CHAPTER 13:

Kernel Machines

Kernel Machines: Motivation

- Discriminant-based: No need to estimate densities first
- Define the discriminant in terms of support vectors
- The use of kernel functions, application-specific measures of similarity
- No need to represent instances as vectors
- Convex optimization problems with a unique solution

Optimal Separating Hyperplane

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \text{ where } r^t = \begin{cases} +1 & \text{if } \mathbf{x}^t \in C_1 \\ -1 & \text{if } \mathbf{x}^t \in C_2 \end{cases}$$

find $\mathbf{w}$ and w_0 such that

$$\mathbf{w}^T \mathbf{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$\mathbf{w}^T \mathbf{x}^t + w_0 \leq +1 \text{ for } r^t = -1$$

which can be rewritten as

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

(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

Distance from the discriminant to the closest instances on either side

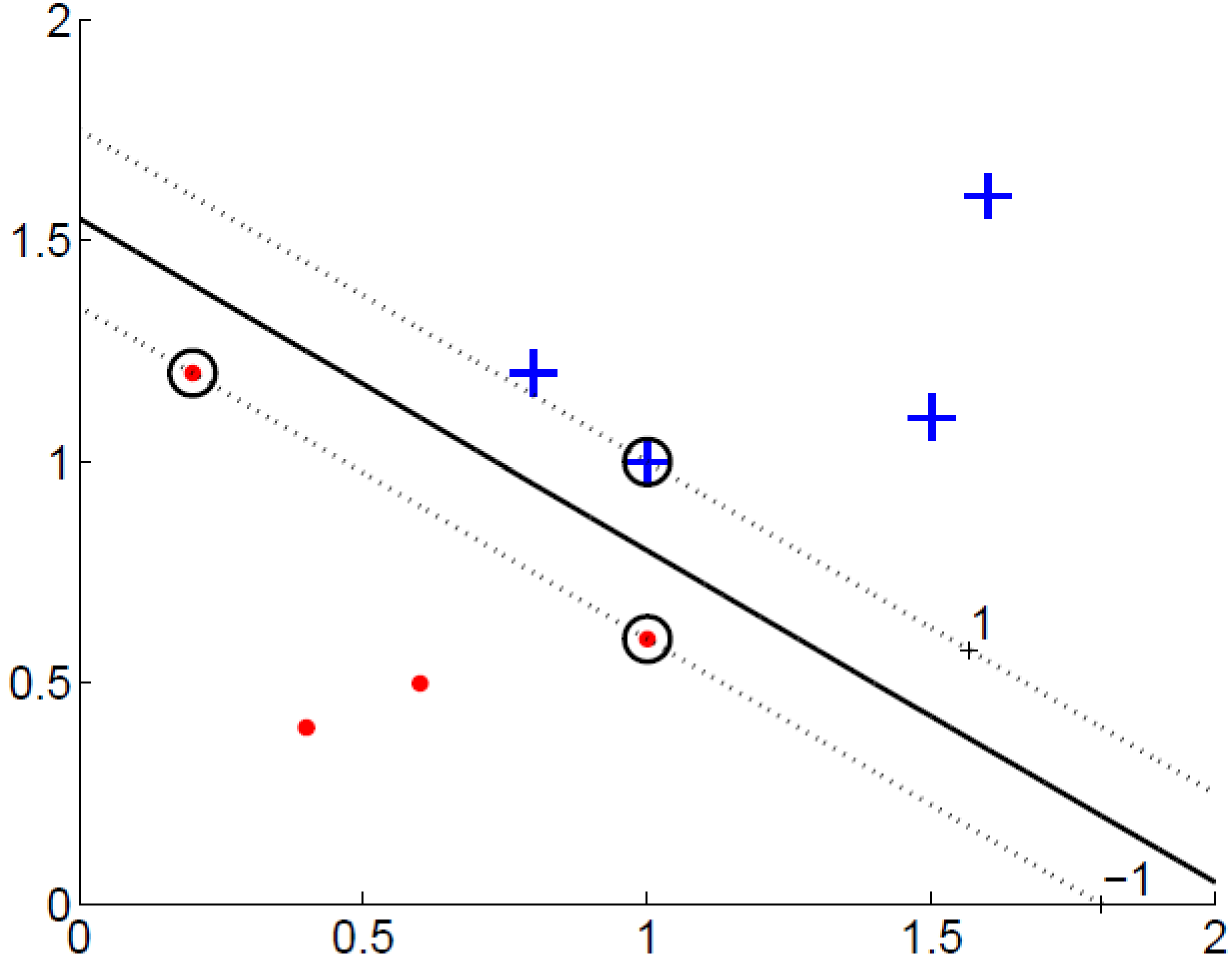
Distance of x to the hyperplane is $\frac{|\mathbf{w}^T\mathbf{x}^t + w_0|}{\|\mathbf{w}\|}$

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

For a unique sol'n, fix $\rho||\mathbf{w}||=1$, and to max margin

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

Margin



Dual Formulation and solution 1/2

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

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

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

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$

Dual Formulation and solution 2/2

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

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

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

$$\text{subject to } \sum_{\text{t}} \alpha^t r^t = 0 \text{ and } \alpha^t \geq 0, \forall t$$

Most α^t are 0 and only a small number have $\alpha^t > 0$; they are the support vectors

Need for soft margin

- Hard margin issues
  - works if the data is linearly separable
  - Sensitive to outliers

- Soft margin promotes more flexible model
  - Flexibility is controlled by parameter *C (*Details in next slide*)*

Soft margin Hyperplane

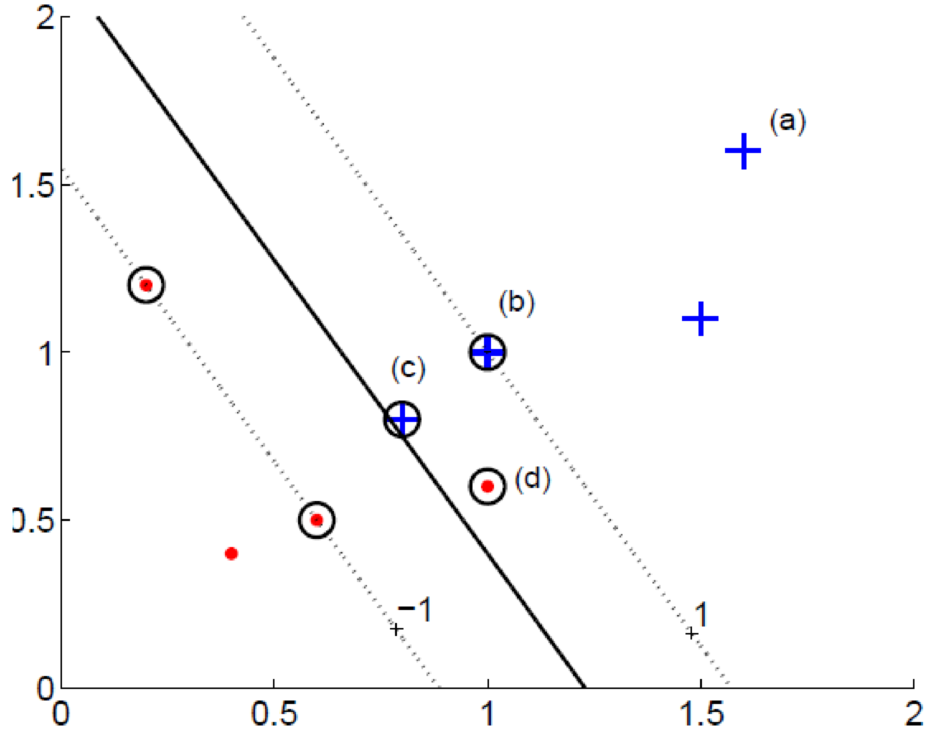
Not linearly separable

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

Soft error

$$\sum_t \xi^t$$

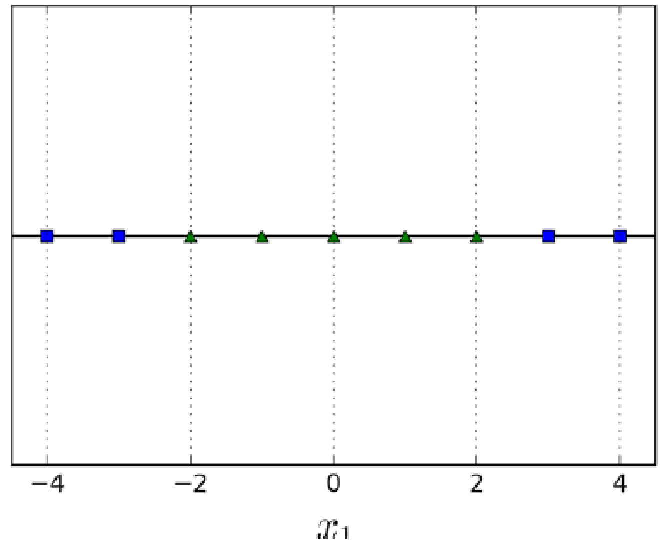
New primal is



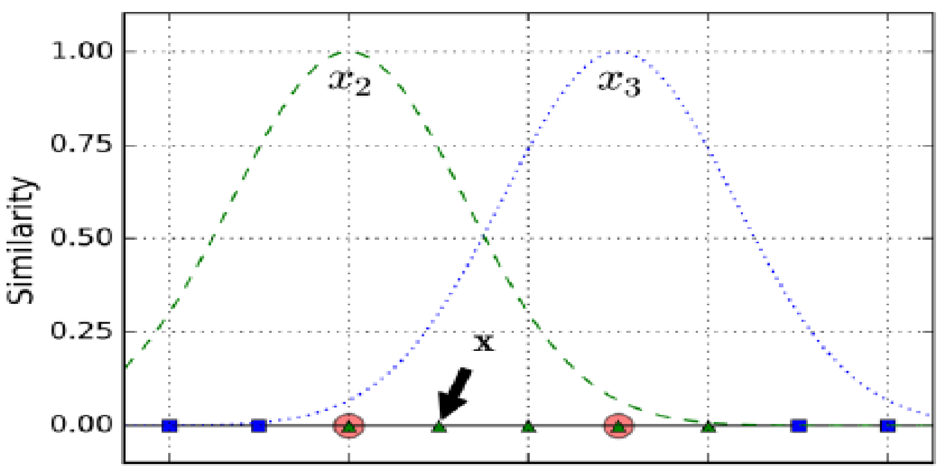
$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_t \alpha^t \left[r^t\left(\mathbf{w}^T x^t + w_0\right) - 1 + \xi^t\right] - \sum_t \mu^t \xi^t$$

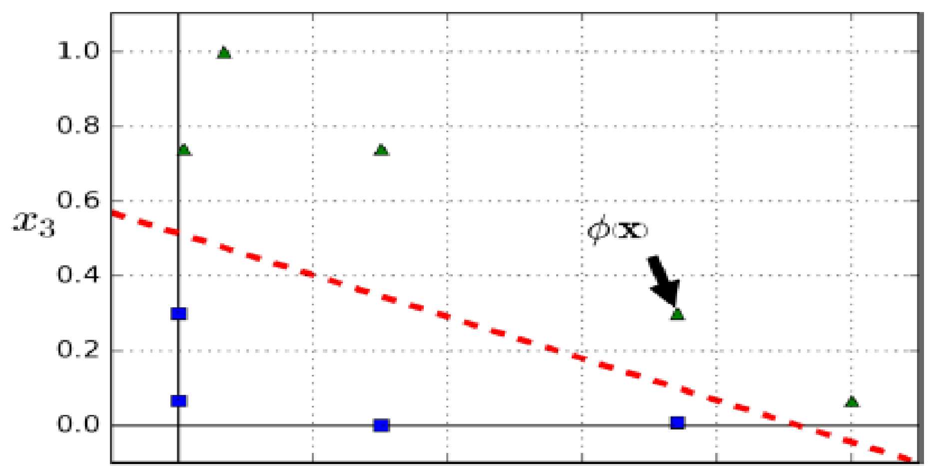
Nonlinear SVM

- Not linearly separable



- Similarity function: $\phi\gamma(\mathbf{x}, \ell) = \exp\left(-\gamma \| \mathbf{x} - \ell \|^2\right)$





Non linear SVM

- When examples are not truly separable.
  - Map attributes to higher dimension
    - ϕ: X-> $\phi(X)$, going d-dimension to D-dimension
    - Then perform SVM

  - Problem: Computation cost of dot operation
    - $X_i.X_j$ requires $O(d^2)$
    - $\phi(X_i).\ \phi(X_j)$ requires $O(D^2)$

Kernel Function

- K(X,Y) = ϕ(X_i). ϕ(X_j)
  - Idea is to compute RHS without expanding individual functions.
- Example:

$$K(\mathbf{x},\mathbf{y}) = \left(\mathbf{x}^T\mathbf{y}+1\right)^2$$
$$= \left(x_1y_1 + x_2y_2 + 1\right)^2$$
$$= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1x_2y_1y_2 + x_1^2y_1^2 + x_2^2y_2^2 \quad = \phi(X).\ \phi(Y)$$
$$\phi(\mathbf{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2\right]^T$$

Kernel Trick: why ?

Preprocess input ***x*** by basis functions

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

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

The SVM solution

$$\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \boldsymbol{\varphi}\left(\mathbf{x}^t\right)$$

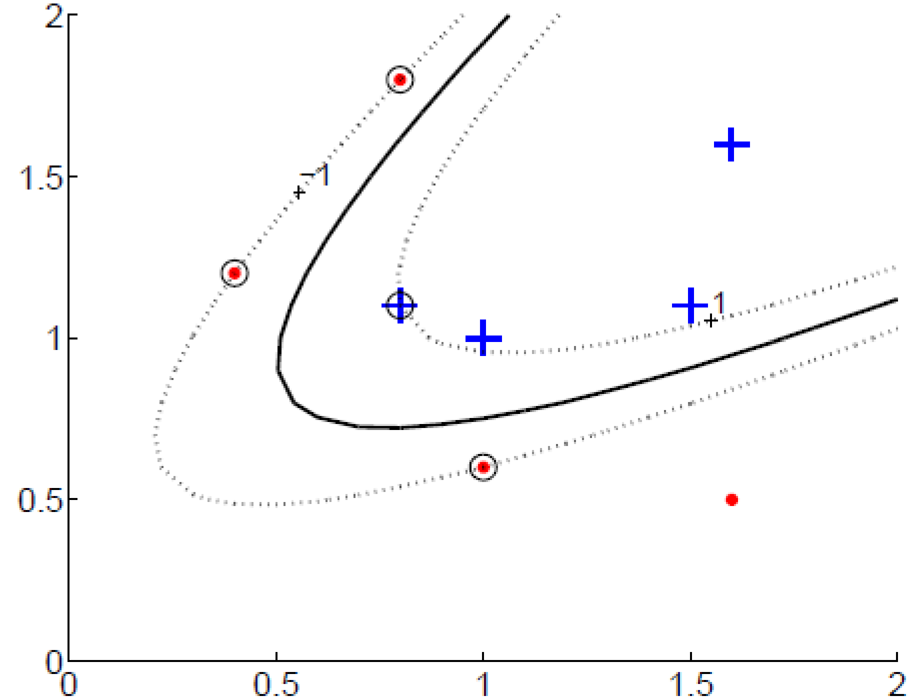
$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\varphi}\left(\mathbf{x}^t\right)^T \boldsymbol{\varphi}(\mathbf{x})}$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{K\left(\mathbf{x}^t, \mathbf{x}\right)}$$

Vectorial Kernels : Popular Kernals 1/2

Polynomials of degree q:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \left(\mathbf{x}^T \mathbf{x}^t + 1\right)^q$$

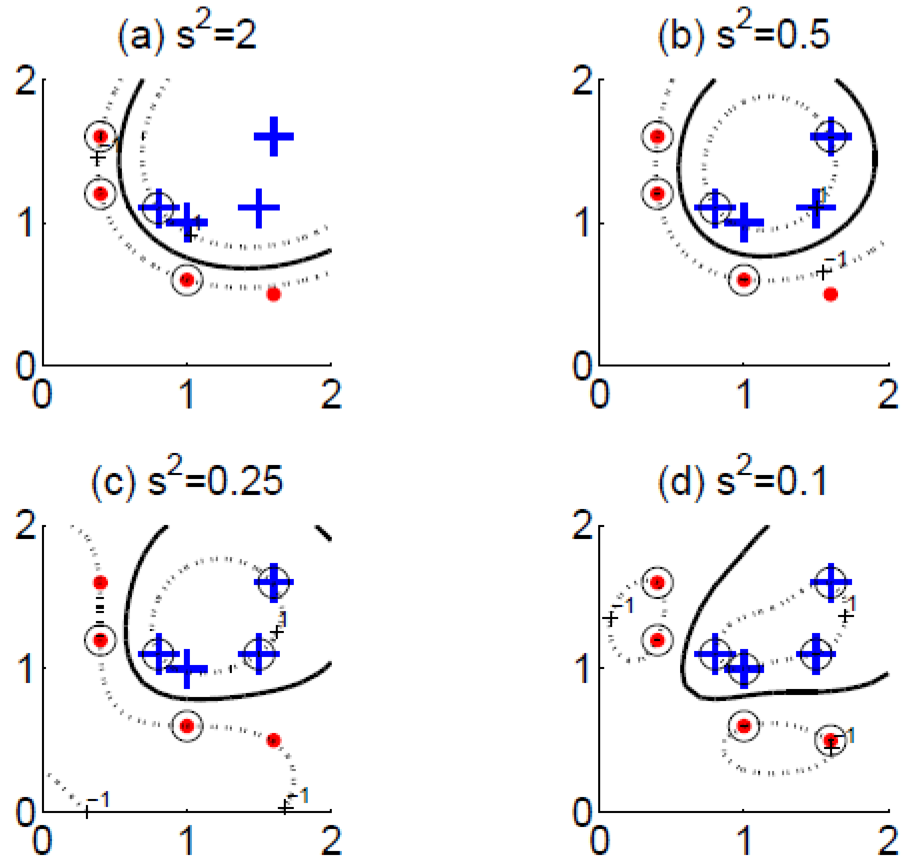


$$K(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^T \mathbf{y} + 1\right)^2$$
$$= \left(x_1 y_1 + x_2 y_2 + 1\right)^2$$
$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$
$$\phi(\mathbf{x}) = \left[1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2\right]^T$$

Vectorial Kernels : Popular Kernals 2/2

Radial-basis functions:

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



Defining kernels

- Kernel “engineering”
- Defining good measures of similarity
- Depending on representation of data, there are String kernels, graph kernels, image kernels, etc...
- Examples: common words between two documents

Multiple Kernel Learning

Replacing a single kernel with multiple kernel

Fixed kernel combination $K(\mathbf{x},\mathbf{y}) = \begin{cases} cK(\mathbf{x},\mathbf{y}) \\ K_1(\mathbf{x},\mathbf{y}) + K_2(\mathbf{x},\mathbf{y}) \\ K_1(\mathbf{x},\mathbf{y})K_2(\mathbf{x},\mathbf{y}) \end{cases}$

Adaptive kernel combination $K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \eta_i K_i(\mathbf{x},\mathbf{y})$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s \sum_i \eta_i K_i(\mathbf{x}^t, \mathbf{x}^s)$$

$$g(\mathbf{x}) = \sum_t \alpha^t r^t \sum_i \eta_i K_i(\mathbf{x}^t, \mathbf{x})$$

Localized kernel combination $g(\mathbf{x}) = \sum_t \alpha^t r^t \sum_i \eta_i(\mathbf{x} \mid \theta) K_i(\mathbf{x}^t, \mathbf{x})$