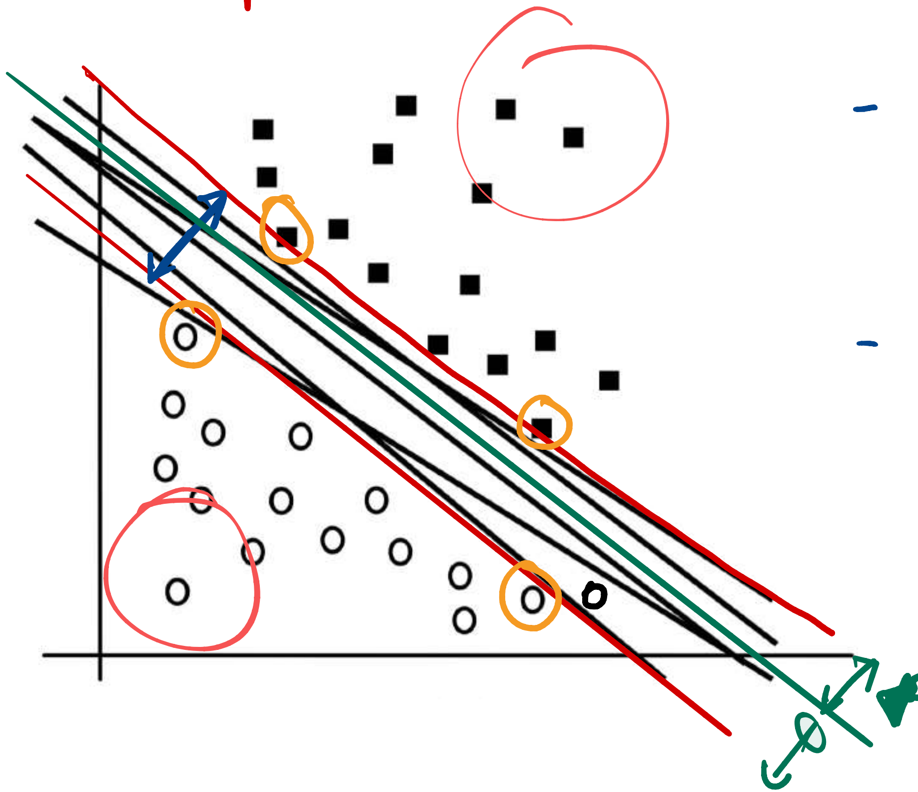
DMML, 5 Mar 2020

Linear Separators – Perceptron



- Which separator is the best one?
- "Maximize the margin"

$$\frac{1}{\|w\|} \times 2$$

Want to find an optimal w^*

Constraints

Assume w^* is scaled appropriately.

$$\forall i.\quad w^* \cdot x_i \cdot l_i > 1$$

Objective function

maximize $\frac{1}{\|w\|} \cdot 2 \rightsquigarrow$ minimize $\frac{\langle w \cdot w \rangle}{2}$

Minimize $\frac{\langle w \cdot w \rangle}{2}$

Subject to

$\forall i. \quad w \cdot x_i \cdot l_i > 1$ $\quad w \cdot x_i \cdot y_i > 1$

Unknowns? w Each constraint is linear

But objective function is quadratic

Not linear programming

Convex optimization

Lagrange Multipliers

One multiplier α_i for each constraint

Minimize

$$\frac{1}{2}\langle w \cdot w\rangle - \sum_{i=1}^{n} \alpha_i \left[w \cdot x_i \cdot l_i - 1 \right]$$

s.t. ---

In general

Minimize $f(x)$

Subject to $g_i(x) \le b_i \quad i = 1, 2, \ldots n$

$$L_p = f(x) + \sum_{i=1}^{n} \alpha_i \left[g(x_i) - b_i \right]$$

If $\alpha_i > 0$ constraint is exact

Solution must satisfy KKT condition

- $\dfrac{\partial L_p}{\partial x_j} = 0 \quad \forall j$
- $g_i(x) - b_i \le 0 \quad \forall i$
- $\alpha_i \ge 0$
- $\alpha_i \left(b_i - g_i(x_i) \right) = 0 \quad \forall i$

Can be solved iteratively

Wolfe Dual

Maximize $L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} l_i l_j \alpha_i \alpha_j \langle x_i \cdot x_j \rangle$

Subject to

$$\left.\begin{aligned} &\sum_{i=1}^{n} l_i \alpha_i = 0 \\ &\alpha_i \geq 0 \end{aligned}\right] \quad \forall i$$

Solving the dual produces a solution of the form

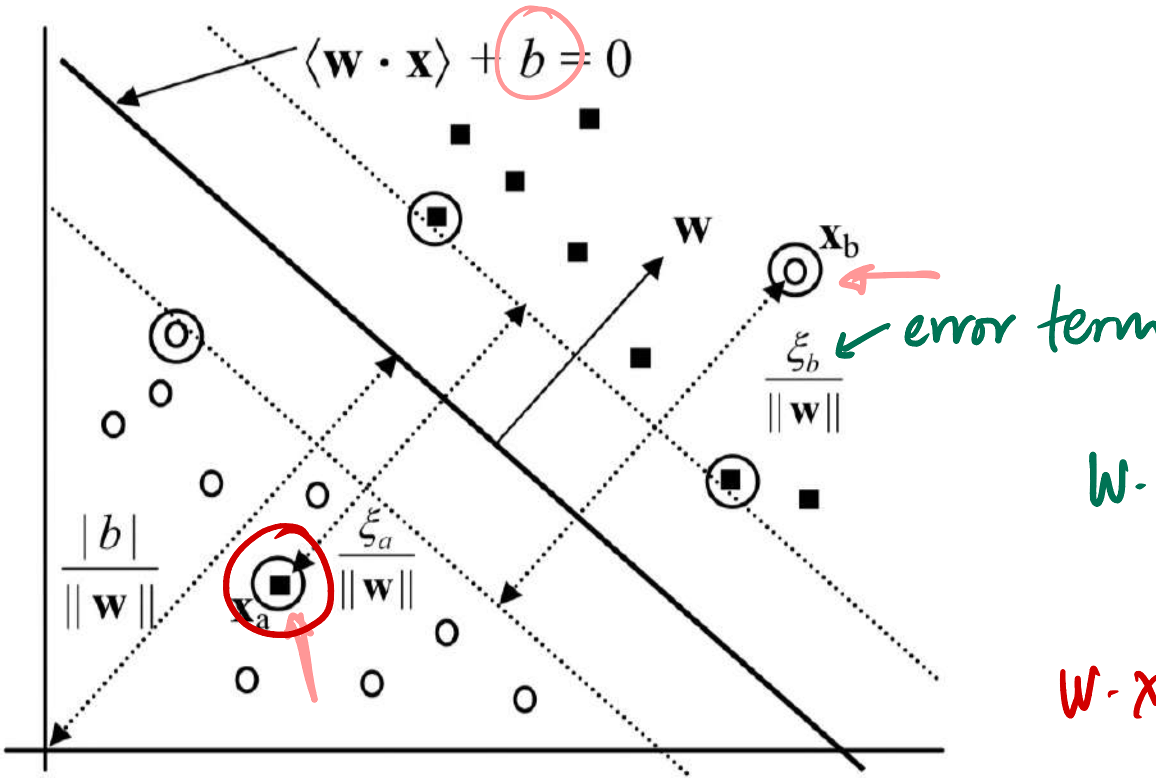
$$\sum_{\substack{i \in SV \\ =}} l_i \alpha_i \langle x_i \cdot x \rangle = 0$$

unknown input

Support vectors $\equiv$ those that satisfy

Support Vector Machine = SVM

Can accommodate non-linearity



$$W \cdot x_i \cdot l_i > 1$$

$$\Downarrow$$

$$W \cdot x_i \cdot l_i + \xi_i > 1$$

Minimize the need for ξ_i

Minimize : $\frac{\langle w \cdot w \rangle}{2} + \sum_{i=1}^{n} \xi_i^2$ or ξ_i^k etc

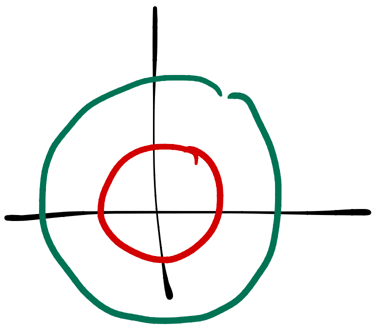
Subject to : $w \cdot x_i \cdot l_i \geq 1 - \underline{\underline{\xi_i}}$

$\xi_i \geq 0$

Soft Margin SVM

Kernel Functions

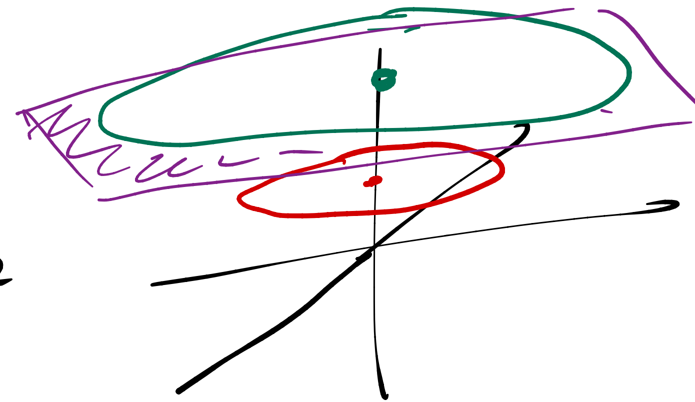
Not linearly separable? $\rightarrow$ Geometric transformation



x, y

$\downarrow$

x, y, x^2+y^2



Final solution depends only on dot products

Solution require $\phi(x_i) \cdot \phi(x_j)$

Find $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$

Suppose $x = \langle x_1, x_2 \rangle$

$$\phi(x) = \langle x_1^2, x_2^2, \sqrt{2}\, x_1 x_2 \rangle$$

$x \quad z \qquad \phi(z) = \langle z_1^2, z_2^2, \sqrt{2}\, z_1 z_2 \rangle$

$$\phi(x) \cdot \phi(z) = x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2$$

$$k(x, z) = \langle x \cdot z \rangle^2 = (x_1 z_1 + x_2 z_2)^2$$

$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2$$

$K(x,y)$ is a kernel function if $\exists \varphi$

s.t. $K(x,y) = \langle \varphi(x) . \varphi(y) \rangle \quad \forall x, y$

Interestingly φ would create infinite dimension

$(x_1, x_2) \rightarrow \langle x_1, x_2, x_1 x_2, x_1^2 x_2 \rangle \ldots \ldots x_1^n x_2^2 \ldots$

Given k, is it a kernel?

Simple case:

Each $\bar{x} = \langle x_1, \ldots, x_m \rangle$, each x_i has finitely many possibilities

(e.g. $x_i = 0, 1$)

Enumerate all possible $\bar{x}$: $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$

Explicitly write $k(\bar{x}, \bar{y})$ as a matrix

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_j & \bar{x}_N \end{array} \\
\begin{array}{c} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_i \\ x_N \end{array} & \begin{bmatrix} & & & & \\ & & & & \\ & & & K(\bar{x}_i, \bar{x}_j) & \\ & & & & \end{bmatrix}
\end{array}
$$

Square Matrix

Symmetric & Positive Semidefinite

$\forall c \in \mathbb{R}^N \quad c^T K c \geq 0$

If K is positive semidefinite, it is a valid kernel

What if x_i's are from unbounded domain?

- Unboundedly many inputs $\bar{x}, \bar{y}$

For any finite set $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ of inputs

k over this subset is an

$n \times n$ matrix

"Gram Matrix"

$$\begin{array}{cc} & \begin{array}{ccc} \bar{x}_1 & \cdots & \bar{x}_n \end{array} \\ \begin{array}{c} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{array} & \begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix} \end{array}$$

For any n, for any subset $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$, the Gram Matrix of k over this subset is positive semidefinite

Typically – k is a similarity measure

Such a k is generally positive semidefinite

Some rules for combining kernels.

1. k is a kernel, so is $c \cdot k$

2. k_1, k_2 are kernels, so is $k_1 + k_2$

3. k_1, k_2 are kernels, so is $k_1 \cdot k_2$

1 is a kernel

$(1+k)^d$ — polynomial kernels

$k(\bar{x},\bar{y}) = \|\bar{x}-\bar{y}\|$

$k(\bar{x},\bar{y}) = e^{-\gamma\|x-y\|}$ Radial Basis Function

RBF

↳ corresponding φ is infinite dimensional

Inputs $\quad x = \langle x_1 \ldots x_m \rangle$

Features

$K(\bar{x}, \bar{y})$ encapsulates all relevant info about features

Individual features no longer matter to perceptron, SVM ...