



Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

© The MIT Press, 2010

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml2e>

CHAPTER 13:

Kernel Machines

Kernel Machines

- 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

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

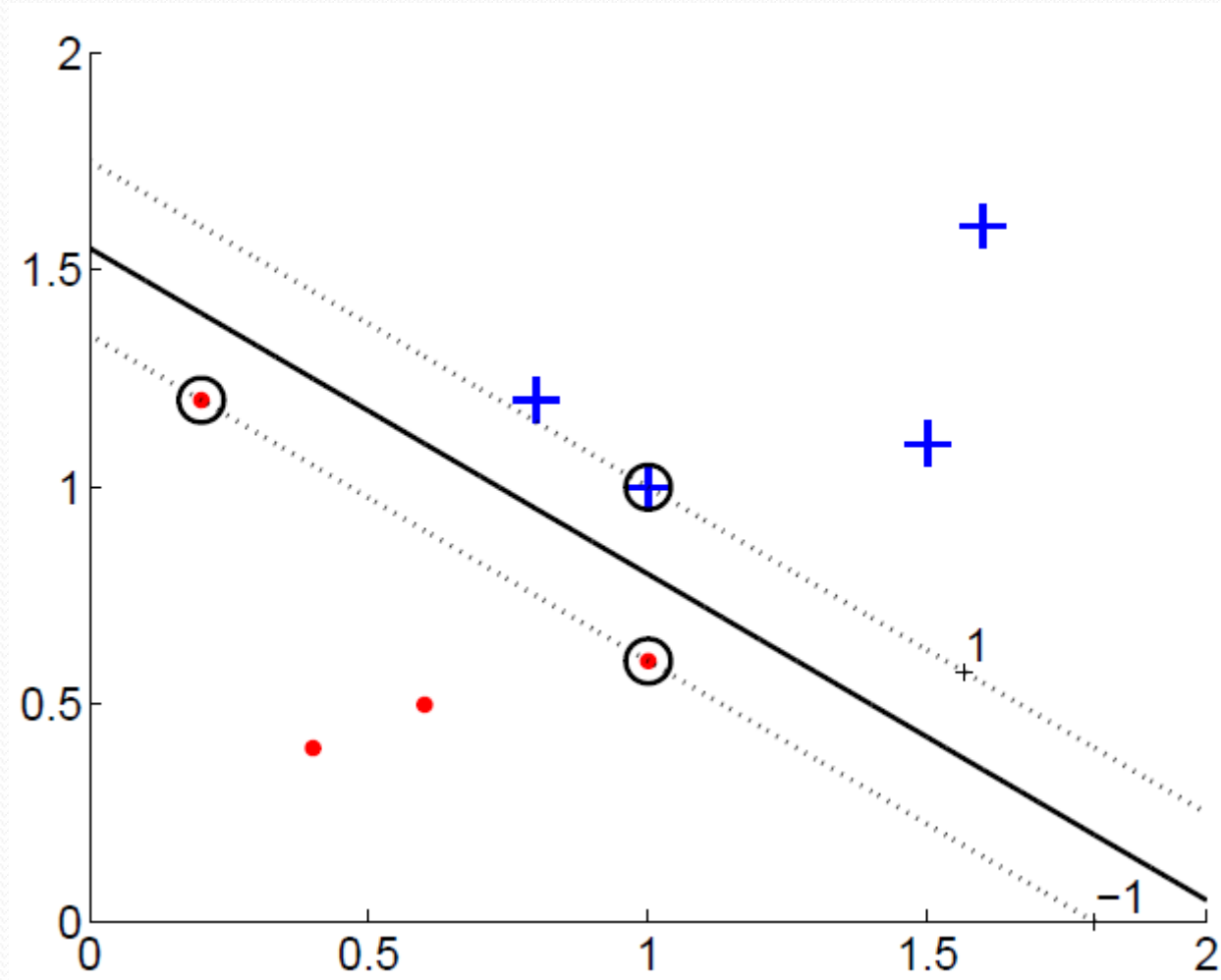
(Cortes and Vapnik, 1995; Vapnik, 1995)

Margin

- Distance from the discriminant to the closest instances on either side
- Distance of \mathbf{x} to the hyperplane is $\frac{|\mathbf{w}^T \mathbf{x}^t + w_0|}{\|\mathbf{w}\|}$
- We require $\frac{r^t(\mathbf{w}^T \mathbf{x}^t + w_0)}{\|\mathbf{w}\|} \geq \rho, \forall t$
- For a unique sol'n, fix $\rho \|\mathbf{w}\| = 1$, and to max margin

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

Margin



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

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

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

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

$$\begin{aligned}
L_d &= \frac{1}{2}(\mathbf{w}^T \mathbf{w}) - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t \\
&= -\frac{1}{2}(\mathbf{w}^T \mathbf{w}) + \sum_t \alpha^t \\
&= -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_t \alpha^t \\
&\text{subject to } \sum_t \alpha^t r^t = 0 \text{ and } \alpha^t \geq 0, \forall t
\end{aligned}$$

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

Soft Margin Hyperplane

- Not linearly separable

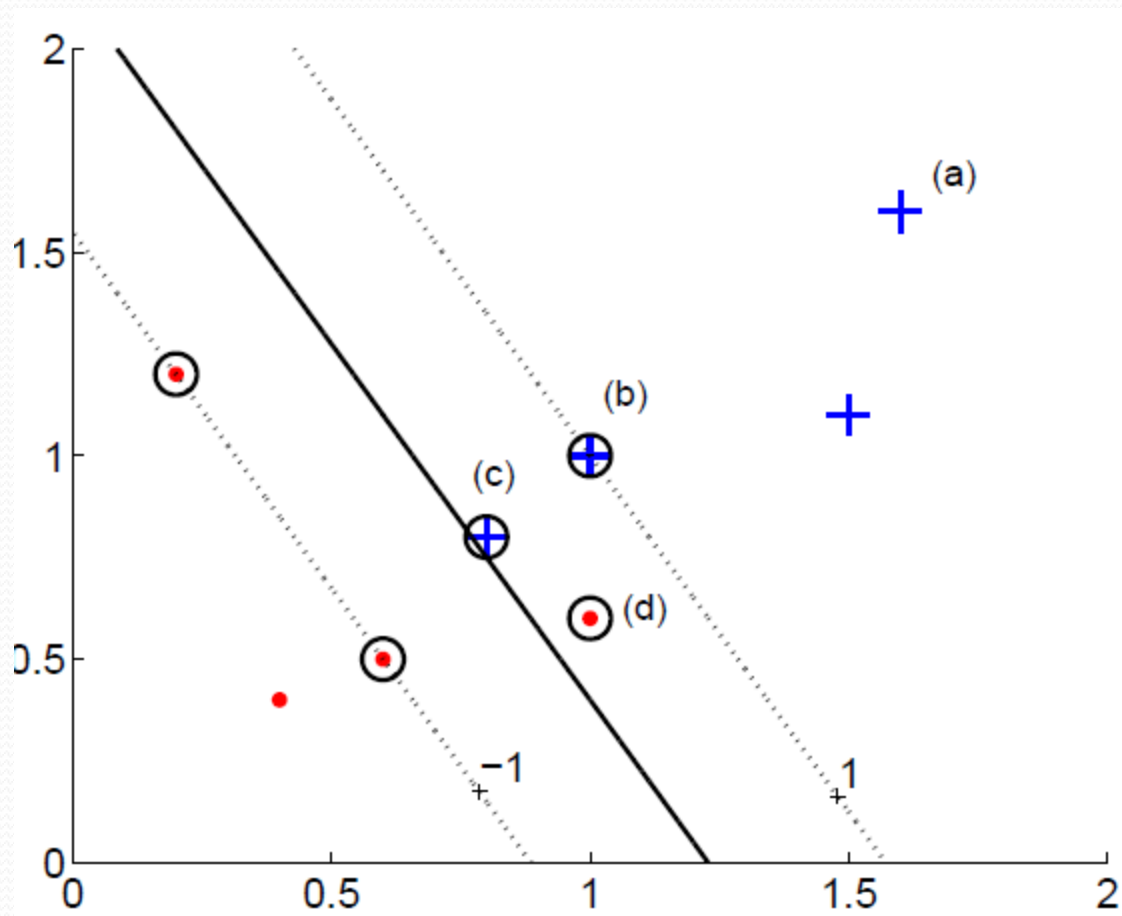
$$r^t(\mathbf{w}^T \mathbf{x}^t + w_0) \geq 1 - \xi^t$$

- Soft error

$$\sum_t \xi^t$$

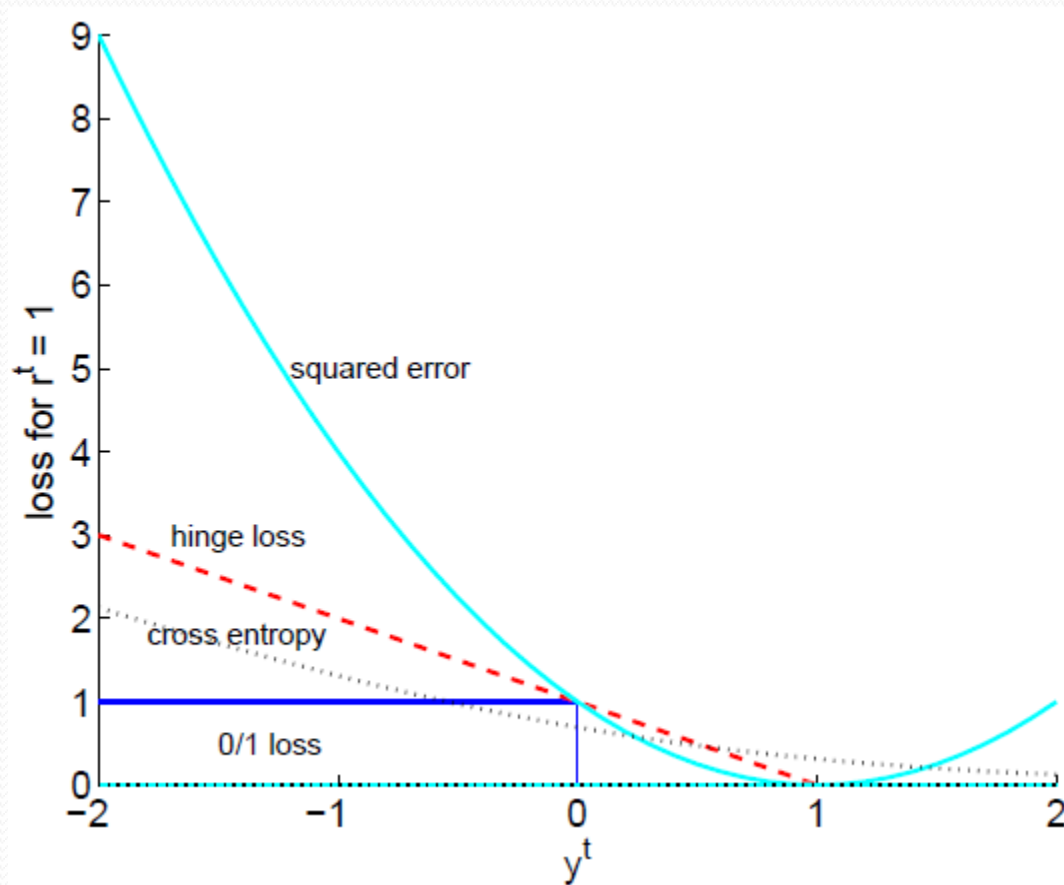
- New primal is

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



Hinge Loss

$$L_{\text{hinge}}(y^t, r^t) = \begin{cases} 0 & \text{if } y^t r^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$



ν -SVM

$$\min \frac{1}{2} \|\mathbf{w}\|^2 - \nu \rho + \frac{1}{N} \sum_t \xi^t$$

subject to

$$r^t (\mathbf{w}^T \mathbf{x}^t + w_0) \geq \rho - \xi^t, \xi^t \geq 0, \rho \geq 0$$

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

subject to

$$\sum_t \alpha^t r^t = 0, 0 \leq \alpha^t \leq \frac{1}{N}, \sum_t \alpha^t \leq \nu$$

ν controls the fraction of support vectors

Kernel Trick

- Preprocess input \mathbf{x} by basis functions

$$\mathbf{z} = \boldsymbol{\varphi}(\mathbf{x})$$

$$g(\mathbf{z}) = \mathbf{w}^T \mathbf{z}$$

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

- The SVM solution

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

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

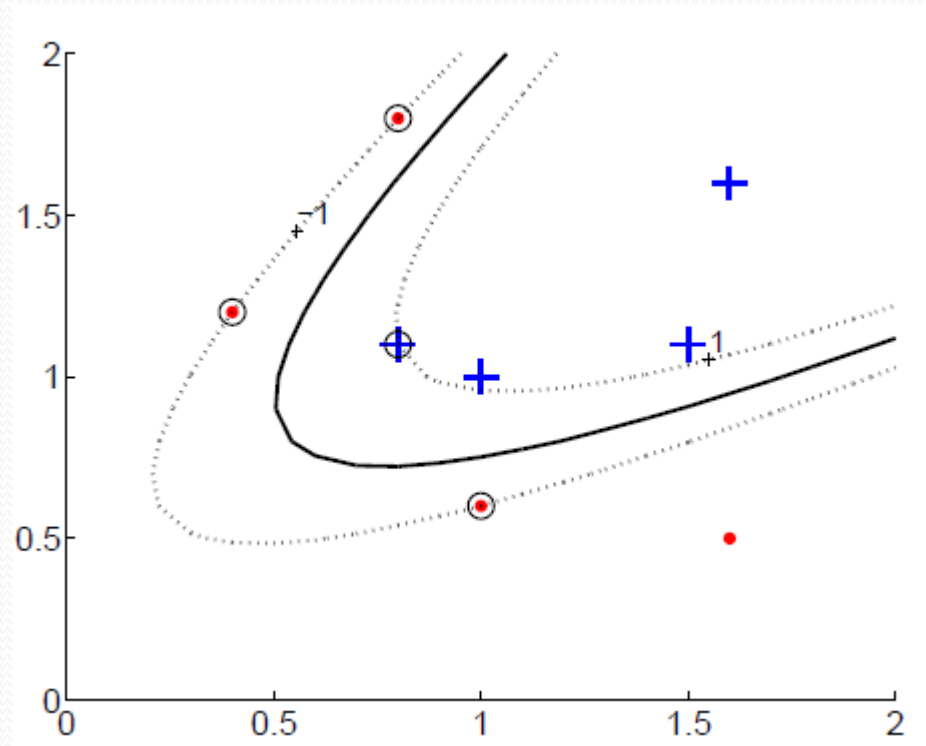
$$g(\mathbf{x}) = \sum_t \alpha^t r^t K(\mathbf{x}^t, \mathbf{x})$$

Vectorial Kernels

- Polynomials of degree q :

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

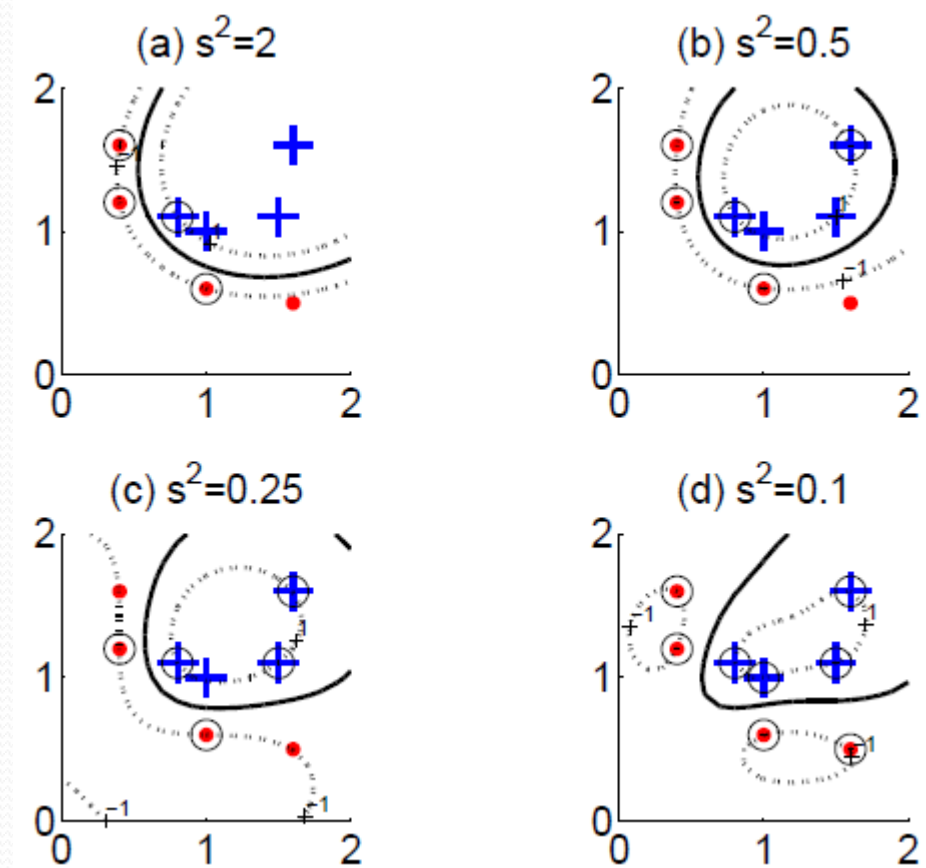
$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}^T \mathbf{y} + 1)^2 \\ &= (x_1 y_1 + x_2 y_2 + 1)^2 \\ &= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 \\ \phi(\mathbf{x}) &= [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^T \end{aligned}$$



Vectorial Kernels

- 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
- String kernels, graph kernels, image kernels, ...
- Empirical kernel map: Define a set of templates \mathbf{m}_i and score function $s(\mathbf{x}, \mathbf{m}_i)$

$$\phi(\mathbf{x}^t) = [s(\mathbf{x}^t, \mathbf{m}_1), s(\mathbf{x}^t, \mathbf{m}_2), \dots, s(\mathbf{x}^t, \mathbf{m}_M)]$$

and

$$K(\mathbf{x}, \mathbf{x}^t) = \phi(\mathbf{x})^T \phi(\mathbf{x}^t)$$

Multiple Kernel Learning

- 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

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

Multiclass Kernel Machines

- 1-vs-all
- Pairwise separation
- Error-Correcting Output Codes (section 17.5)
- Single multiclass optimization

$$\min \frac{1}{2} \sum_{i=1}^K \|\mathbf{w}_i\|^2 + c \sum_i \sum_t \xi_i^t$$

subject to

$$\mathbf{w}_{z^t}^T \mathbf{x}^t + w_{z^t 0} \geq \mathbf{w}_i^T \mathbf{x}^t + w_{i0} + 2 - \xi_i^t, \forall i \neq z^t, \xi_i^t \geq 0$$

SVM for Regression

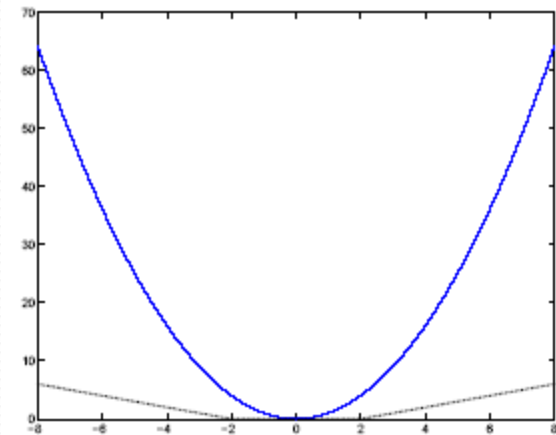
- Use a linear model (possibly kernelized)

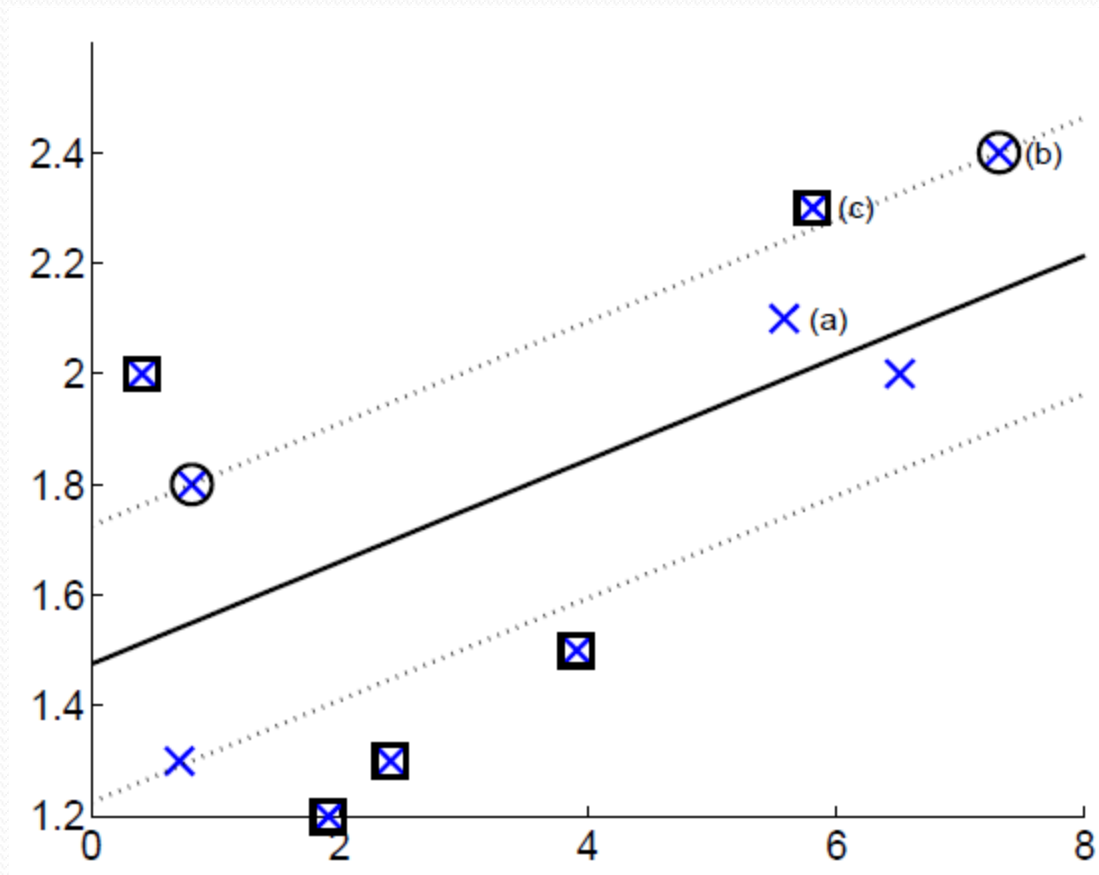
$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

- Use the ϵ -sensitive error function

$$e_\epsilon(r^t, f(\mathbf{x}^t)) = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| < \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$

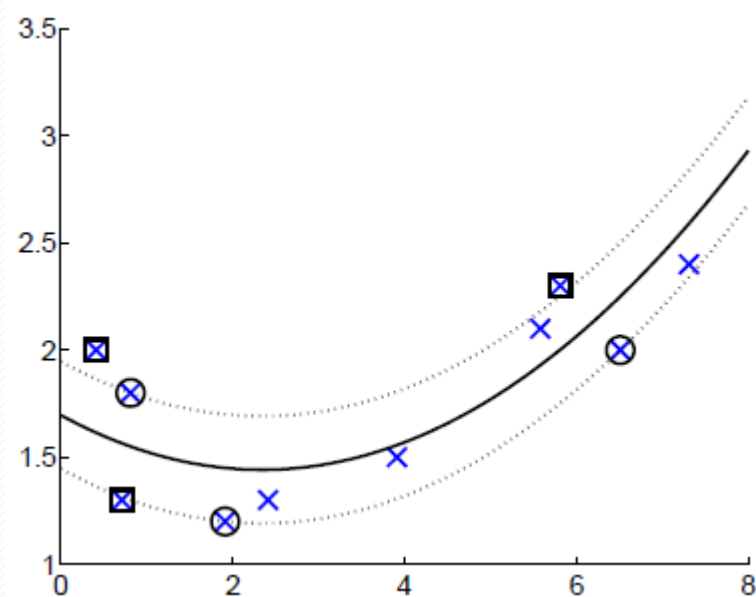
- $$\min \frac{1}{2} \|\mathbf{w}\|^2 + c \sum_t (\xi_+^t + \xi_-^t)$$
$$r^t - (\mathbf{w}^T \mathbf{x} + w_0) \leq \epsilon + \xi_+^t$$
$$(\mathbf{w}^T \mathbf{x} + w_0) - r^t \leq \epsilon + \xi_-^t$$
$$\xi_+^t, \xi_-^t \geq 0$$



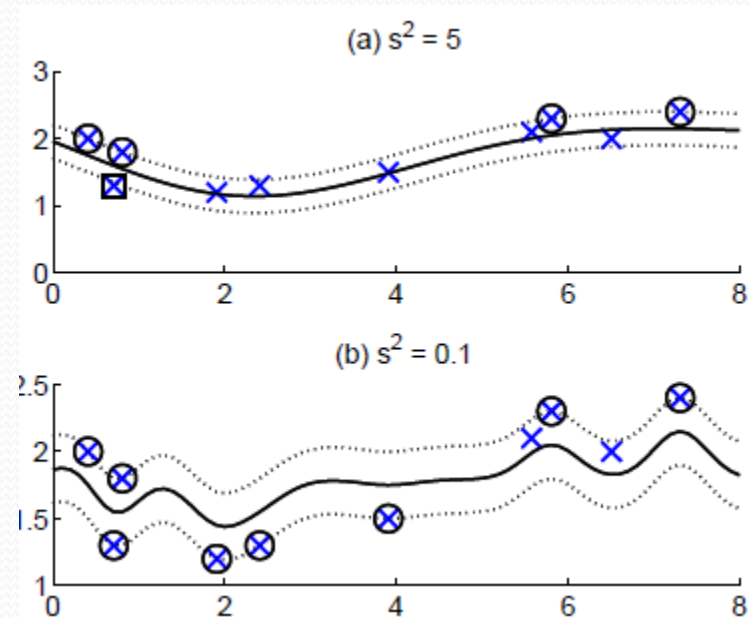


Kernel Regression

- Polynomial kernel



- Gaussian kernel



One-Class Kernel Machines

- Consider a sphere with center \mathbf{a} and radius R

$$\min R^2 + C \sum_t \xi^t$$

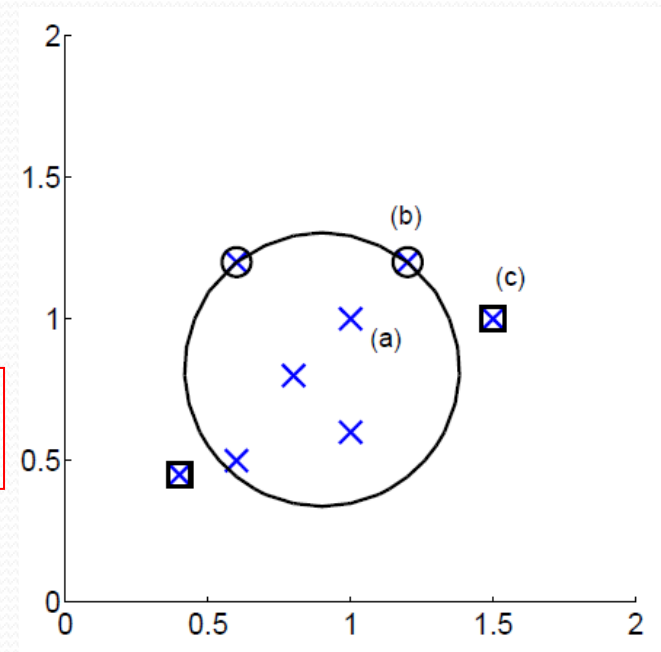
subject to

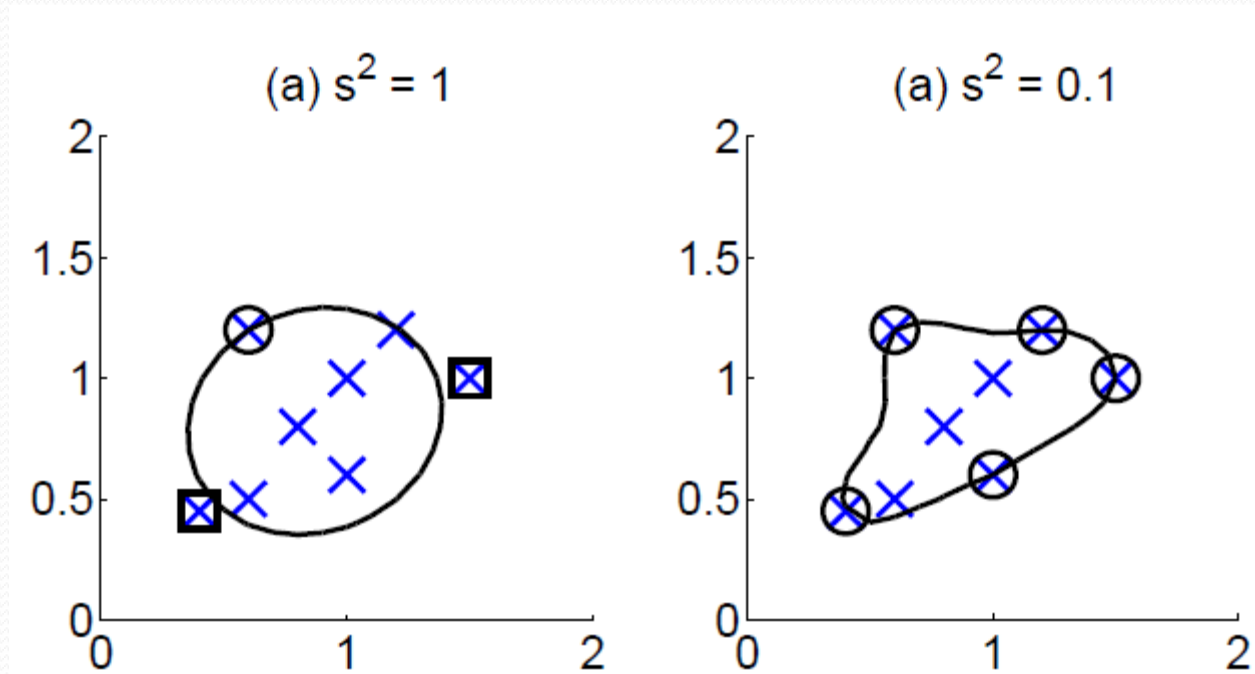
$$\|\mathbf{x}^t - \mathbf{a}\| \leq R^2 + \xi^t, \xi^t \geq 0$$

$$L_d = \sum_t \alpha^t \boxed{(\mathbf{x}^t)^T \mathbf{x}^s} - \sum_{t=1}^N \sum_s \alpha^t \alpha^s r^t r^s \boxed{(\mathbf{x}^t)^T \mathbf{x}^s}$$

subject to

$$0 \leq \alpha^t \leq C, \sum_t \alpha^t = 1$$





Kernel Dimensionality Reduction

- Kernel PCA does PCA on the kernel matrix (equal to canonical PCA with a linear kernel)
- Kernel LDA

