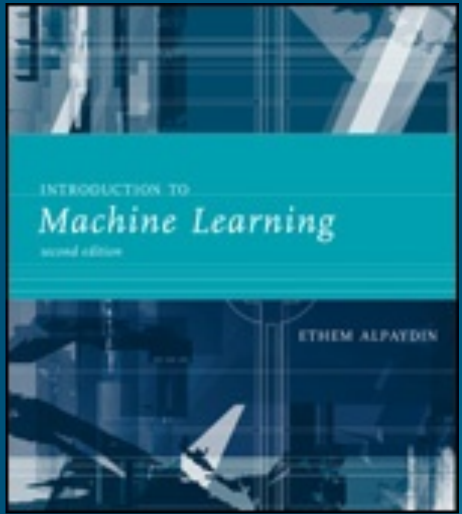


Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

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

- # google scholar results with "support vector machine"
- 39800, 26700, 26400 in 2014, 2013, 2012 respectively
- MLP/neural network 17200,16600,16700
- Bayesian network: 7500, 9050, 9250

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)

Review: Lagrange Multipliers

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} f(\mathbf{x})$$

$$\text{subject to } h_i(\mathbf{x}) = 0, \forall i = 1, \ldots, m$$

$$\text{subject to } g_i(\mathbf{x}) \leq 0, \forall i = 1, \ldots, n$$

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}}\, L(\mathbf{x}, \lambda, \mu) = \underset{\mathbf{x}}{\operatorname{argmin}}\, f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{i=1}^{n} \mu_i g_i(\mathbf{x}),$$

Review: Karush-Kuhn-Tucker Conditions (needed when we have inequality constraints)

- **Stationarity**

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \nabla_{\mathbf{x}} \lambda_i h_i(\mathbf{x}) + \sum_{i=1}^{n} \mu_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) = 0 \text{ (minimization)}$$

- **Equality constraints**

$$\nabla_{\lambda} f(\mathbf{x}) + \sum_{i=1}^{m} \nabla_{\lambda} \lambda_i h_i(\mathbf{x}) + \sum_{i=1}^{n} \mu_i \nabla_{\lambda} g_i(\mathbf{x}) = 0$$

- **Inequality constraints a.k.a. complementary slackness condition**

$$\mu_i g_i(\mathbf{x}) = 0, \forall i = 1, \ldots, n$$

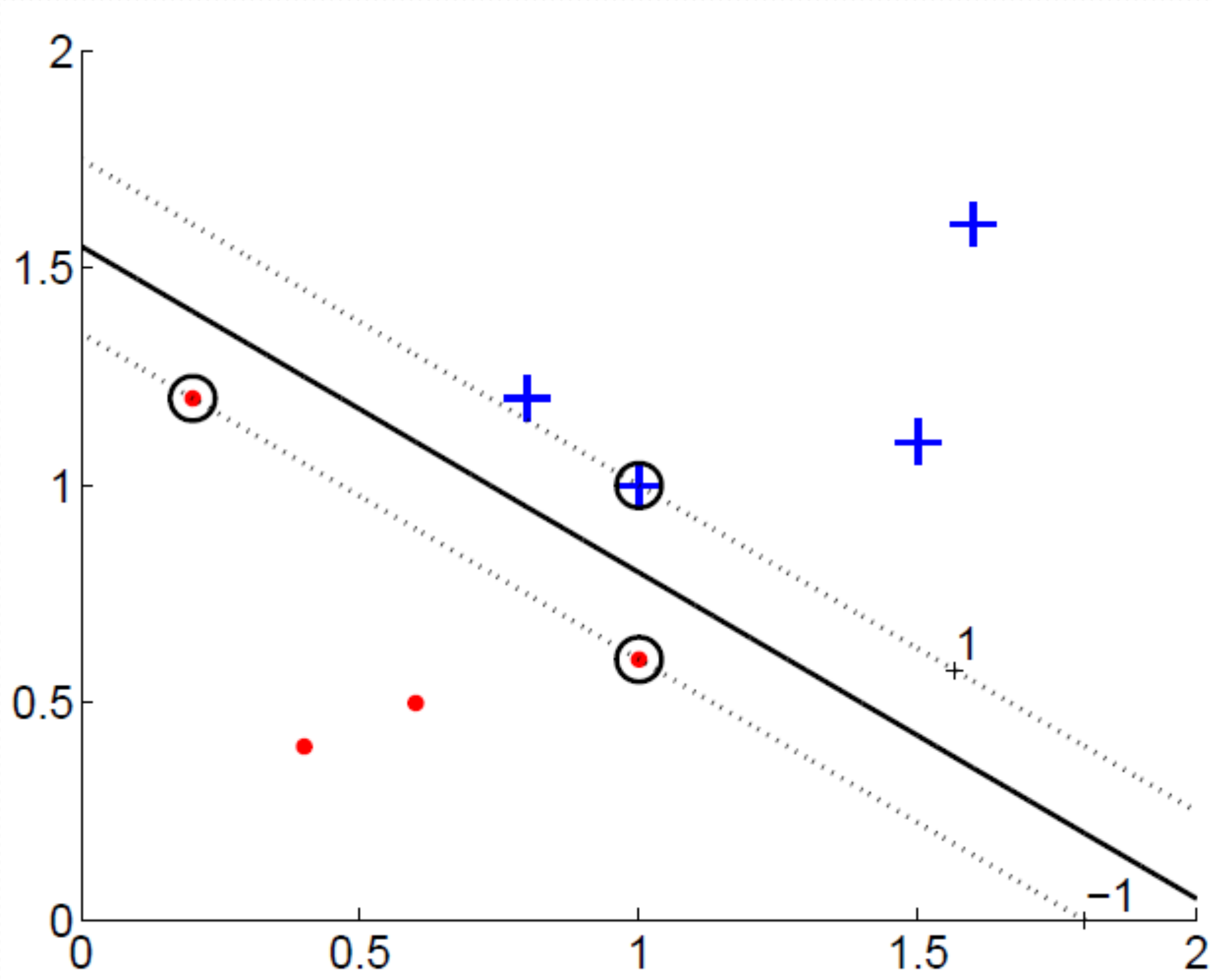
$$\mu_i \geq 0, \forall i = 1, \ldots, n$$

Margin

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

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

Margin



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

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

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

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

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

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

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

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

subject to $\sum_t \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$

minimize $\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x}.$

subject to $A\mathbf{x} \leq \mathbf{b}$

Solve using quadratic programming

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

Soft Margin Hyperplane

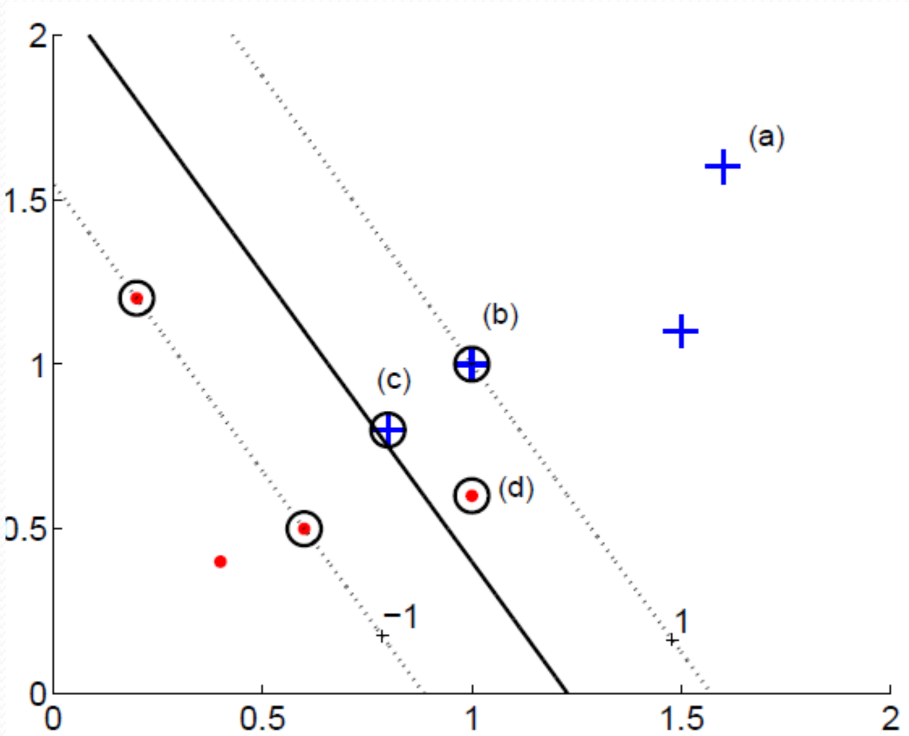
- Not linearly separable

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

- Soft error

$$\sum_t \xi^t$$

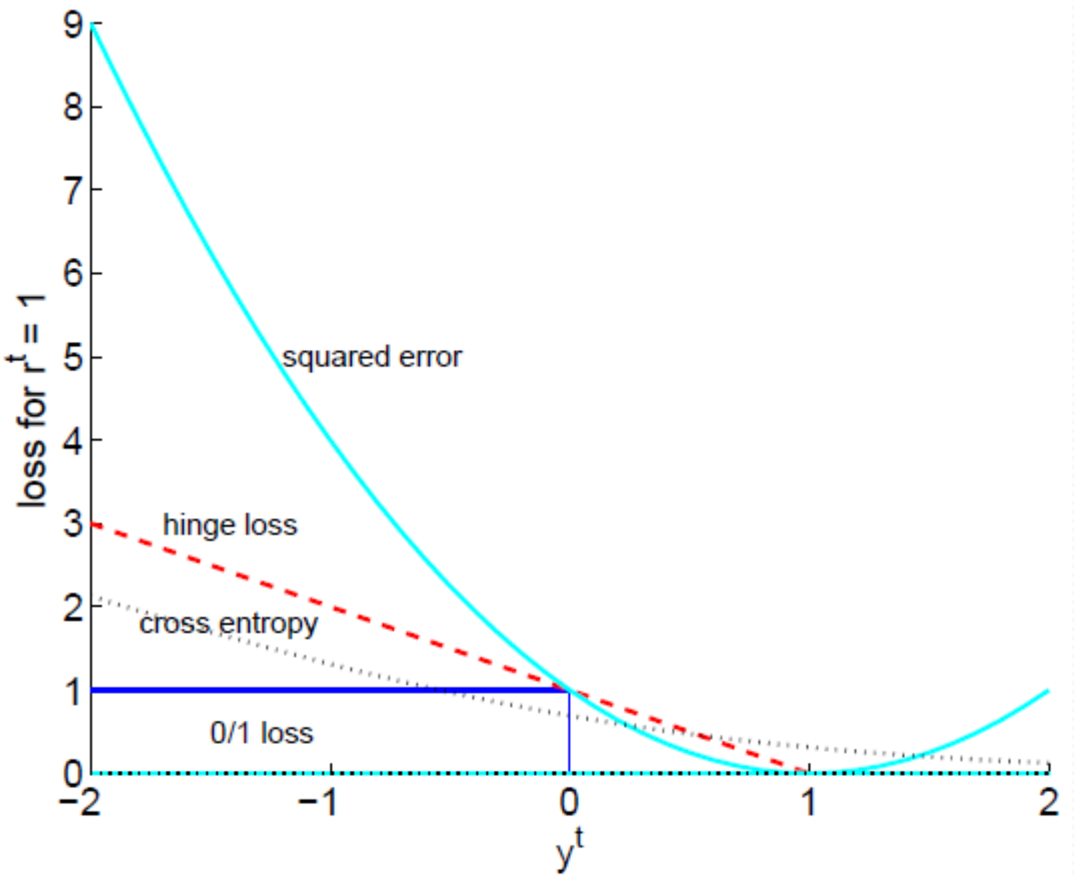
- New primal is



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

Hinge Loss

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



ν-SVM

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

subject to

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

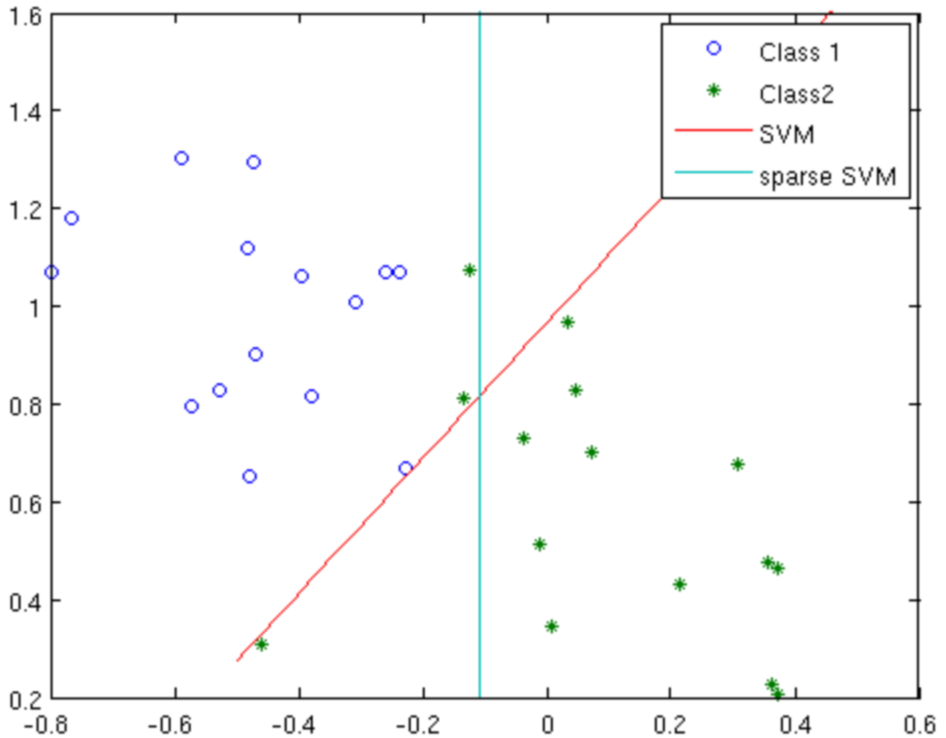
$$L_d = -\frac{1}{2}\sum_{t=1}^{N}\sum_s \alpha^t\alpha^s r^t r^s \left(x^t\right)^T x^s$$

subject to

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

ν controls the fraction of support vectors

Sparse SVM



$$\min_{m,b} \sum_{i=1}^{n} \text{hinge}(\text{label}_i \cdot (x_i^T m - b)) + \lambda_B \|m\|_1$$

the l1 term drives small coefficients to zero
http://cvxr.com/tfocs/demos/sparsesvm/

Kernel Trick

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

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

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

- The SVM solution

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

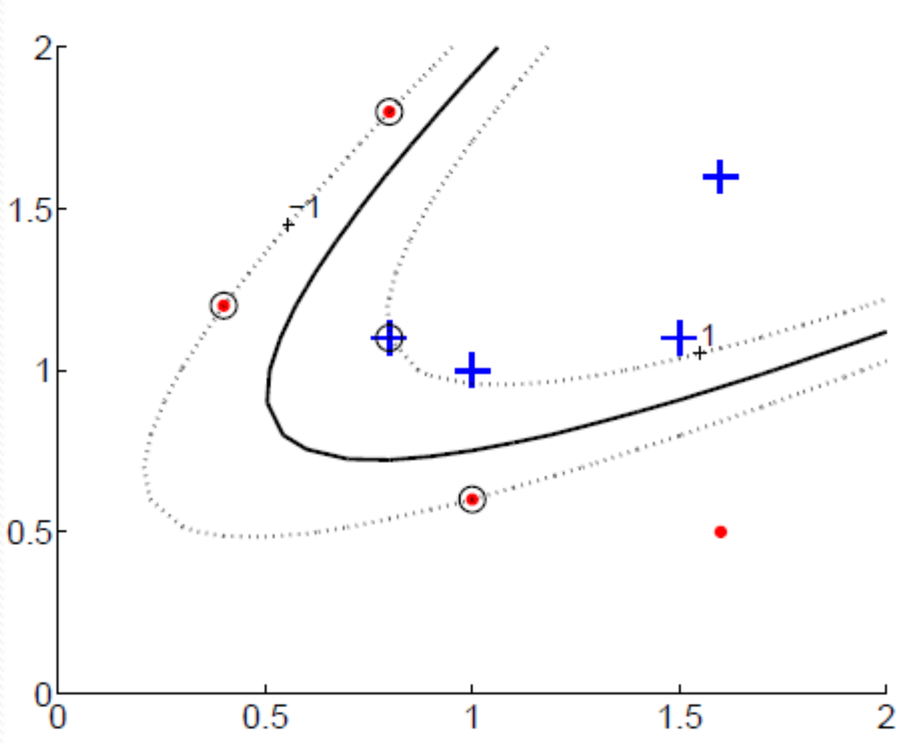
$$g(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_t \alpha^t r^t \boxed{\boldsymbol{\varphi}\left(\mathbf{x}^t\right)^T \boldsymbol{\varphi}(\mathbf{x})}$$

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

Vectorial Kernels

- Polynomials of degree q:

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



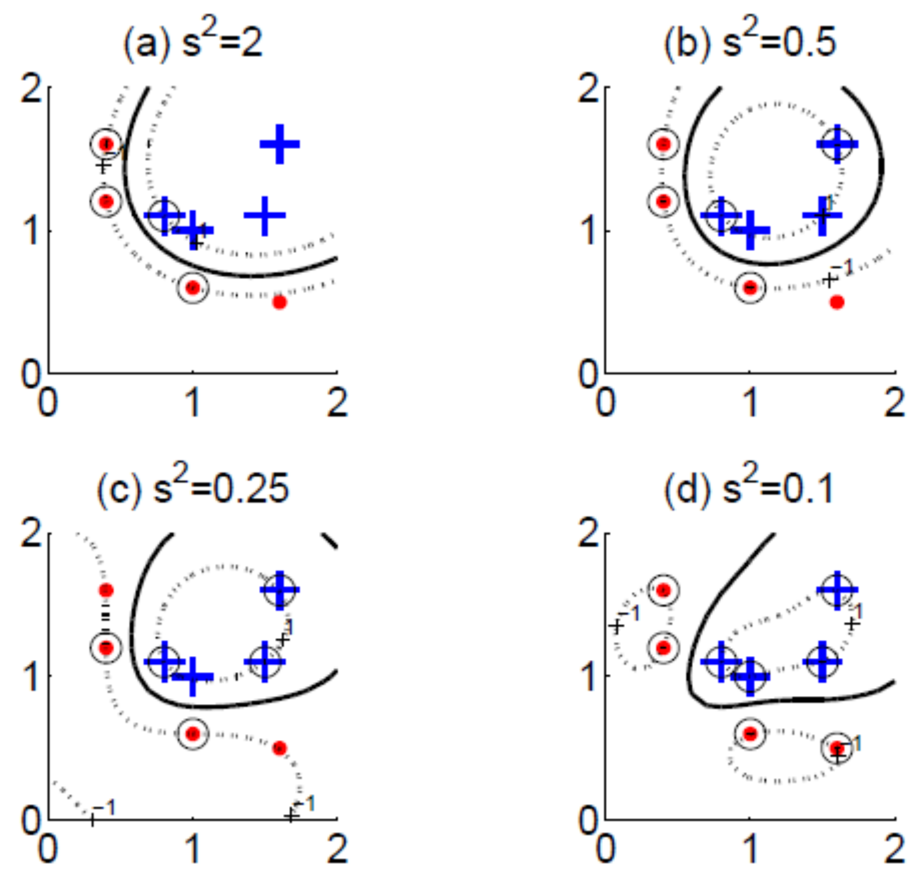
$$\begin{aligned}K(\mathbf{x},\mathbf{y}) &= \left(\mathbf{x}^T\mathbf{y}+1\right)^2 \\ &= \left(x_1y_1+x_2y_2+1\right)^2 \\ &= 1+2x_1y_1+2x_2y_2+2x_1x_2y_1y_2+x_1^2y_1^2+x_2^2y_2^2\end{aligned}$$

$$\phi(\mathbf{x})=\left[1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2\right]^T$$

Vectorial Kernels

- Radial-basis functions:

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



Defining kernels

- Kernel "engineering"
- Defining good measures of similarity
- String kernels, graph kernels, image kernels, ...
- Empirical kernel map: Define a set of templates $\boldsymbol{m}_i$ and score function $s(\boldsymbol{x},\boldsymbol{m}_i)$

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

  and

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

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

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

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

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

Learn α s and kernel weights η from data

- Localized kernel combination

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

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

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

zt: class index

SVM for Regression

- Use a linear model (possibly kernelized)

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

- Use the ϵ-sensitive error function

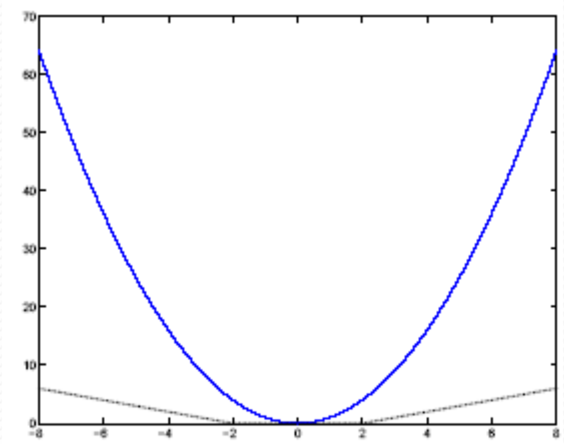
$$e_\varepsilon\left(r^t, f(\mathbf{x}^t)\right)=\begin{cases}0 & \text{if } \left|r^t - f(\mathbf{x}^t)\right| < \varepsilon \\ \left|r^t - f(\mathbf{x}^t)\right| - \varepsilon & \text{otherwise}\end{cases}$$

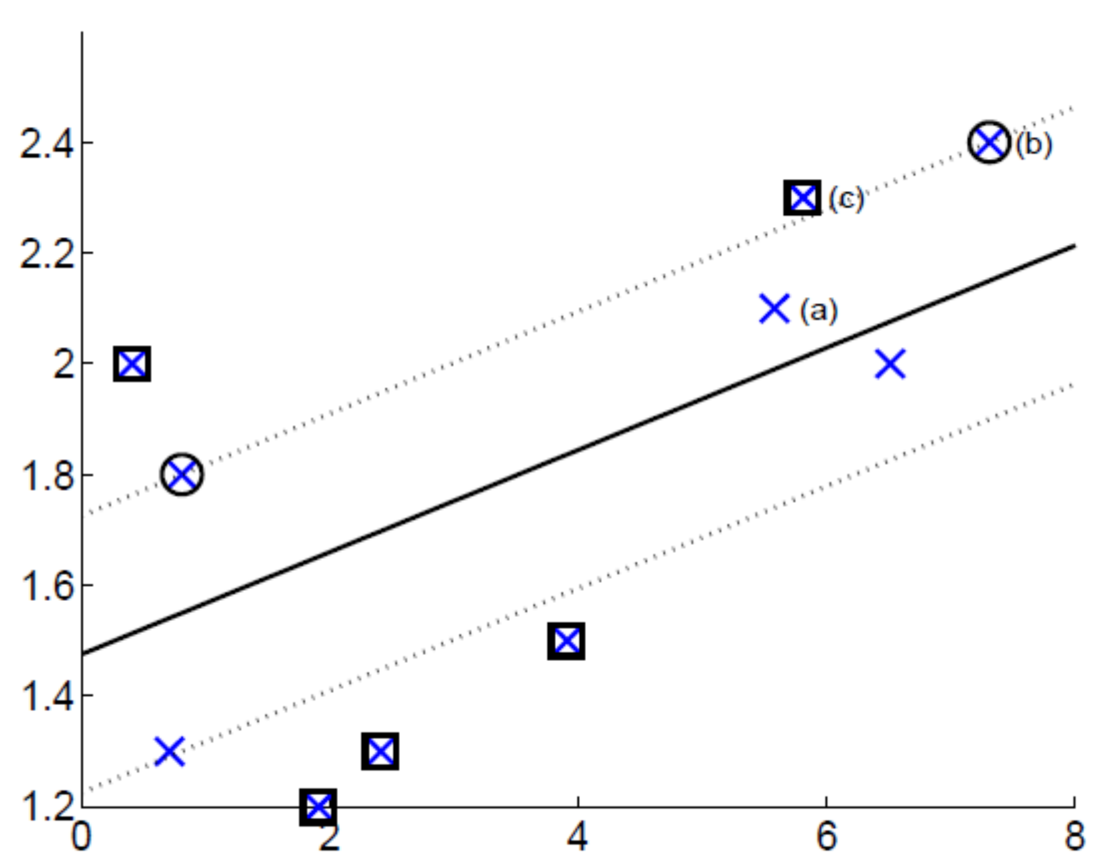
- $$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \left(\xi_+^t + \xi_-^t\right)$$

$$r^t - \left(\mathbf{w}^T\mathbf{x} + w_0\right) \le \varepsilon + \xi_+^t$$

$$\left(\mathbf{w}^T\mathbf{x} + w_0\right) - r^t \le \varepsilon + \xi_-^t$$

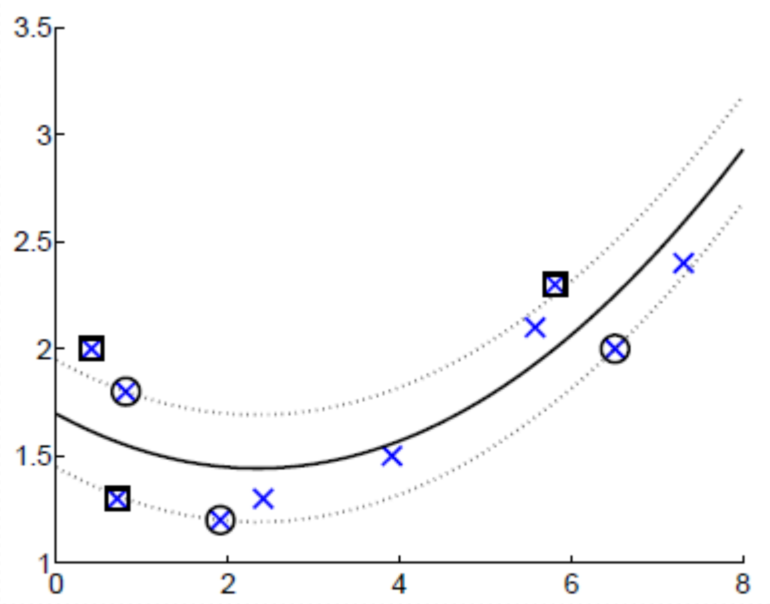
$$\xi_+^t, \xi_-^t \ge 0$$

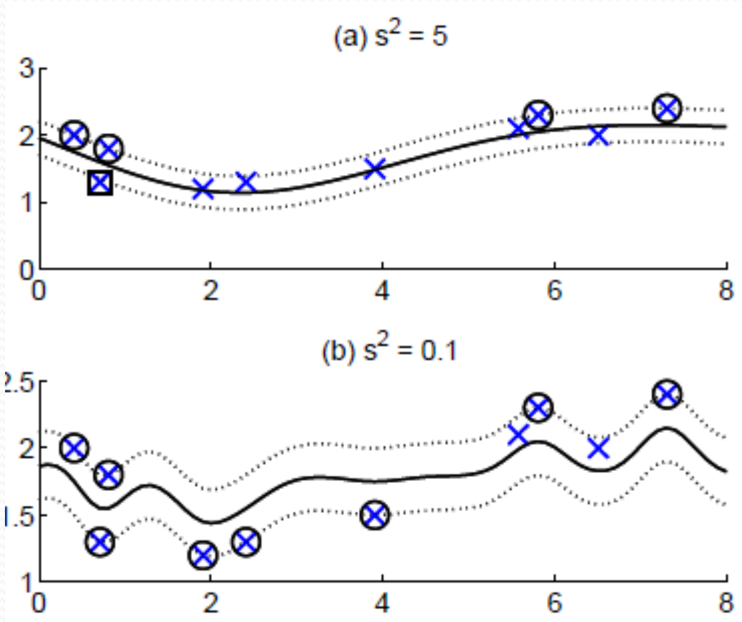



2.4
2.2
2
1.8
1.6
1.4
1.2
0
2
4
6
8
(a)
(b)
(c)

Kernel Regression

- Polynomial kernel
- Gaussian kernel





One-Class Kernel Machines

- Consider a sphere with center ***a*** and radius *R*

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

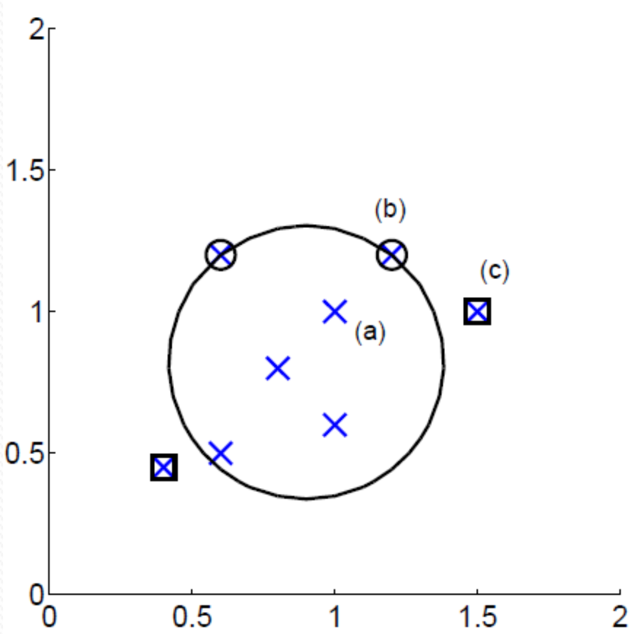
subject to

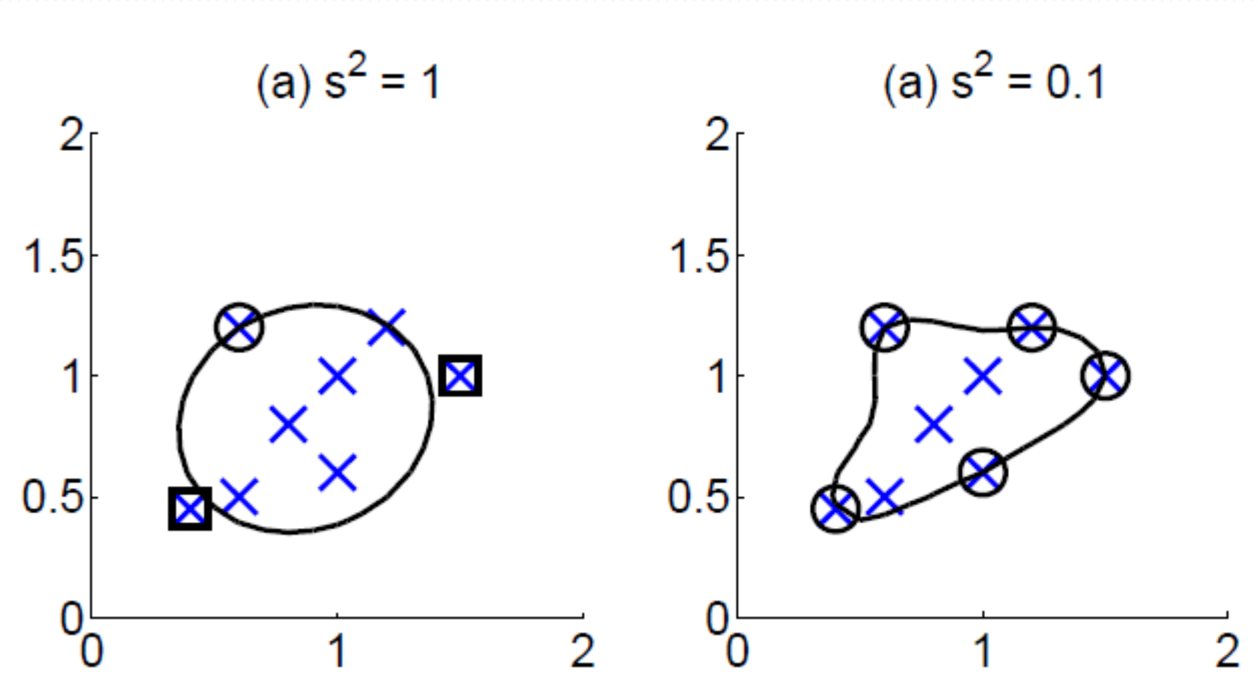
$$\left\|\mathbf{x}^t - a\right\| \le R^2 + \xi^t, \xi^t \ge 0$$

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

subject to

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




(a) s2 = 1
(a) s2 = 0.1
2
1.5
1
0.5
0
0
1
2

Kernel Dimensionality Reduction

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