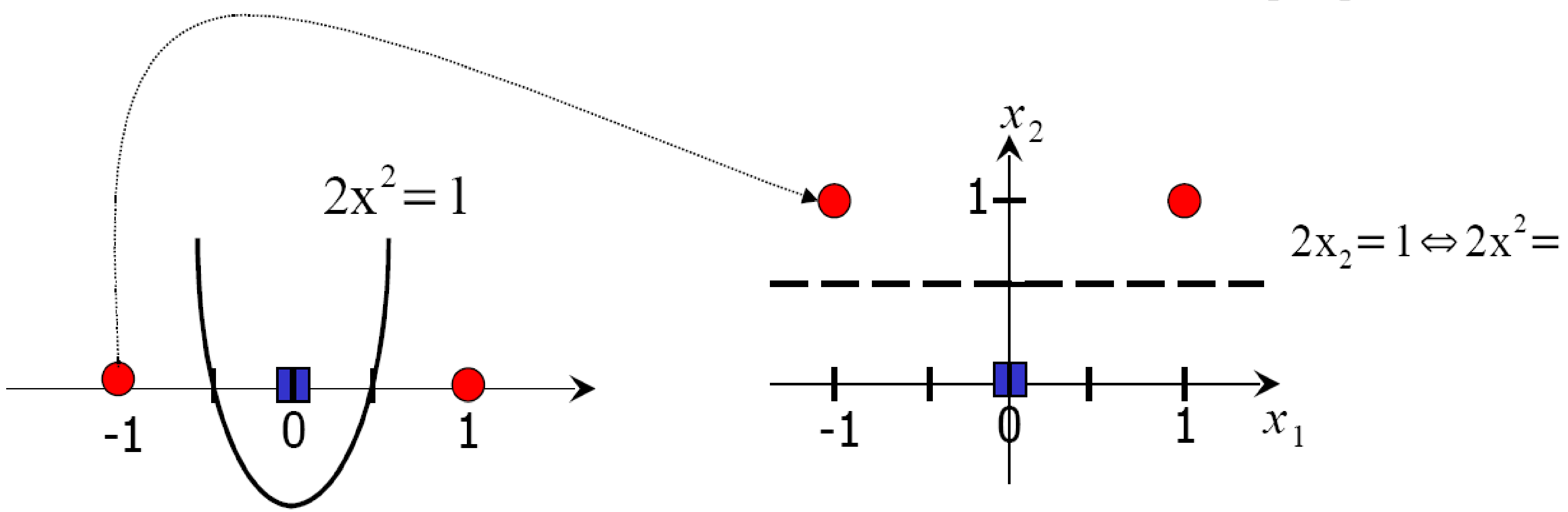


- Solutions:

1) Nonlinear classifiers

2) **Increase dimensionality** of dataset and add a **non-linear mapping Φ**

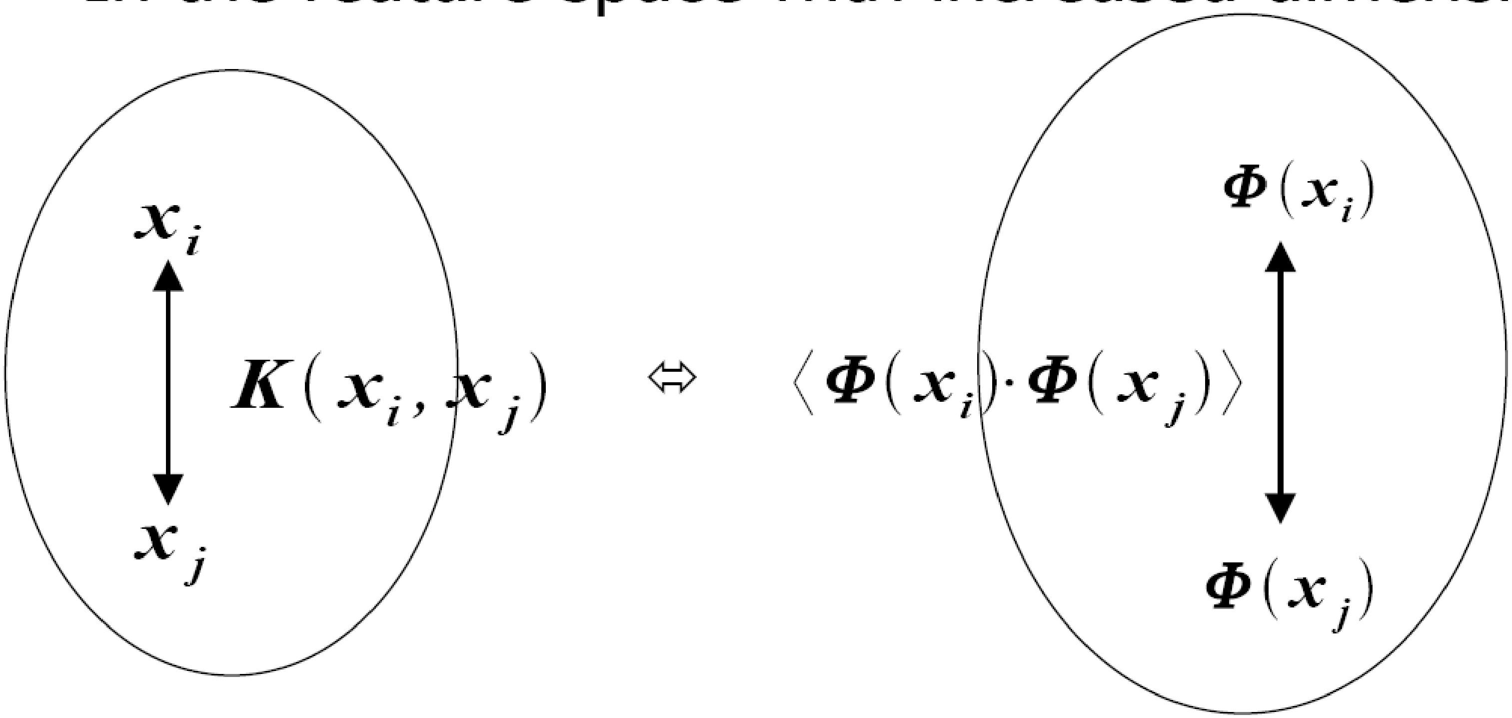
$$\begin{bmatrix} x \end{bmatrix} \rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$



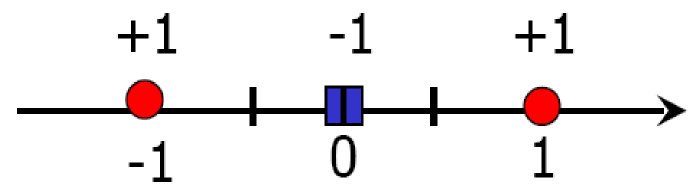
Kernel Trick

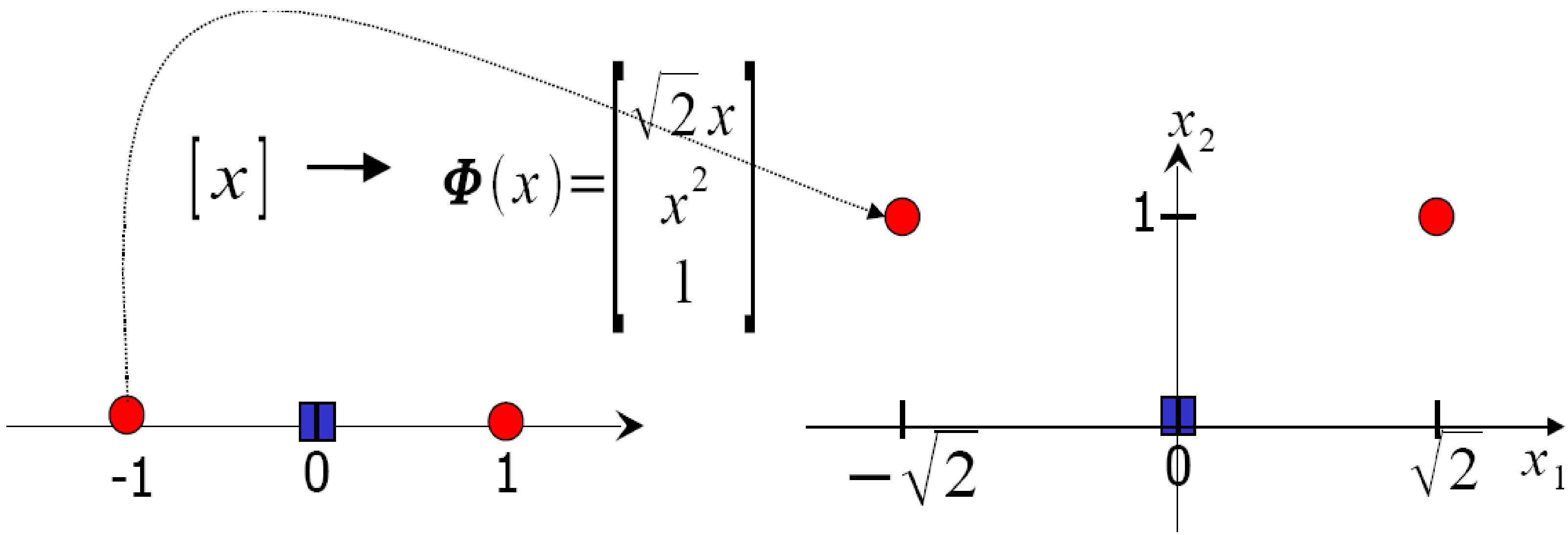
- Kernel function
  - in the original space
- Inner product
  - In the feature space with increased dimension

"similarity measure"
between 2 data samples



Kernel Trick Illustrated





$$\boldsymbol{K}(x_i, x_j) = (x_i x_j + 1)^2$$

$$\langle \boldsymbol{\Phi}(x_i) \cdot \boldsymbol{\Phi}(x_j) \rangle = 2\mathrm{x}_i x_j + x_i^2 x_j^2 + 1 = (x_i x_j + 1)^2 = \boldsymbol{K}(x_i, x_j)$$

Curse of Dimensionality Due to the Non-Linear Mapping

- Primal space
  - Makes optimization much harder
- Dual space
  - Can be avoided

$$\min_{w,b} \frac{1}{2}\langle \boldsymbol{\Phi}^T(\boldsymbol{w})\cdot\boldsymbol{\Phi}(\boldsymbol{w})\rangle$$

$$y_i(\langle \boldsymbol{\Phi}^T(\boldsymbol{w})\cdot\boldsymbol{\Phi}(\boldsymbol{x}_i)\rangle + b) \geq 1$$

$$\max_{\lambda} Q(\lambda) = -0.5\,\lambda^T \boldsymbol{H}\lambda + \boldsymbol{f}^T\lambda$$

$$\boldsymbol{y}^T\lambda = 0$$

$$\lambda \geq 0$$

$$where,\ H_{ij} = y_i y_j \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\boldsymbol{f}\ is\ a\ unit\ vector$$

Positive Symmetric Definite Kernels (Mercer Condition)

- Dual form is convex
  - H is P.S.D.
  - Kernel must be P.S.D.

$$Q(\lambda) = -0.5\,\lambda^T \boldsymbol{H}\,\lambda + \boldsymbol{f}^T \lambda$$
$$where,\ H_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Mercer kernels
  - Polynomial
  - Gaussian

$$K(\boldsymbol{x}, \boldsymbol{y}) = [\langle \boldsymbol{x}^T \boldsymbol{y} \rangle + 1]^p$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{e}^{-(\boldsymbol{x}-\boldsymbol{y})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{y})/2}$$

Advantages of SVM

- Work very well...
- Error bounds easy to obtain:
  - Generalization error **small** and **predictable**

$$E_{test} = E_{train} + E_{generalization} \longleftarrow \frac{|SV|}{N}$$

- Fool-proof method:
  - (Mostly) three kernels to choose from:
    - Gaussian
    - Linear and Polynomial
    - Sigmoid
  - Very small number of parameters to optimize

Limitations of SVM

- Size limitation:
  - Size of kernel matrix is quadratic with the number of training vectors
- Speed limitations:
  - 1) During training:
    very large quadratic programming problem solved numerically
    - Solutions:
      - **Chunking**
      - **Sequential Minimal Optimization** (SMO) breaks QP problem into many small QP problems solved analytically
      - **Hardware** implementations
  - 2) During testing:
    number of support vectors
    - Solution: **Online SVM**