Computational learning and discovery

GEORGE MASON UNIVERSITY

CSI 873 / MATH 689

Instructor: I. Griva

Wednesday 7:20 - 10 pm

Support vector machine

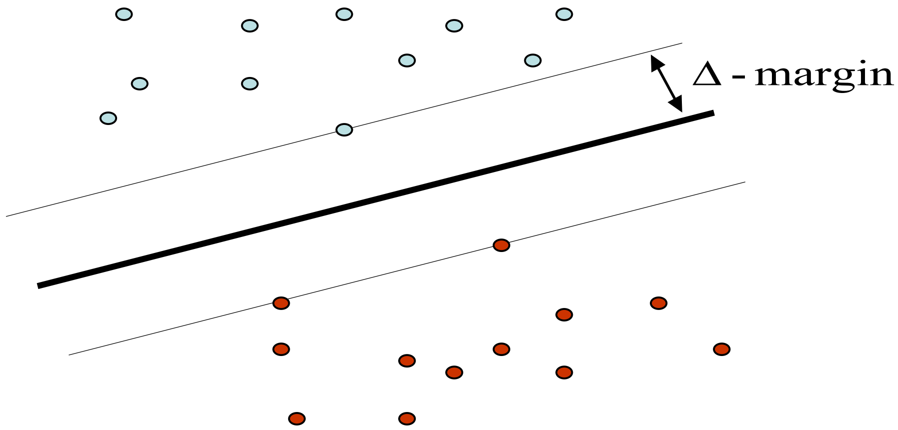
Given a set of training data

$$(x_1, y_1), \ldots, (x_l, y_l),\ x_i \in \Re^n,\ y_i \in \{+1, -1\}$$

find a function that can estimate

$$y_j^* \in \{+1, -1\} \text{ given new } x_j^* \in \Re^n$$

and minimize the frequency of the future error.



Support vector machine

based on fundamentals of statistical learning theory
(Vapnik-Chervonenkis theory)

$$X \longrightarrow \boxed{Y=f(X)} \xrightarrow{Y}$$

Instead of identifying the unknown function (what classical statistics does), the main goal of VC theory is to imitate the unknown function.

The key discovery of VC theory:

- Two and only two factors are responsible for generalization:
  - -One (empirical loss) defines how well the function approximates data
  - -Another (capacity, VC dimension) defines the diversity of the set of functions from which one chooses an approximation function

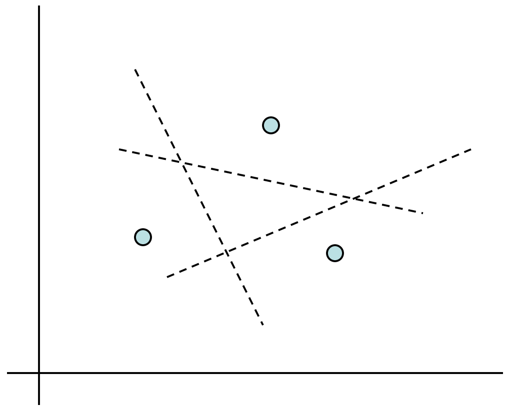
- If VC dimension is finite, then one can achieve a good generalization. If it is not finite the generalization is impossible.

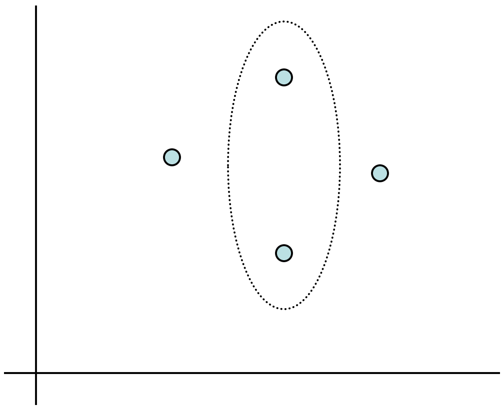
Examples

The VC dimension of linear indicator functions

$$I(x) = \mathrm{sgn}(x^T w + b), \quad x \in \Re^n, w \in \Re^n$$

is equal $n+1$



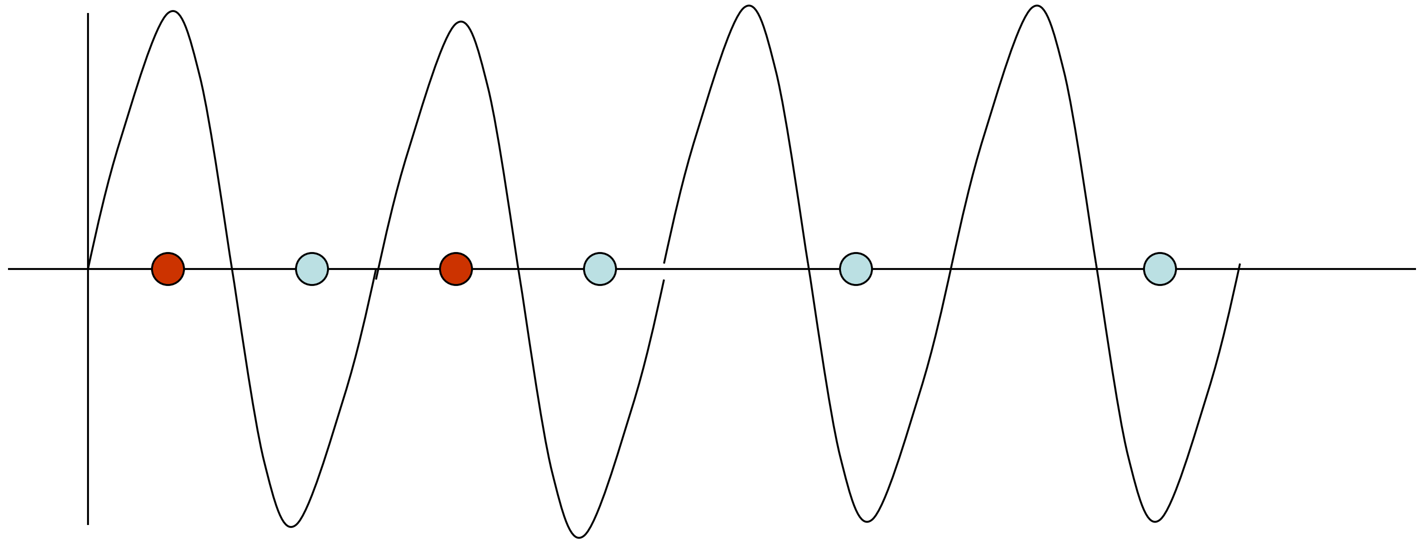


Examples

The VC dimension of the set of functions

$$I(x) = \operatorname{sgn}(\sin ax), \quad x \in \Re^1, w \in \Re^1$$

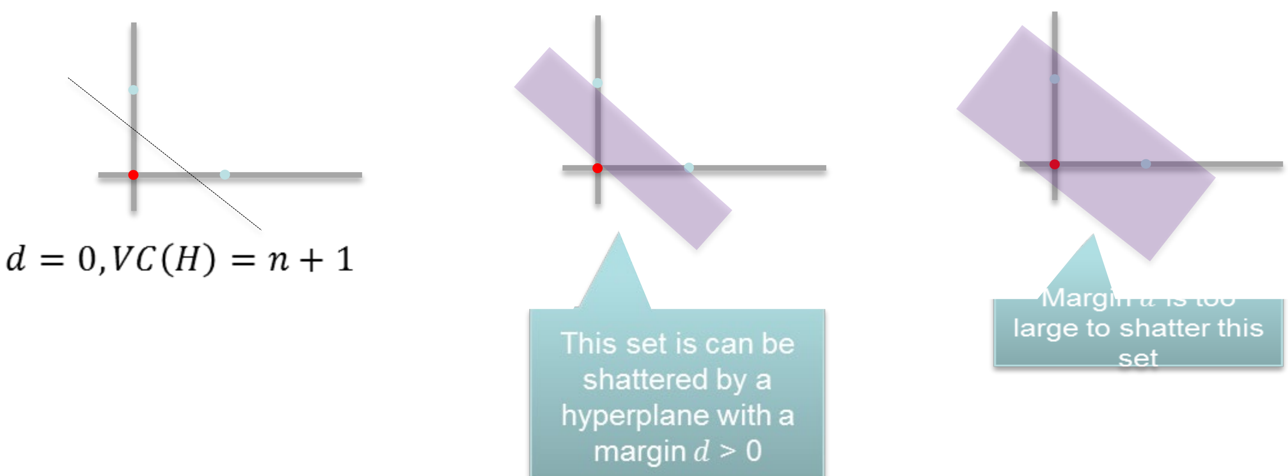
is infinity



Support vector machine

Let the vector $x \in \Re^n$ belong to a sphere of radius R. Then the set of Δ - margin separating hyperplaines has a VC dimention bounded as follows

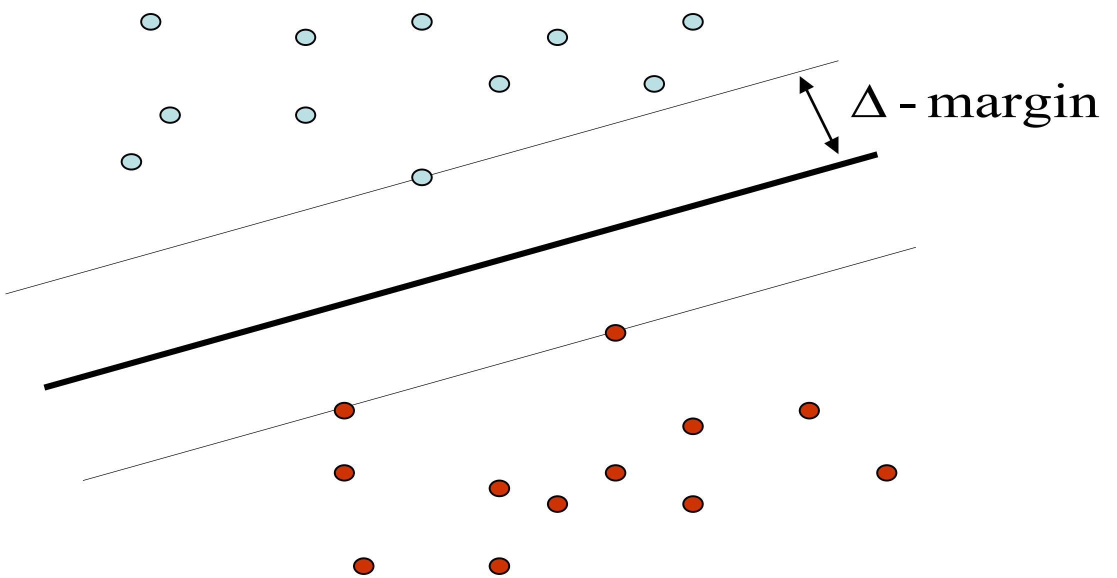
$$VC_{\dim} \leq \min\left\{\frac{R^2}{\Delta^2}, n\right\} + 1$$



Support vector machine

Let the vector $x \in \Re^n$ belong to a sphere of radius R. Then the set of Δ - margin separating hyperplaines has a VC dimention bounded as follows

$$VC_{\dim} \leq \min\left\{\frac{R^2}{\Delta^2}, n\right\} + 1$$



Support vector machine

Theorem. With probability $1-\delta$ one can assert that the probability that a test example will not be separated correctly by the Δ - margin hyperplane has the bound

$$P_{error} \le \frac{m}{l} + \frac{\kappa}{2}\left(1 + \sqrt{1 + \frac{4m}{l\kappa}}\right),$$

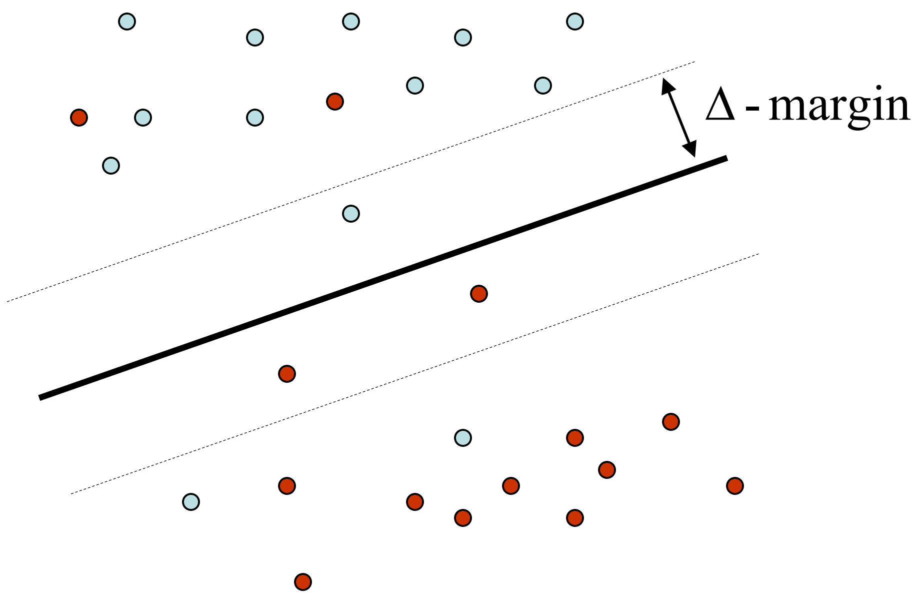
where

$$\kappa = 4\frac{VC_{\dim}(\ln\frac{2l}{VC_{\dim}} + 1) - \ln\frac{\delta}{4}}{l},$$

m is the number of training examples that are not separated correctly by the Δ - margin hyperplane and the VC dimention bounded as follows

$$VC_{\dim} \le \min\left\{\frac{R^2}{\Delta^2}, n\right\} + 1$$

Support vector machine



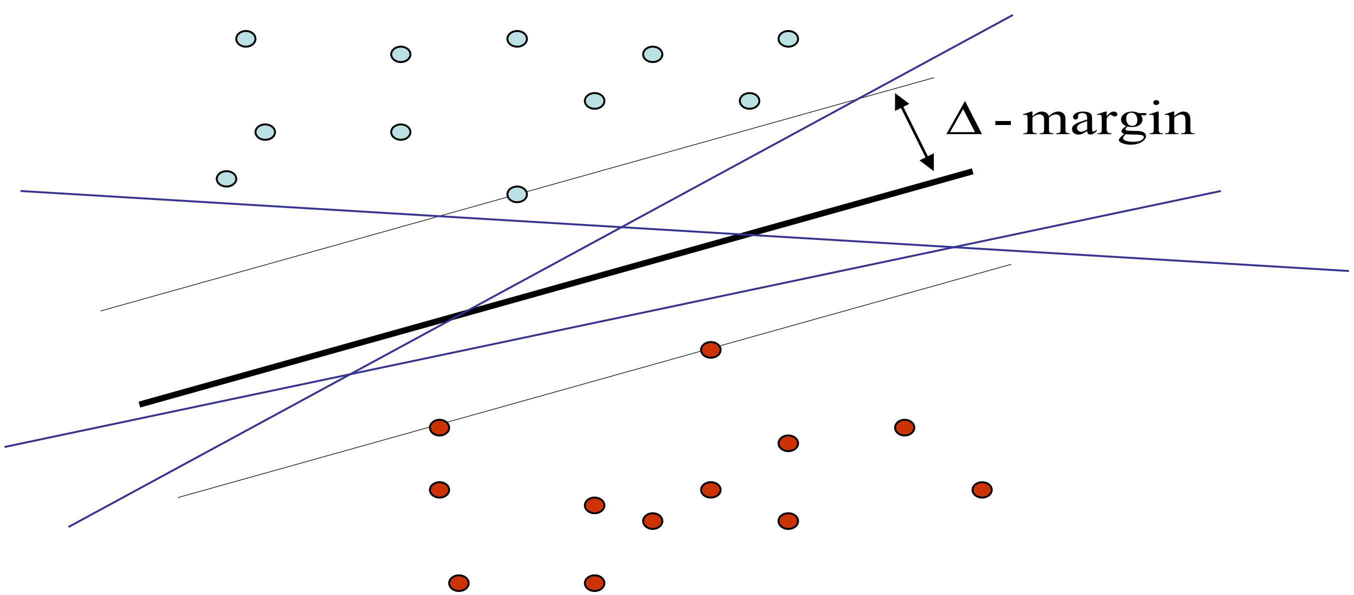
Support vector machine

Suppose that the data

$$(y_1, x_1), \ldots, (y_l, x_l) \qquad x \in \Re^n \qquad y \in \{+1; -1\}$$

can be separated by a hyperplane

$$(w \cdot x) - b = 0$$

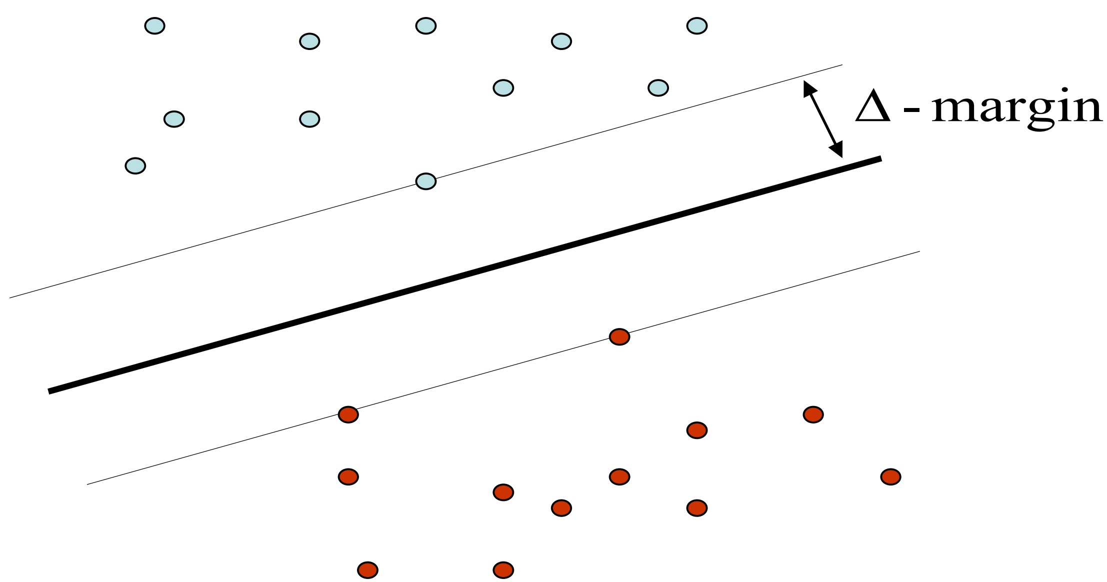


Support vector machine

Blue dots: $(w \cdot x_i) - b \geq 0, \quad y_i = +1$

Red dots: $(w \cdot x_i) - b \leq 0, \quad y_i = -1$

Combined: $y_i \left[(w \cdot x_i) - b \right] \geq 0, \quad \forall i$ Variables: w and b



$(w \cdot x) - b = 0$

$\|w\| = 1$

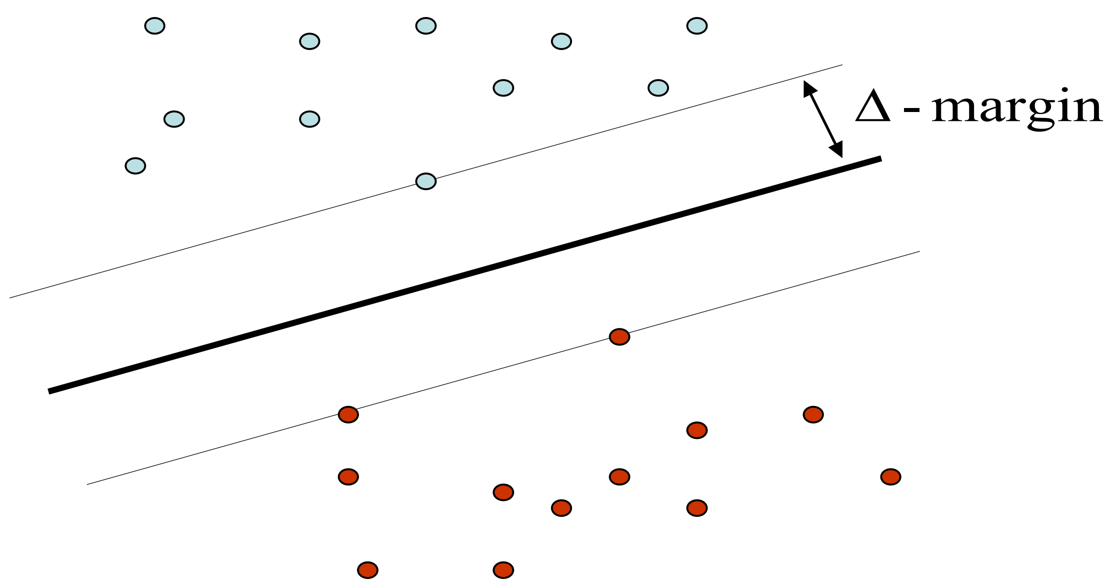
Support vector machine

Blue dots: $(w \cdot x_i) - b \geq +\Delta, \quad y_i = +1$

$$\Delta \geq 0$$

Red dots: $(w \cdot x_i) - b \leq -\Delta, \quad y_i = -1$

Combined: $y_i\left[(w \cdot x_i) - b\right] \geq \Delta, \quad \forall i$ Variables: w and b



$$(w \cdot x) - b = 0$$

$$\|w\| = 1$$

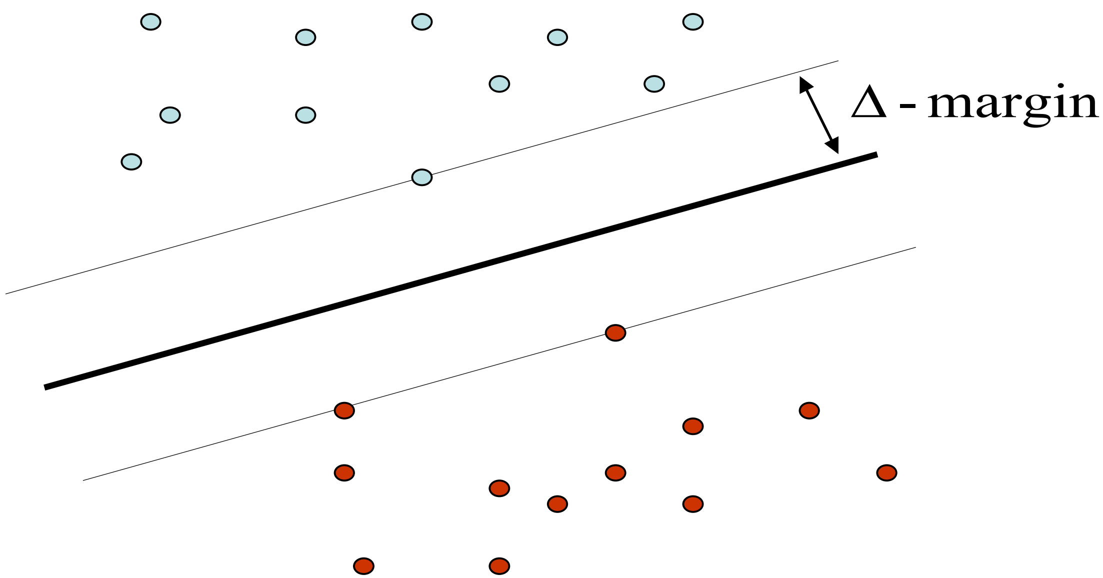
Support vector machine

Maximize the margin: $\max \Delta$

s.t. $$y_i\left[(w \cdot x_i) - b\right] \geq \Delta, \quad \forall i$$

$$w_1^2 + \cdots + w_n^2 = 1, \quad \Delta \geq 0$$

Variables: w, Δ and b



$$(w \cdot x) - b = 0$$

$$\|w\| = 1$$

Support vector machine

$$y_i\left[(\frac{w}{\Delta}\cdot x_i)-\frac{b}{\Delta}\right]\geq 1, \quad \forall i \qquad \text{or:} \qquad y_i\left[(\overline{w}\cdot x_i)-\overline{b}\right]\geq 1, \quad \forall i$$

$$\overline{w}=w/\Delta, \quad \left\|\overline{w}\right\|=1/\Delta$$

Maximize the margin: $\min \left\|\overline{w}\right\|$

s.t. $y_i\left[(\overline{w}\cdot x_i)-\overline{b}\right]\geq 1, \quad \forall i$

Variables: $\overline{w}$ and $\overline{b}$

Support vector machine

Maximize the margin: $\min \|w\|^2$

s.t. $y_i\left[(w \cdot x_i) - b\right] \geq 1, \quad \forall i$

Variables: w and b

Support vector machine

Maximize the margin: $\min\ (w \cdot w)$

s.t. $y_i\left[(w \cdot x_i) - b\right] \geq 1, \quad \forall i$

Variables: w and b

Support vector machine

Maximize the margin: $\min\ 0.5(w \cdot w)$

s.t. $y_i\left[(w \cdot x_i) - b\right] \geq 1, \quad \forall i$

Variables: w and b

Non separable case:

Maximize the margin: $\min\ 0.5(w \cdot w) + C\left(\sum_{i=1}^{l} \xi_i\right)$

s.t. $y_i\left[(w \cdot x_i) - b\right] \geq 1 - \xi_i, \quad \forall i$

$\xi_i \geq 0$

Variables: w, b and ξ

Support vector machine

$(y_1, x_1), \ldots, (y_l, x_l)$ $\qquad x \in \Re^n$

Primal problem

$$\min 0.5(w \cdot w) + C\sum_{i=1}^{l} \xi_i$$

$$\text{s.t.} \quad \xi_i \geq 0$$

Variables: w, b and ξ

$$y_i\left[(x_i \cdot w) - b\right] \geq 1 - \xi_i,\ i = 1, \ldots, l$$

Dual problem

$$\max \sum_{i=1}^{l} \alpha_i - 0.5\sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

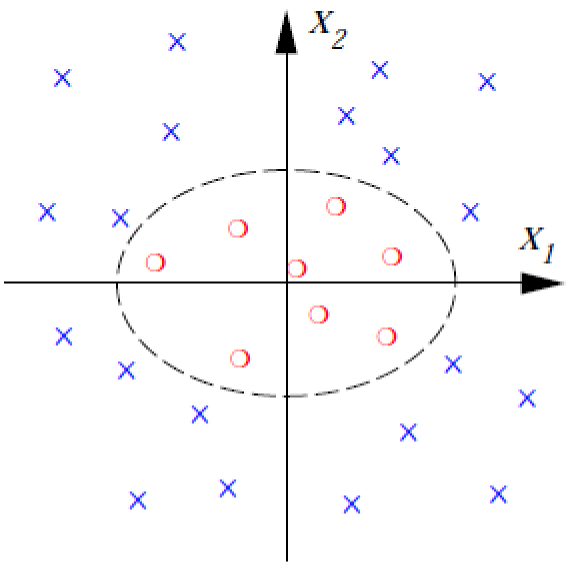
s.t.

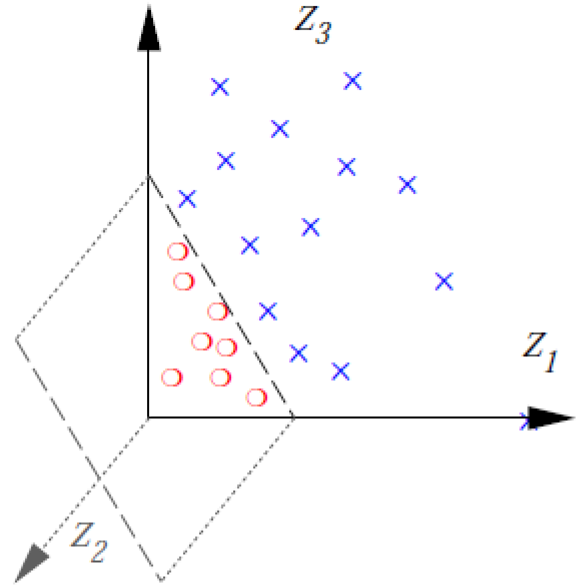
Variables: α

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, l$$

Kernels

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$





$$\begin{aligned}
\langle \Phi(x), \Phi(x') \rangle &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)(x_1'^2, \sqrt{2}\, x_1' x_2', x_2'^2)^\top \\
&= \langle x, x' \rangle^2 \\
&= : k(x, x')
\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed in $\mathbb{R}^2$

Support vector machine

Optimization problem for finding support vectors

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - 0.5 \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

s.t.

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \; i = 1,2,\ldots,l$$

Kernels

Polynomial machine:

$$K(x_i, x_j) = \left[\alpha(x_i \cdot x_j) + \beta\right]^d$$

A radial basis function machine:

$$K(x_i, x_j) = \exp\left\{-\gamma \left\| x_i - x_j \right\|^2\right\}$$

Support vector machine

Optimization problem for finding support vectors

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - 0.5 \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

s.t.

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \; i = 1,2,\ldots,l$$

Decision rules with a kernel using found α_i^*

$$\text{if } \left(\sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x)\right) - b \ge 0 \text{ then } x \text{ is } \mathit{blue}$$

$$\text{if } \left(\sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x)\right) - b < 0 \text{ then } x \text{ is } \mathit{red}$$

$$b = \left(\sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_{i_0})\right) - y_{i_0} \; \text{ for some } \alpha_{i_0}^* : \; 0 < \alpha_{i_0}^* < C, \; (\alpha_{i_0}^* \ne 0, \; \alpha_{i_0}^* \ne C)$$

$$\text{if there is no such } \alpha_{i_0}^*, \text{ increase } C \text{ and train again}$$

Support vector machine

x_i that correspond to positive α_i^* are called the support vectors!!!

Only the support vectors carry inportant information!!!

They correspond to the active constraints of the primal problem!!!

Let $I^* = \{i : \alpha_i > 0\}$ be the set of support vectors

Decision rules using only the support vectors

$$\text{if } \left(\sum_{i\in I^*} y_i \alpha_i^* K(x_i, x)\right) - b \geq 0 \text{ then } x \text{ is } \textit{blue}$$

$$\text{if } \left(\sum_{i\in I^*} y_i \alpha_i^* K(x_i, x)\right) - b < 0 \text{ then } x \text{ is } \textit{red}$$

$$b = \left(\sum_{i\in I^*} y_i \alpha_i^* K(x_i, x_{i_0})\right) - y_{i_0} \text{ for some } \alpha_{i_0}^* : 0 < \alpha_{i_0}^* < C, \ (\alpha_{i_0}^* \neq 0, \ \alpha_{i_0}^* \neq C)$$

if there is no such $\alpha_{i_0}^*$, increase C and train again

Support vector machine

Optimization problem for finding support vectors

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - 0.5 \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

s.t.

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \;\; i = 1,2,\ldots,l$$

Matlab QP setting

$$\min_{\alpha} 0.5 \alpha^T M \alpha - e^T \alpha$$

s.t.

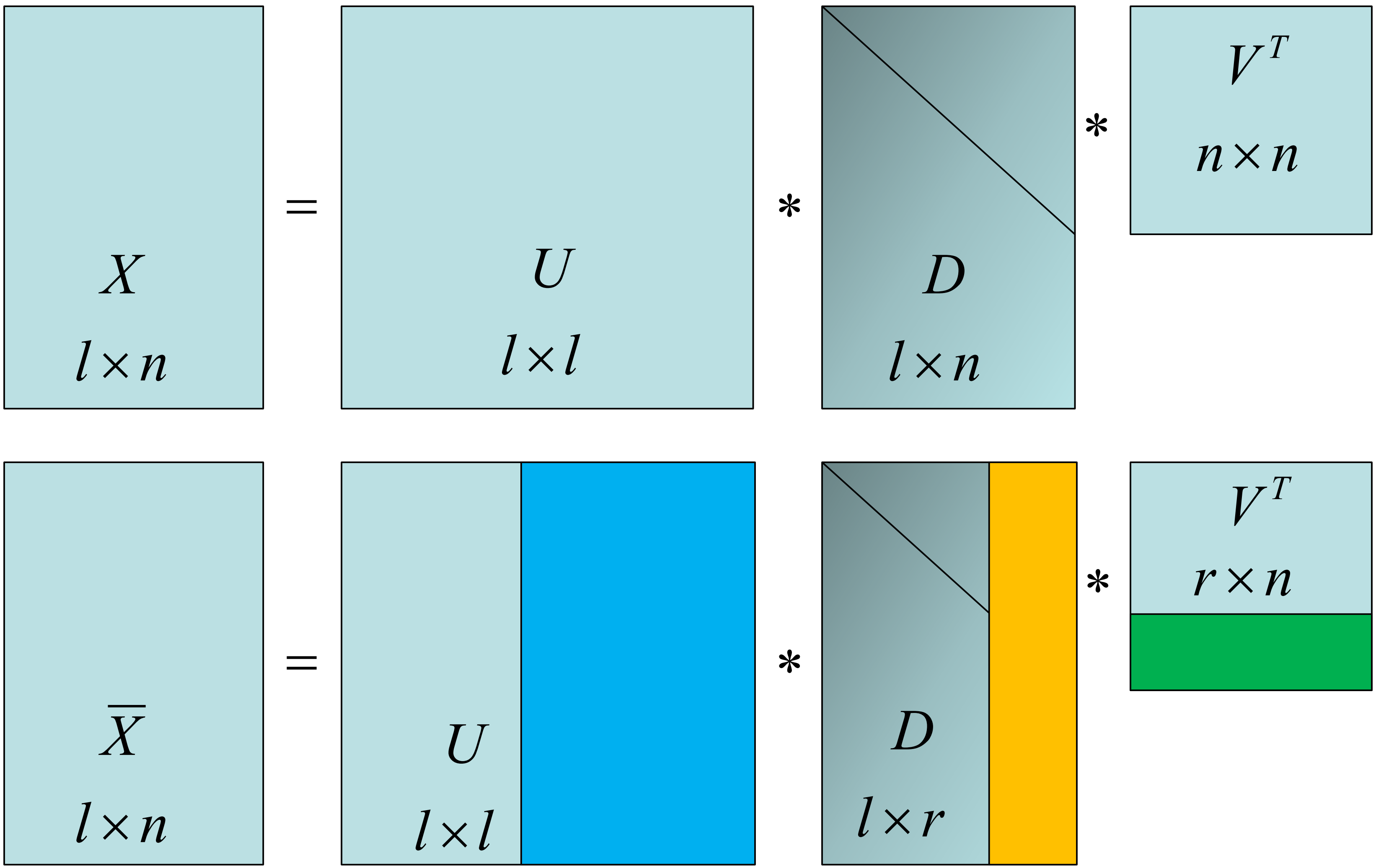
$$y^T \alpha = 0, \; 0 \le \alpha \le Ce$$

where $M_{ij} = y_i y_j K(x_i, x_j)$,

$$y = (y_1, \ldots, y_l)^T, e = (1, \ldots, 1)^T$$

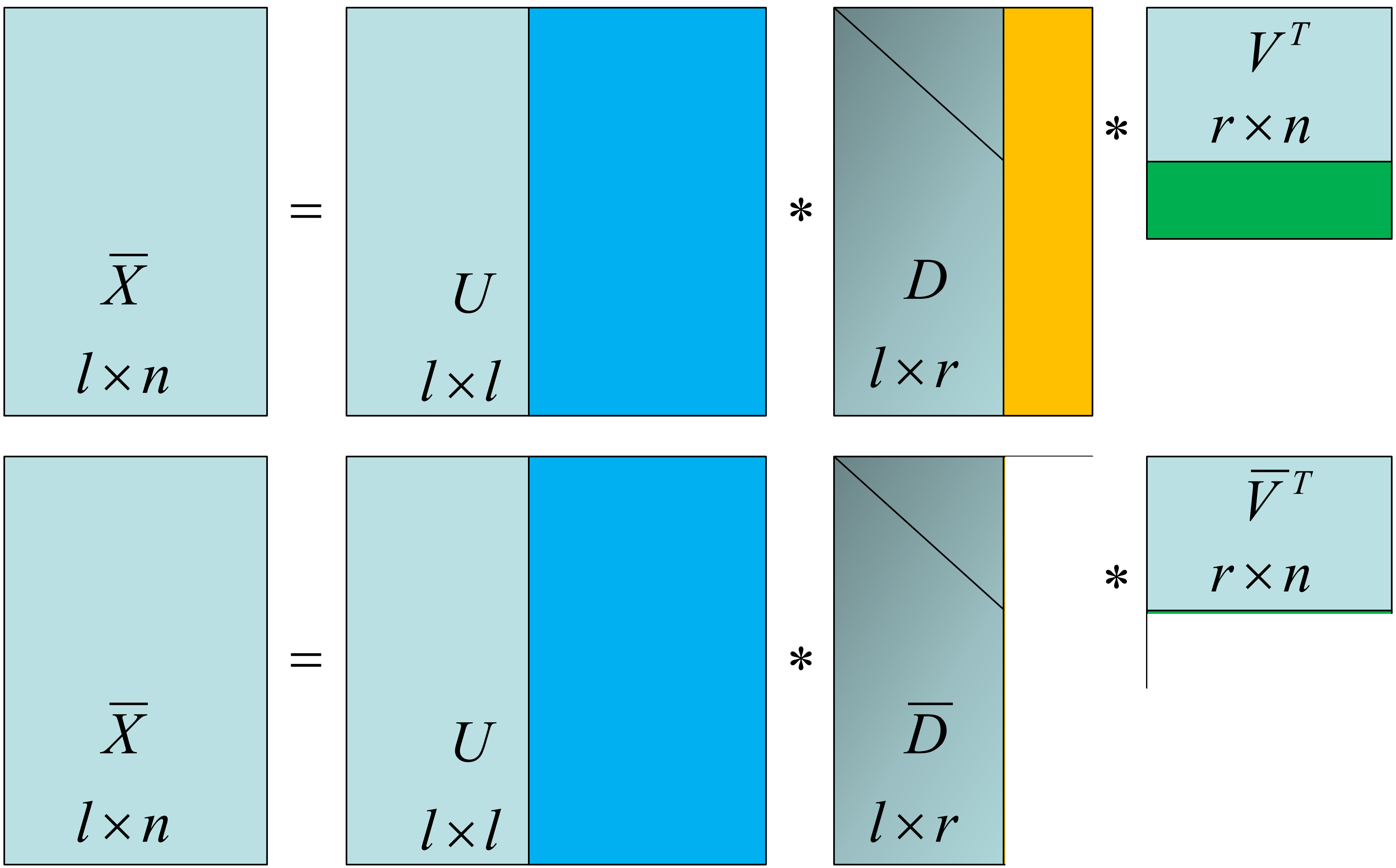
Linear Principle Component Analysis = Singular Value Decomposition of X

$$X = UDV^T \qquad\qquad XV = UD$$



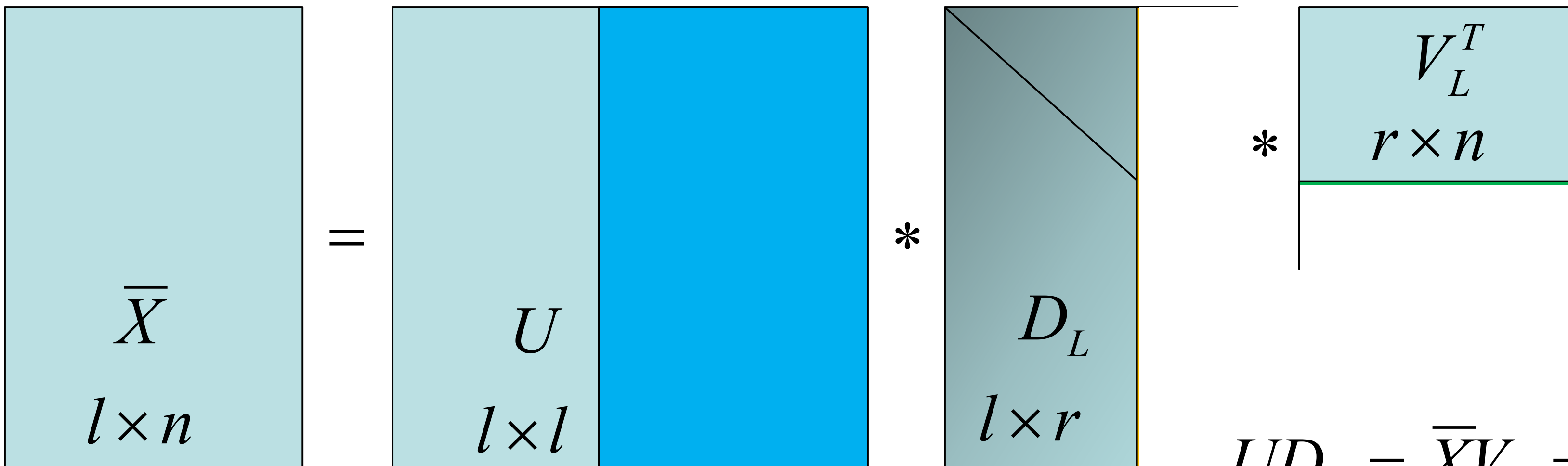
Linear Principle Component Analysis = Singular Value Decomposition of X

$$\bar{X} = UD_L V_L^T \qquad \bar{X}V_L = UD_L = X_L$$

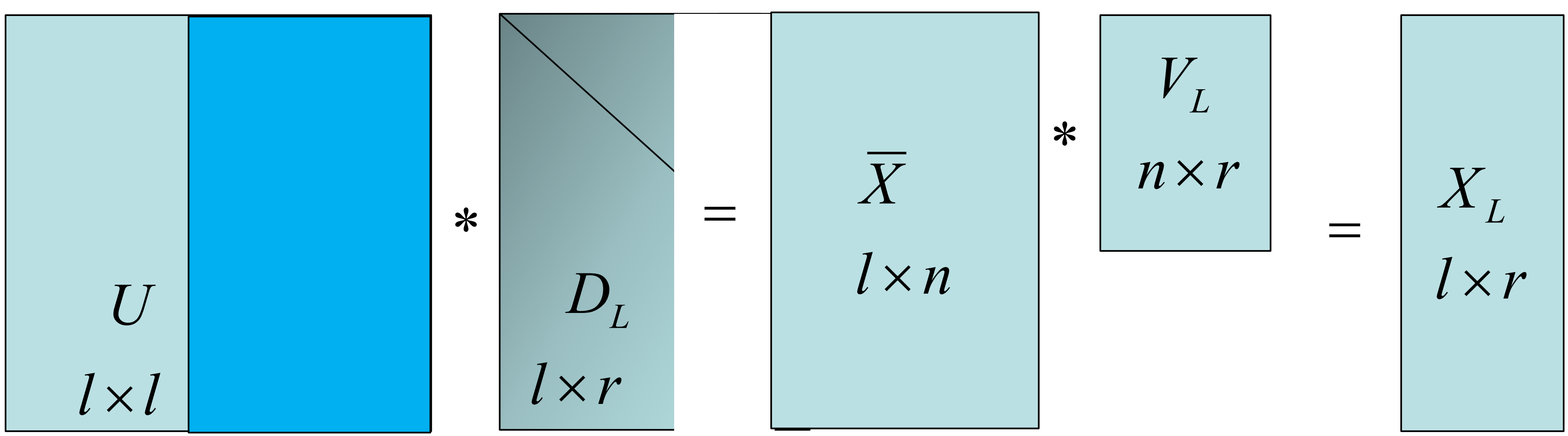


Linear Principle Component Analysis = Singular Value Decomposition of X

$$\bar{X} = UD_L V_L^T$$



$$UD_L = \bar{X}V_L = X_L$$



SVM testing with PCA

1. Calculate the SVD: $X = UDV^T$

2. Reduce the dimensionality of the feature space:

$$X_L^{training} = UD_L, \text{ for training data}$$

$$X_L^{testing} = X^{testing} V_L, \text{ for testing data,}$$

$$V_L \text{ is calculated on the training data}$$

3. Perform the SVM as usual