Support Vector Machines

Emre Ugur, BM 33

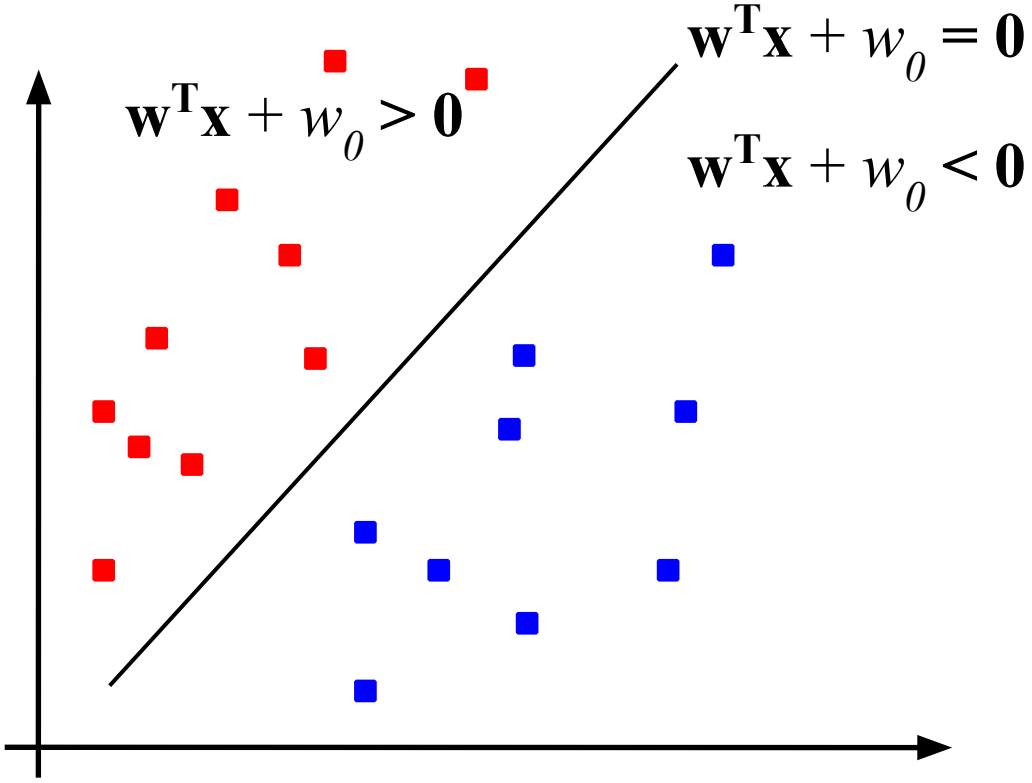
emre.ugur@boun.edu.tr

http://www.cmpe.boun.edu.tr/~emre/courses/cmpe462

cmpe462@listeci.cmpe.boun.edu.tr

Perceptron Revisited: Linear Separators

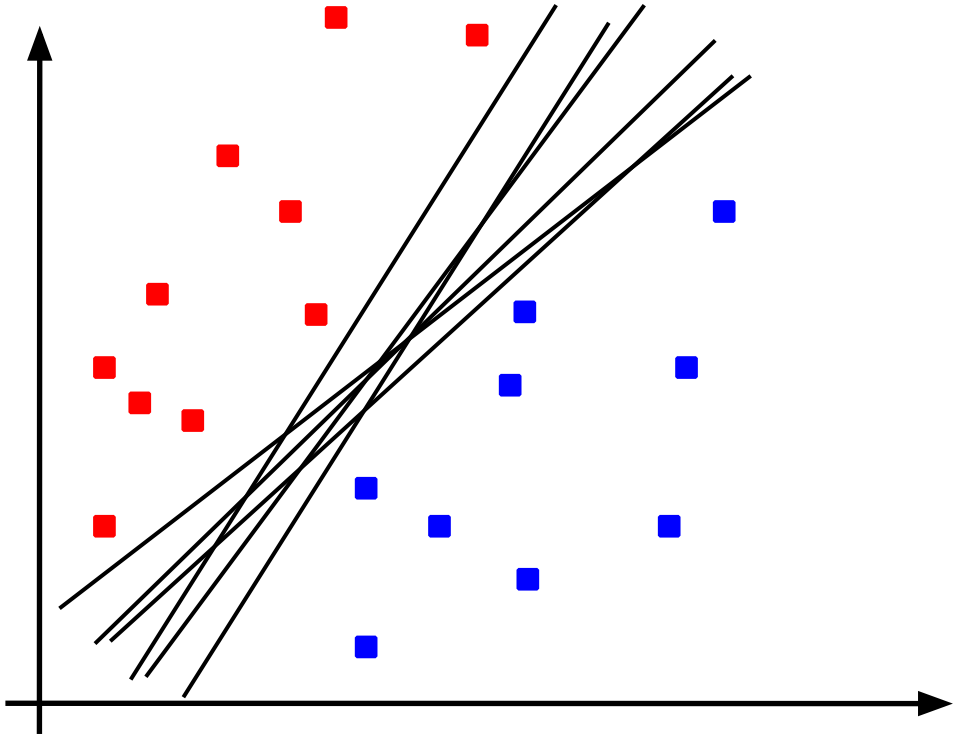
- Binary classification can be viewed as the task of separating classes in feature space:



$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + w_0)$$

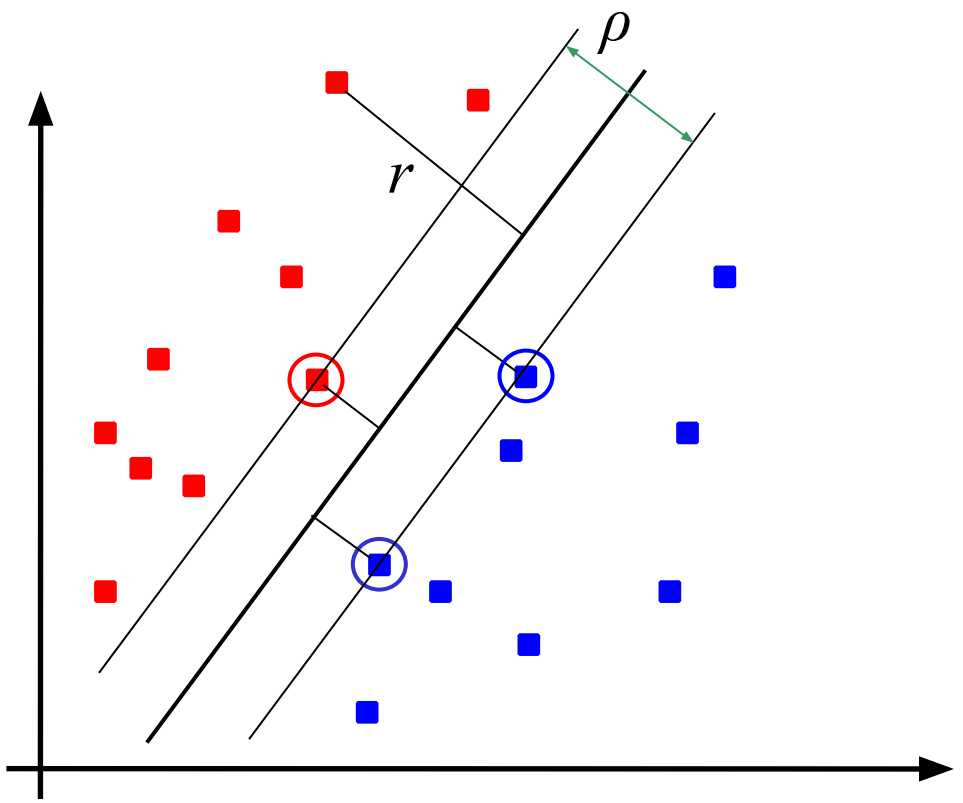
Linear Separators

- Which of the linear separators is optimal?
- Perceptron?



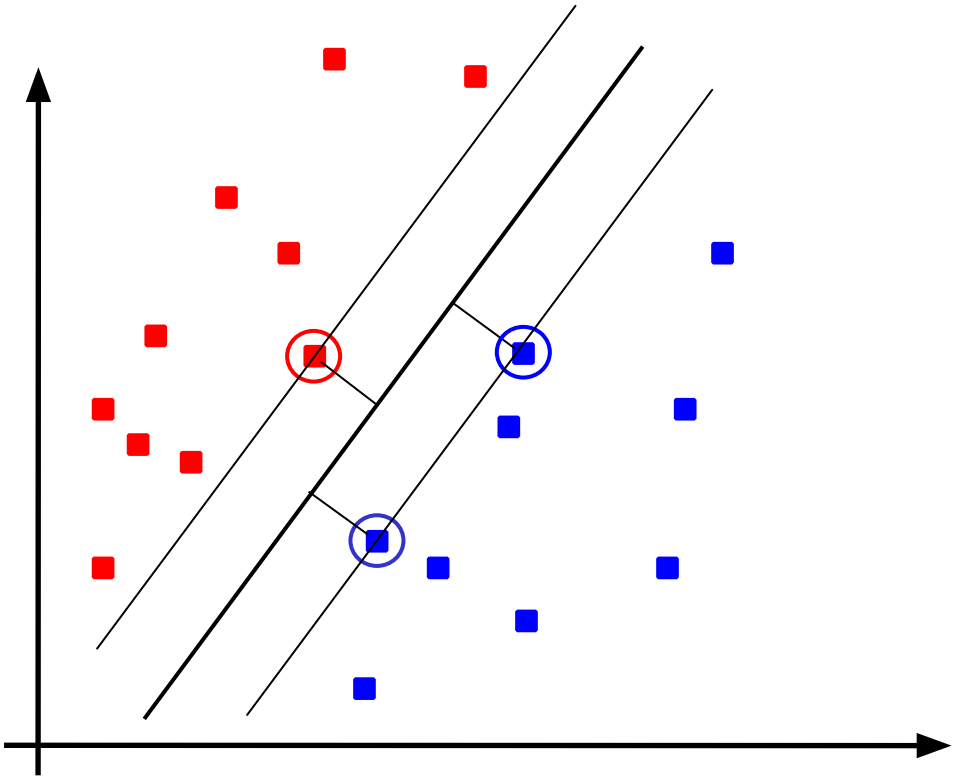
Margin, support vectors

- Margin: the smallest distance between the decision boundary and any of the samples
    - Choose decision boundary to maximize the margin
- Location determined by a set of data points: support vectors



Maximum Margin Classification

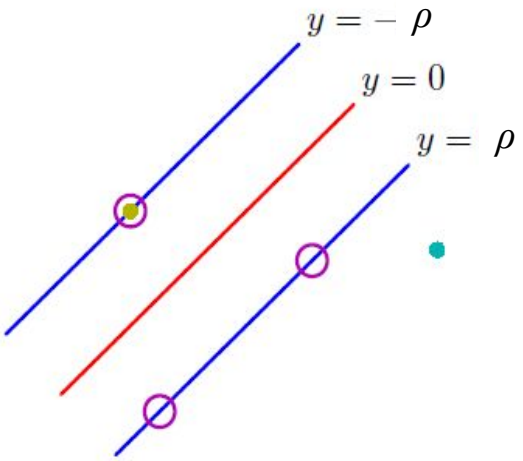
- Maximizing the margin is the aim.
- Implies that only support vectors matter; other training examples are ignorable.



Margin

- The smallest distance between the decision boundary and any of the samples
- Optimal separating hyperplane?

$$\mathcal{X} = \{\boldsymbol{x}^t, r^t\}$$



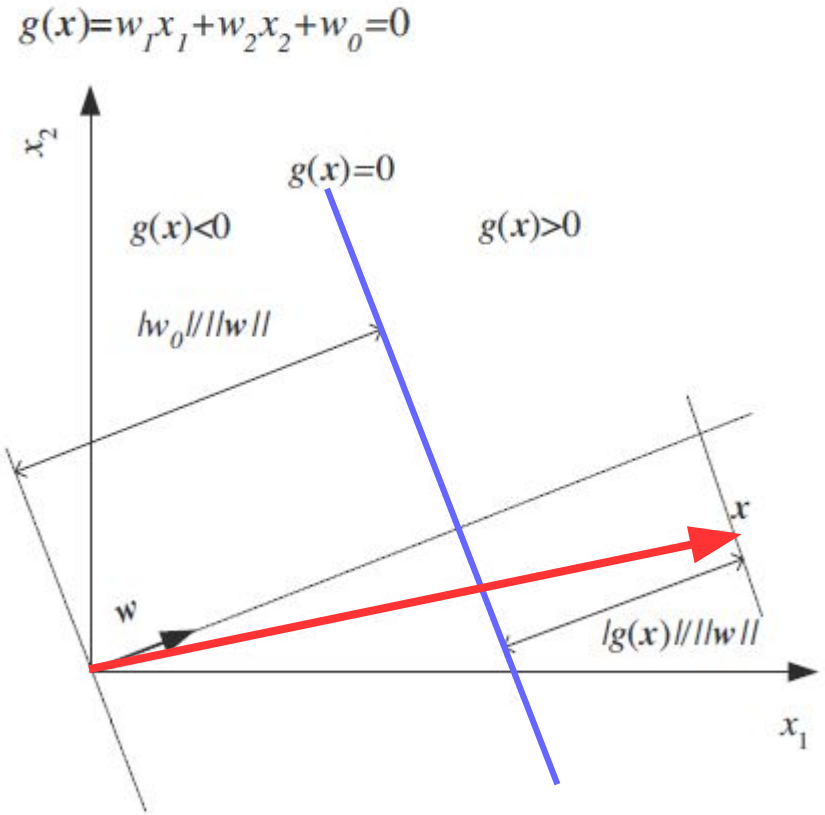
Not only ≥ 0

$$\boldsymbol{w}^T\boldsymbol{x}^t + w_0 \geq +\rho \quad \text{for} \quad r^t = +1$$
$$\boldsymbol{w}^T\boldsymbol{x}^t + w_0 \leq -\rho \quad \text{for} \quad r^t = -1$$

$$r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0) \geq +\rho$$

Reminder

- Distance to decision boundary



$$\frac{|\boldsymbol{w}^T\boldsymbol{x}^t + w_0|}{\|\boldsymbol{w}\|}$$

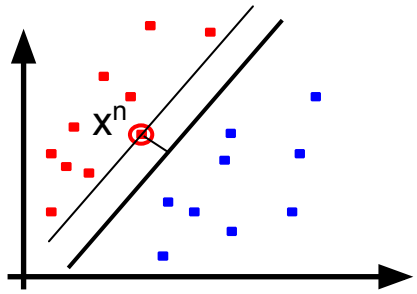
For all data points, at least some value:

$$\frac{r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0)}{\|\boldsymbol{w}\|} \geq \rho, \forall t$$

Maximize ρ

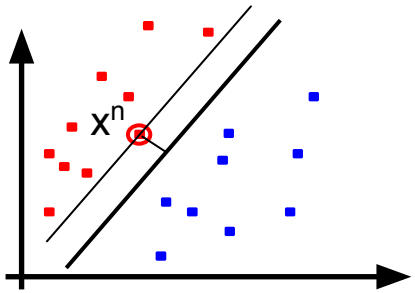
Minimization method

$$\frac{r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0)}{\|\boldsymbol{w}\|} \geq \rho, \forall t$$



Minimization method

$$\frac{r^t(\mathbf{w}^T\mathbf{x}^t + w_0)}{\|\mathbf{w}\|} \geq \rho, \forall t$$



Scale w for a unique solution. Fix $\rho\|\mathbf{w}\| = 1$, minimize ||w|| to maximize ρ

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \text{ subject to } r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t$$

New formulation using Lagrange multipliers.

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t [r^t(\mathbf{w}^T\mathbf{x}^t + w_0) - 1]$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_t \alpha^t r^t(\mathbf{w}^T\mathbf{x}^t + w_0) + \sum_t \alpha^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_t \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_t \alpha^t r^t = 0$$

$$L_d = \frac{1}{2}(\mathbf{w}^T\mathbf{w}) - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}(\mathbf{w}^T\mathbf{w}) + \sum_t \alpha^t$$

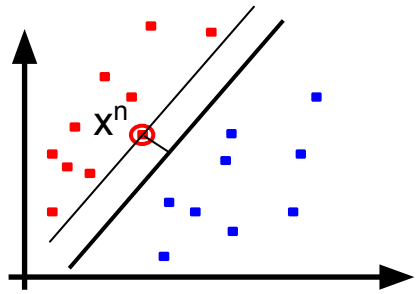
$$= -\frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_t \alpha^t$$

subject to constraints:

$$\sum_t \alpha^t r^t = 0, \text{ and } \alpha^t \geq 0, \forall t$$

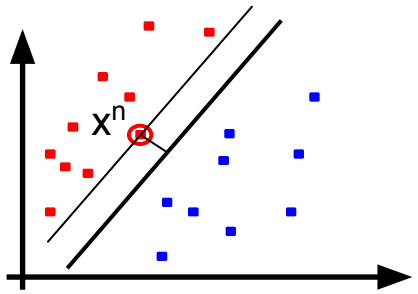
Minimization method

$$-\frac{1}{2}\sum_t\sum_s \alpha^t\alpha^s r^t r^s (\boldsymbol{x}^t)^T \boldsymbol{x}^s + \sum_t \alpha^t$$



Minimization method

$$-\frac{1}{2}\sum_t\sum_s \alpha^t\alpha^s r^t r^s (\boldsymbol{x}^t)^T\boldsymbol{x}^s + \sum_t \alpha^t$$



We solve for α^t , we see that though there are N of them, most vanish with $\alpha^t = 0$
Only a small percentage have $\alpha^t > 0$. The set of x^t whose $\alpha^t > 0$ are the support vectors.

$$\boldsymbol{w} = \sum_t \alpha^t r^t \boldsymbol{x}^t$$

$$r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0) = 1$$

Each support vector should satisfy

$$w_0 = r^t - \boldsymbol{w}^T\boldsymbol{x}^t$$

Therefore, find w_0 from each support vector and take average.

$$w_0$$

Discriminant:

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

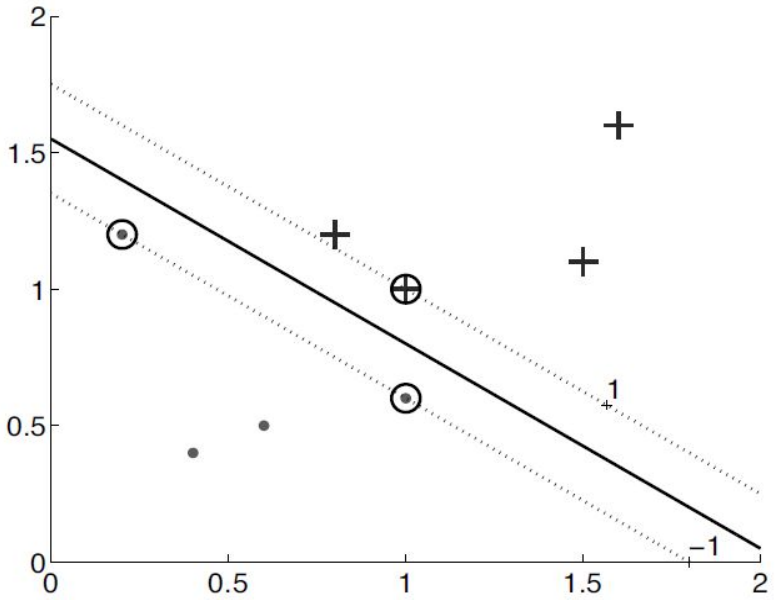
Class of any point/instance x:

Choose C_1 if $g(\boldsymbol{x}) > 0$ and C_2 otherwise

Hyperplane and support vectors

$$\boldsymbol{w} = \sum_t \alpha^t r^t \boldsymbol{x}^t$$

support vectors...



$$r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0) = 1$$

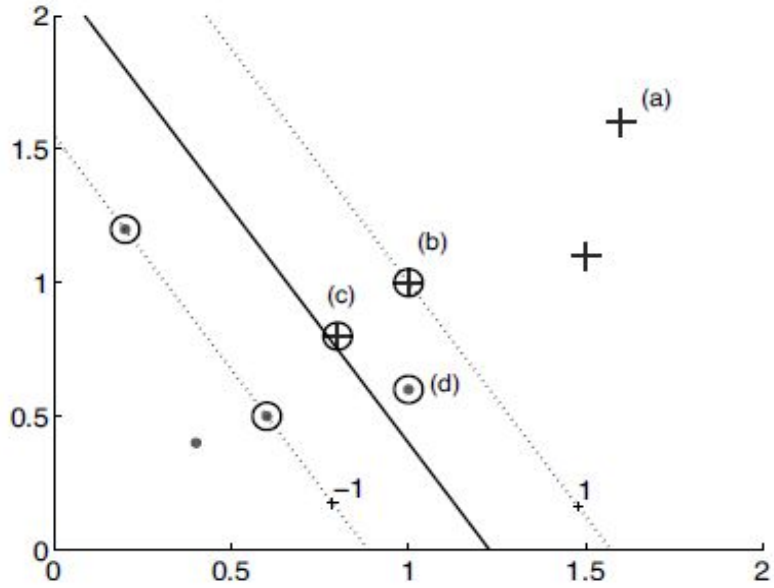
$$w_0 = r^t - \boldsymbol{w}^T\boldsymbol{x}^t$$

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

Choose C_1 if $g(\boldsymbol{x}) > 0$ and C_2 otherwise

Non-separable case: soft margin hyperplane

- Define slack variable $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t$



$\xi^t = 0$ → no problem with x^t

$0 < \xi^t < 1$ → within the margin

$\xi^t >= 1$ → x^t is misclassified

Define soft error add as penalty

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t$$

subject to

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t$$

Non-separable case: soft margin hyperplane

- Lagrangian equation (enforcing positive slack variables)

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t [r^t(\boldsymbol{w}^T\boldsymbol{x}^t + w_0) - 1 + \xi^t] - \sum_t \mu^t \xi^t$$

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_t \alpha^t r^t \boldsymbol{x}^t = 0 \Rightarrow \boldsymbol{w} = \sum_t \alpha^t r^t \boldsymbol{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = \sum_t \alpha^t r^t = 0$$

$$\frac{\partial L_p}{\partial \xi^t} = C - \alpha^t - \mu^t = 0$$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s (\boldsymbol{x}^t)^T \boldsymbol{x}^s$$

subject to

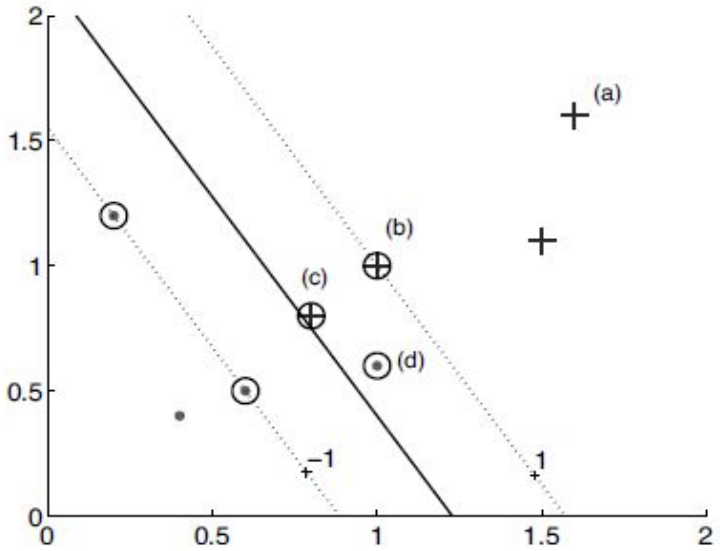
$$\sum_t \alpha^t r^t = 0 \text{ and } 0 \le \alpha^t \le C, \forall t$$

Non-separable case: soft margin hyperplane

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s (\boldsymbol{x}^t)^T \boldsymbol{x}^s$$

subject to

$$\sum_t \alpha^t r^t = 0 \text{ and } 0 \le \alpha^t \le C, \forall t$$



- $\alpha^t = 0$, vanished
- $0 < \alpha^t$, support vectors
  - $0 < \alpha^t < C$, on the margin
  - $\alpha^t = C$, in the margin or misclassified

$$\boldsymbol{w} = \sum_t \alpha^t r^t \boldsymbol{x}^t$$

$$r^t (\boldsymbol{w}^T \boldsymbol{x}^t + w_0) = 1$$

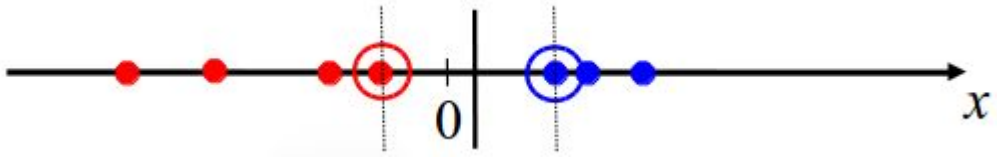
Non-separable case: soft margin hyperplane

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t$$

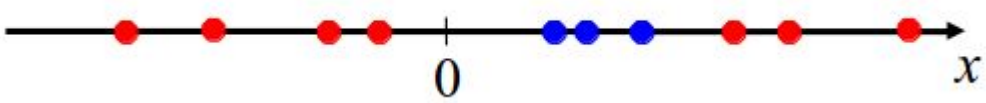
- C is the hyper-parameter
  - Margin maximization vs. error minimization
  - Too large?
  - Too small?

Kernel trick – linear separability

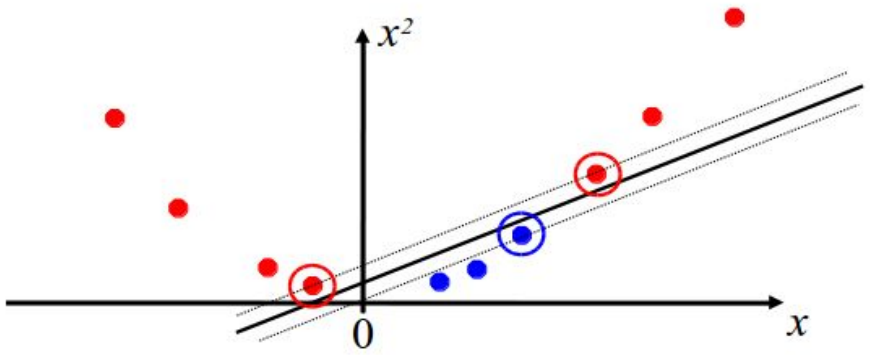
- Datasets that are linearly separable with some noise work out great:



- But what are we going to do if the dataset is just too hard?

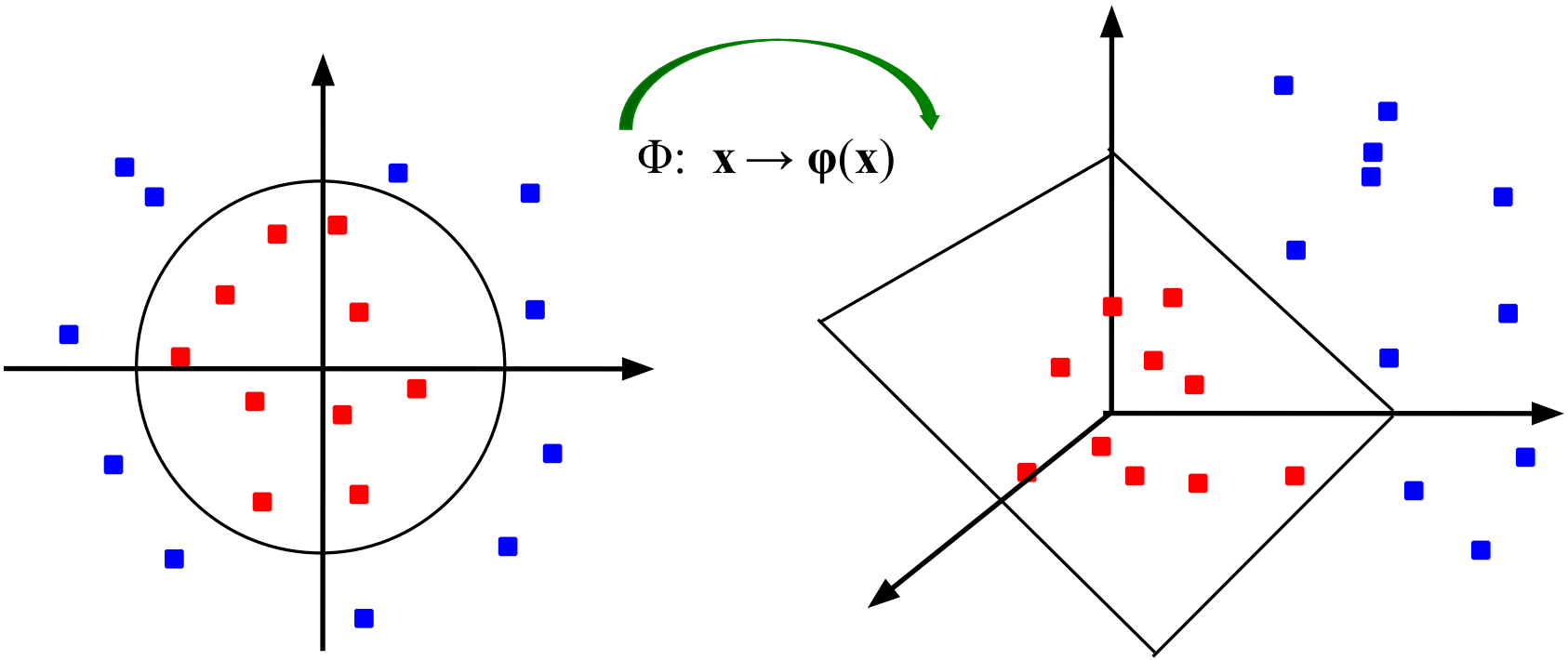


- How about… mapping data to a higher-dimensional space:



Kernel trick : change feature space

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



Kernel Trick

If every datapoint is mapped into high-dimensional space via some transformation Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$

$$\boldsymbol{z} = \boldsymbol{\phi}(\boldsymbol{x}) \text{ where } z_j = \phi_j(\boldsymbol{x}), j = 1, \ldots, k$$

The new discriminant:

$$\begin{aligned} g(\boldsymbol{z}) &= \boldsymbol{w}^T \boldsymbol{z} \\ g(\boldsymbol{x}) &= \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \end{aligned}$$

k is much larger than d (and may be larger than N).

Problem is same:

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t$$

constraints in new space

$$r^t \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^t) \geq 1 - \xi^t$$

Kernel Trick

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t \qquad r^t \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^t) \geq 1 - \xi^t$$

The Lagrangian is

$$L_p = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t \left[r^t \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^t) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

The dual is now

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \boldsymbol{\phi}(\boldsymbol{x}^t)^T \boldsymbol{\phi}(\boldsymbol{x}^s)$$

Kernel trick:

$$K(\boldsymbol{x}^t, \boldsymbol{x}^s)$$

subject to

$$\sum_t \alpha^t r^t = 0 \text{ and } 0 \leq \alpha^t \leq C, \forall t$$

Kernel Trick

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \boldsymbol{\phi}(\boldsymbol{x}^t)^T \boldsymbol{\phi}(\boldsymbol{x}^s)$$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s K(\boldsymbol{x}^t, \boldsymbol{x}^s)$$

$$\begin{aligned} g(\boldsymbol{x}) &= \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) = \sum_t \alpha^t r^t \boldsymbol{\phi}(\boldsymbol{x}^t)^T \boldsymbol{\phi}(\boldsymbol{x}) \\ &= \sum_t \alpha^t r^t K(\boldsymbol{x}^t, \boldsymbol{x}) \end{aligned}$$

- We do not map the feature space at all!

Kernel functions

- Polynomial

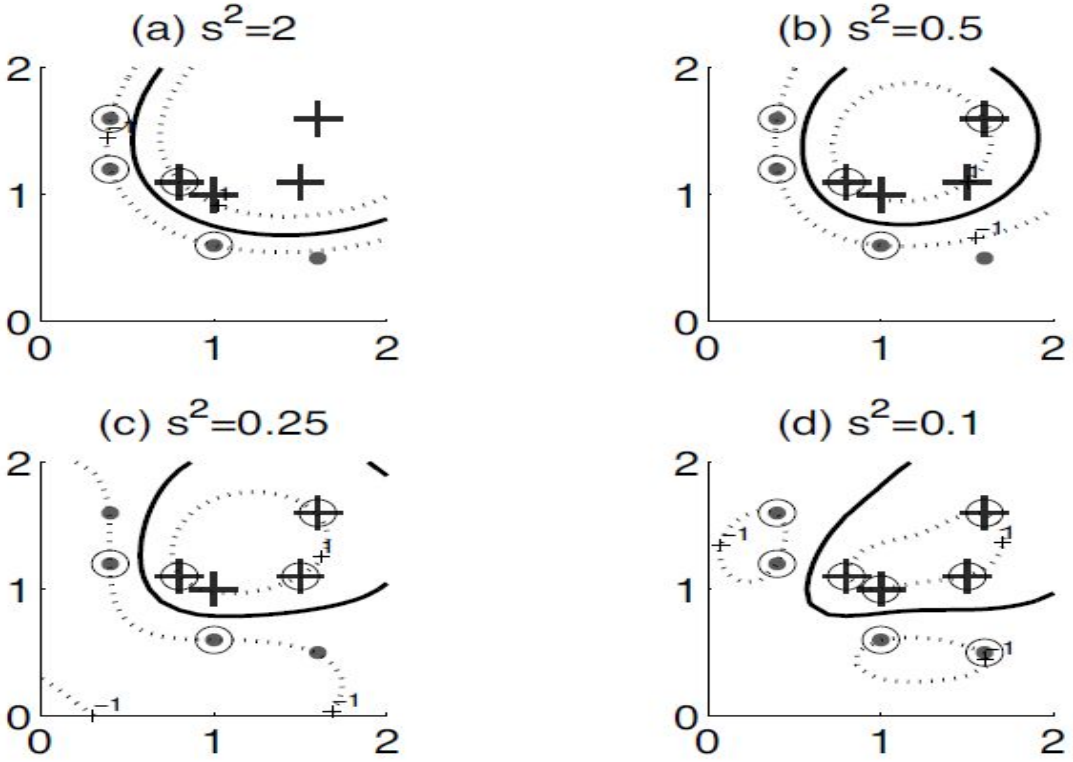
$$K(\boldsymbol{x}^t, \boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{x}^t + 1)^q$$
$$= (x_1 y_1 + x_2 y_2 + 1)^2$$
$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$
$$\phi(\boldsymbol{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^T$$

- Radial-basis functions

$$K(\boldsymbol{x}^t, \boldsymbol{x}) = \exp\left[-\frac{\|\boldsymbol{x}^t - \boldsymbol{x}\|^2}{2s^2}\right]$$

- Sigmoidal functions

$$K(\boldsymbol{x}^t, \boldsymbol{x}) = \tanh(2\boldsymbol{x}^T \boldsymbol{x}^t + 1)$$



Time for demo?