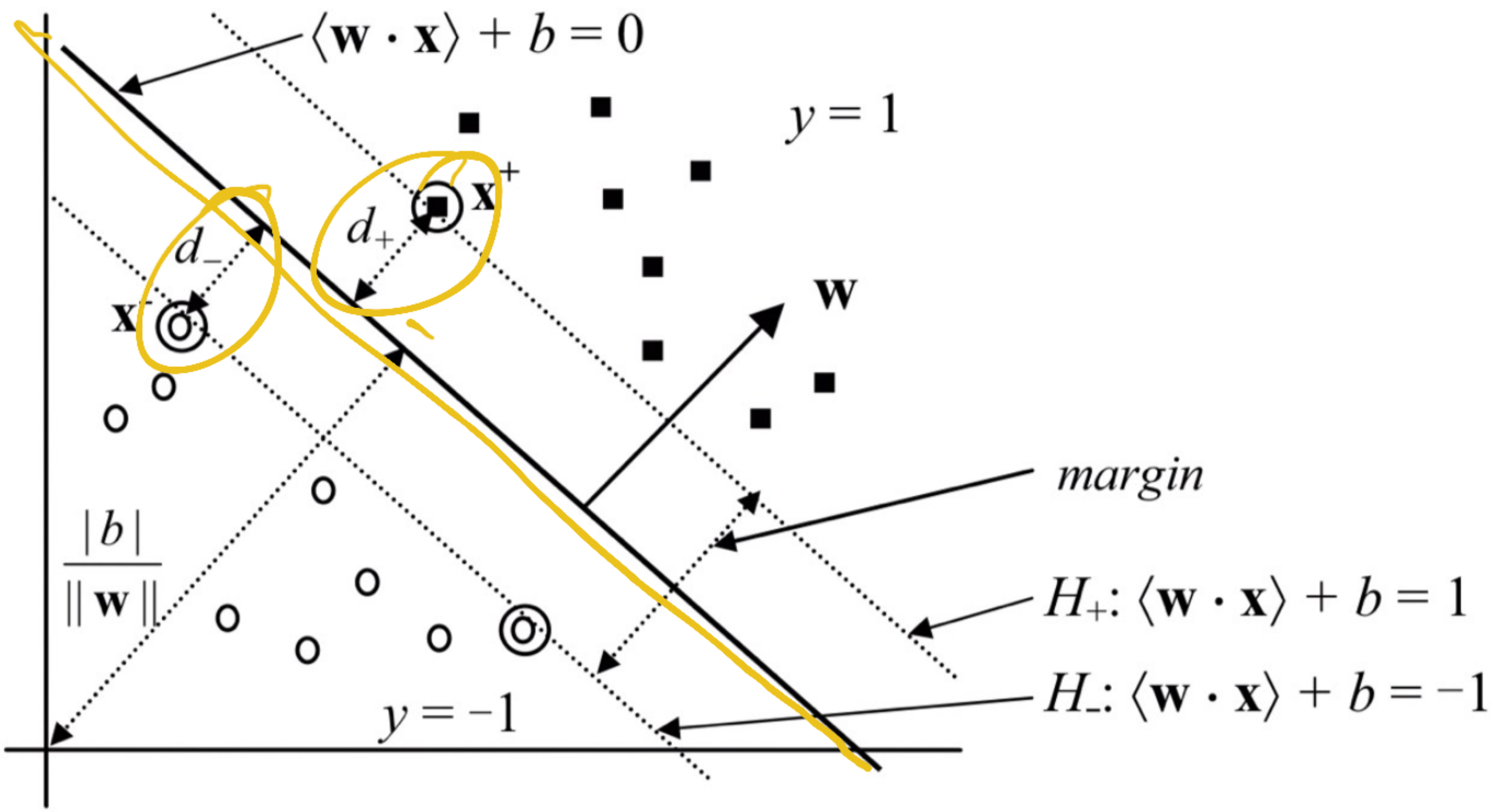
DMML, 21 Feb 2019

Support vector machines



Find a maximum margin separating hyperplane

$\frac{1}{\text{distance}}$

$$\text{Minimize}: \quad \frac{\langle \mathbf{w}\cdot\mathbf{w}\rangle}{2} + \cancel{C\sum_{i=1}^{n}\xi_i}$$

— Objective

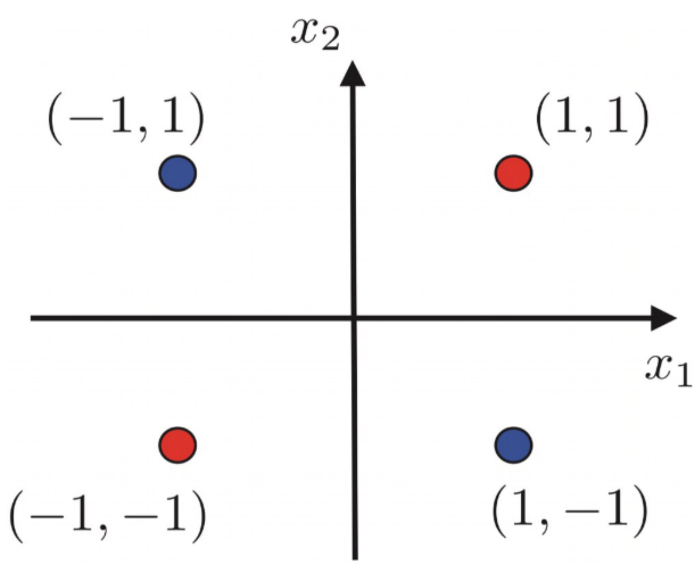
$$\text{Subject to}: \quad y_i(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b) \geq 1 - \cancel{\xi_i}, \quad i = 1, 2, \ldots, n$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, n.$$

— Constraints

± 1

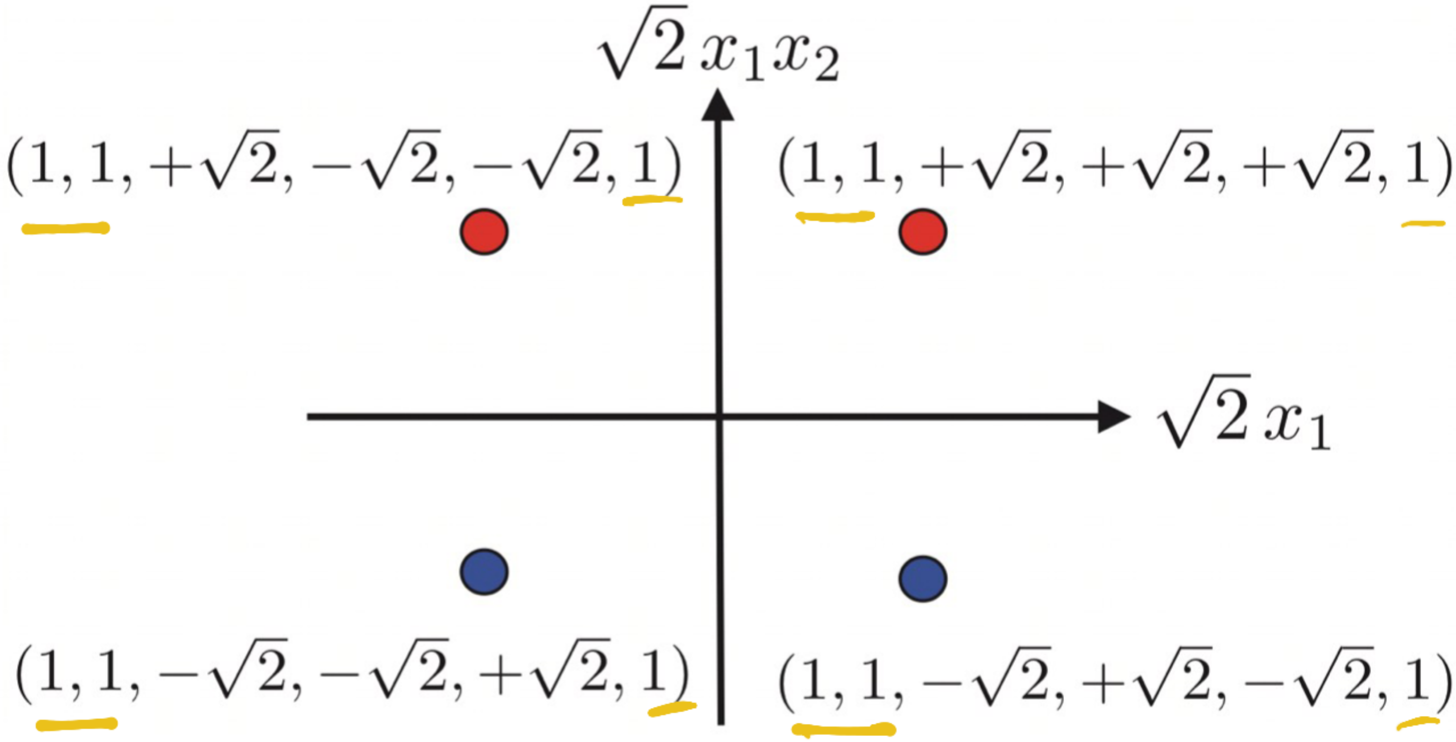
Why is there a separating hyperplane?



XOR

(Exclusive OR)

Solution - transform the space $\quad x \mapsto \varphi(x)$



$$\underbrace{(x_1, x_2)}_{x} \mapsto \underbrace{(x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1)}_{\varphi(x)}$$

1. How to find the transformation?
2. How to solve the problem in terms of more complex $\varphi(x)$?
3. Will φ always exist?

Dual formulation of optimization

$$\text{Maximize: } L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$$

$$\text{Subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, n.$$

$\langle \varphi(x_i) \cdot \varphi(x_j) \rangle$

Back from dual to primal

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \sum_{i=1}^{n} y_i \alpha_i \langle \mathbf{x}_i \cdot \mathbf{x} \rangle + b = 0.$$

$$\varphi((x_1, x_2)) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1)$$

$$\varphi((x_1, x_2)) \cdot \varphi((z_1, z_2))$$

$$(z_1^2, z_2^2, \sqrt{2}z_1, \sqrt{2}z_2, \sqrt{2}z_1z_2, 1)$$

$$x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1z_1 + 2x_2z_2 + 2x_1x_2z_1z_2 + 1$$

$$= (1 + x_1z_1 + x_2z_2)^2$$

Claim Can express $\varphi(x) \cdot \varphi(z)$ as a function $K(x, z)$

Such a function K is called the kernel of the transformation

Dual uses only dot products

$$\text{Maximize: } L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$$

$K(x_i, x_j)$ instead of $\langle \varphi(x_i), \varphi(x_j) \rangle$

$$\text{Subject to: } \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, n.$$

Same is true for Perceptron

In perceptron: $w = x_1 - x_2 + x_3 \ldots$

In each iteration,

$$x^T \cdot w = \underbrace{x^T (x_1 - x_2 + \ldots)}_{\text{Bunch of dot products}}$$

Apply Perceptron to $\varphi(x_1), \varphi(x_2) \ldots, \varphi(x_n)$

Replace each $\varphi(x_i) \cdot \varphi(x_j)$ by $K(x_i, x_j)$

Reversing the argument

$$K(x,z) = (1 + x_1 z_1 + x_2 z_2)^2$$

$$= (1 + x \cdot z)^2$$

is the kernel for $\varphi(x) = \text{---}$

In general, $(1 + x \cdot z)^d$ will be the kernel for <u>some</u> transformation

Have a kernel, ignore the actual transformation

Polynomial kernels

└ Expands the # of dimensions

When is K a kernel?

$$K \quad \begin{array}{cc} & x_j \\ x_i & \begin{bmatrix} n \times n \\ \text{matrix} \end{bmatrix} \end{array}$$

<u>Mercer's Theorm</u>

If K is positive definite then K is a valid kernel

K positive definite

- Symmetric
- For all vectors a $\quad a^T K a \geq 0$

Equivalently, all eigenvalues ≥ 0

Mercer's theorem is <u>not</u> constructive

Given valid K, we need not be able to determine φ

In particular, $\varphi(x)$ may be infinite dimensional

Exponential kernel

$$e^{-\|x-z\|^2/2\sigma}$$

How to find a good kernel? Empirical

- Try a kernel and see if it works
- Some types of kernels work well for some categories of data (images, text, ...)

Similarity as a kernel seems to work

$$K(x,y) = \text{similarity of } x \text{ \& } y$$

Algebraic properties of kernels

K_1 & K_2 are kernels

$\rightarrow$ $K_1 + K_2$ is a kernel

$\rightarrow$ cK_i is a kernel

$\rightarrow$ $K_1 \cdot K_2$ is a kernel