

# Support Vector Machines

Chapter 9, Learning from Data

1

## Introduction (1)

- Universal constructive learning procedure
  - ◆ Based on statistical learning theory (Vapnik, 1995)
  - ◆ Used to learn a variety of representations
    - neural nets, radial basis functions, splines, polynomial estimators
  - ◆ Provides a new form of parameterization of functions.
  - ◆ Provides a meaningful characterization of the function's complexity that is *independent* of the problem's dimensionality.

3

## Contents

- 9.1 Optimal Separating Hyperplane
- 9.2 High-Dimensional Mapping and Inner Product Kernels
- 9.3 Support Vector Machine for Classification
- 9.4 Support Vector Machine for Regression
- 9.5 Summary

2

## Introduction (2)

- Motivation
  - ◆ For nonlinear models
    - 1) VC-dimension cannot be accurately estimated.
    - 2) Implementation of structural risk minimization leads to nonlinear optimization.
  - ◆ For linear models of large multivariate problems
    - The curse of dimensionality

4

## Introduction (3)

- SVM overcomes two problems

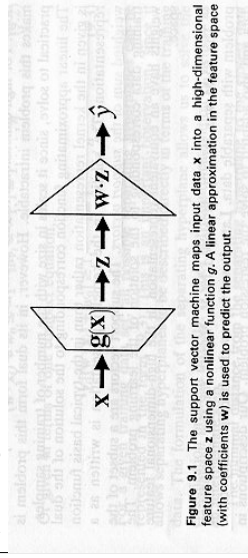
- 1) Conceptual problem
  - How to control the complexity of the set of approximating functions in a high-dimensional space in order to provide good generalization ability.
  - Using penalized linear estimators with a large number of basis functions.
- 2) Computational problem
  - How to perform numerical optimization in a high-dimensional space.
  - Using the dual kernel representation of linear functions.

5

## Introduction (5)

2. Input samples mapped onto a very high-dimensional space using a set of nonlinear basis functions defined a priori

- ◆ In ordinary learning problem, feature space is usually made for the purpose of reduction of complexity.



7

## Introduction (4)

- SVM combines four distinct concepts

1. New implementation of the SRM inductive principle.
  - ◆ SVM can analytically estimate the VC-dim.
    - Minimize the VC-dim, keeping the value of the empirical risk nearly zero.
  - ◆ Ordinary SRM implementation.
    - About each  $VC_1 < VC_2 < \dots < VC_n$  models,
    - Minimize each empirical risk.
    - Choose the best model of which guaranteed risk is small.

6

## Introduction (6)

3. Linear functions with constraints on complexity used to approximate or discriminate the input samples in the high-dimensional space

- ◆ Accurate estimates for model complexity can be obtained for linear estimators.
- ◆ The drawbacks of nonlinear estimators
  - lack of complexity measures
  - lack of optimization approaches

8

## Introduction (7)

4. Duality theory of optimization used to make estimation of model parameters in a high-dimensional feature space computationally tractable.

- ◆ In SVM, a quadratic programming is used for optimization.
- ◆ In original problem, large number of parameter must be estimated, which makes the problem intractable.
- ◆ The size of dual problem scales in size with the number of training samples.
- ◆ The solution of dual problem becomes the support vectors' weights

9

## 9.1. Optimal Separating Hyperplane (1)

### • Separating hyperplane

- ◆ A linear function that is capable of separating the training data

$$D(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + w_0$$
$$y_i [(\mathbf{w} \cdot \mathbf{x}_i) + w_0] \geq 1, \quad i = 1, \dots, n \quad (9.3)$$

- ◆ Note that when linearly separable case,  $\mathbf{w}$ ,  $w_0$  can be scaled so that next condition holds.

$$(\mathbf{w} \cdot \mathbf{x}) + w_0 \geq +1 \quad \text{if } y_i = +1$$

$$(\mathbf{w} \cdot \mathbf{x}) + w_0 \leq -1 \quad \text{if } y_i = -1, \quad i = 1, \dots, n$$

10

## 9.1. Optimal Separating Hyperplane (2)

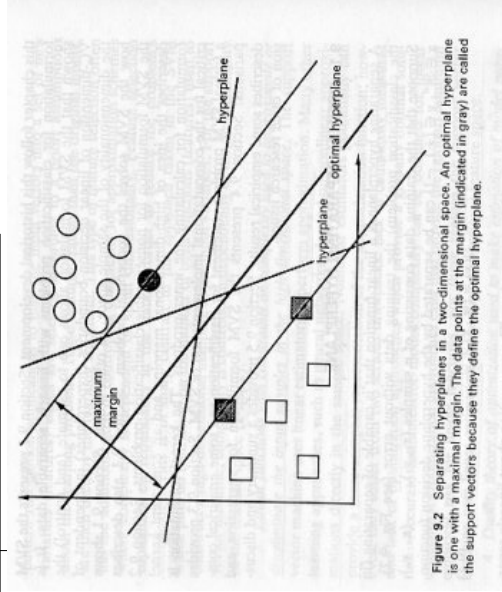


Figure 9.2 Separating hyperplanes in a two-dimensional space. An optimal hyperplane is one with a maximal margin. The data points at the margin (indicated in grey) are called the support vectors because they define the optimal hyperplane.

11

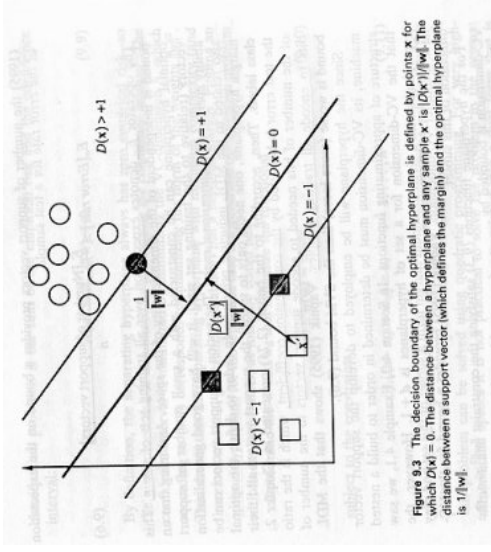
## 9.1. Optimal Separating Hyperplane (3)

- Margin:  $\tau$
- ◆ Minimal distance from the separating hyperplane to the closest data
- Optimal separating hyperplane (s.h.)
- ◆ When the margin is the maximum size.
- Distance between s.h. and a sample  $\mathbf{x}'$   
$$\frac{|D(\mathbf{x}')|}{\|\mathbf{w}\|}$$
- All patterns obey the inequality

$$\frac{y_k D(\mathbf{x}_k)}{\|\mathbf{w}\|} \geq \tau, \quad k = 1, \dots, n$$

12

## 9.1. Optimal Separating Hyperplane (4)



13

## 9.1. Optimal Separating Hyperplane (5)

- Maximizing the margin = Minimizing  $||\mathbf{w}'||$

$$\tau = \frac{1}{||\mathbf{w}'||}$$

- Support Vector
  - ◆ The data that exist at the margin (when the equality condition of (9.3) is satisfied).
  - ◆ Dimensionality independent generalization error bound  $E_n[\text{Error rate}] \leq \frac{E_n[\text{Number of support vectors}]}{n}$
  - ◆ Number of SVs is much smaller than number of patterns in most cases.

14

## 9.1. Optimal Separating Hyperplane (6)

- The VC-dim of hyperplane of (9.3) satisfying  $c \geq ||\mathbf{w}'||^2$

$$h \leq \min(r^2 c, d) + 1$$

- SRM implementation
  - ◆ S.h. always has zero empirical risk
  - ◆  $\Phi$  is minimized by minimizing the VC-dim  $h$ , which corresponds to minimizing  $||\mathbf{w}'||^2$

15

## 9.1. Optimal Separating Hyperplane (7)

- Quadratic optimization problem

$$\text{minimize}_{\mathbf{w}} \quad \eta(\mathbf{w}) = \frac{1}{2} ||\mathbf{w}'||^2$$

subject to

$$y_i [(\mathbf{w} \cdot \mathbf{x}_i) + w_0] \geq 1, \quad i = 1, \dots, n$$

- ◆ Minimizing quadratic function with linear constraints.
- ◆ The solution consists of  $d+1$  parameters.

16

## 9.1. Optimal Separating Hyperplane (8)

- Dual problem
  - ◆ The solution consists of  $n$  parameters.
  - ◆ Convertible if cost and constraint are convex.
- Step 1 of conversion
  - ◆ Construct *Lagrangian* function

$$Q(\mathbf{w}, w_0, \alpha) = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^n \alpha_i \{y_i [(\mathbf{w} \cdot \mathbf{x}_i) + w_0] - 1\}$$

- Step 2 of conversion
  - ◆ Using the optimal condition

17

## 9.1. Optimal Separating Hyperplane (9)

$$\frac{\partial Q(\mathbf{w}^*, w_0^*, \alpha^*)}{\partial w_0} = 0 \quad (9.13)$$

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i, \quad \alpha_i^* \geq 0, \quad i = 1, \dots, n \quad (9.16)$$

$$\frac{\partial Q(\mathbf{w}^*, w_0^*, \alpha^*)}{\partial \mathbf{w}} = 0 \quad (9.14)$$

$$\sum_{i=1}^n \alpha_i^* y_i = 0, \quad \alpha_i^* \geq 0, \quad i = 1, \dots, n \quad (9.15)$$

- ◆ Kuhn-Tucker theorem
    - The data corresponding nonzero  $\alpha_i^*$  are support vectors.
- $$\alpha^* [y_i(\mathbf{w}^* \cdot \mathbf{x}_i + w_0^*) - 1] = 0, \quad i = 1, \dots, n$$

18

## 9.1. Optimal Separating Hyperplane (10)

- Dual problem
  - ◆ The solution consists of  $n$  parameters.
  - ◆ Convertible if cost and constraint are convex.
- Step 1 of conversion
  - ◆ Construct *Lagrangian* function

$$Q(\alpha) = -\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^n \alpha_i$$

- Step 2 of conversion
  - ◆ Using the optimal condition

17

## 9.1. Optimal Separating Hyperplane (11)

- The resulting equation s.h.
  - ◆  $D(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + w_0^*$
  - ◆  $y_s [(\mathbf{w}^* \cdot \mathbf{x}_s) + w_0^*] = 1$
  - ◆  $w_0^* = y_s - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_s)$

19

## 9.1. Optimal Separating Hyperplane (11)

- The resulting equation s.h.
  - ◆  $D(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + w_0^*$
  - ◆  $y_s [(\mathbf{w}^* \cdot \mathbf{x}_s) + w_0^*] = 1$
  - ◆  $w_0^* = y_s - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_s)$

20

## 9.1. Optimal Separating Hyperplane (12)

- Nonseparable problem
  - ◆ Certain data point where doesn't satisfy (9.3) exists.
- Introducing positive slack variables  $\xi_i$

$$y_i[(\mathbf{w} \cdot \mathbf{x}_i) + w_0] \geq 1 - \xi_i \quad (9.25)$$

- Optimization problem

$$Q(\mathbf{w}) = \sum_{i=1}^n I(\xi_i > 0) \quad (9.26)$$

- ◆ (9.26) is combinatorial optimization and very difficult because of the nonlinearity.

21

## 9.1. Optimal Separating Hyperplane (13)

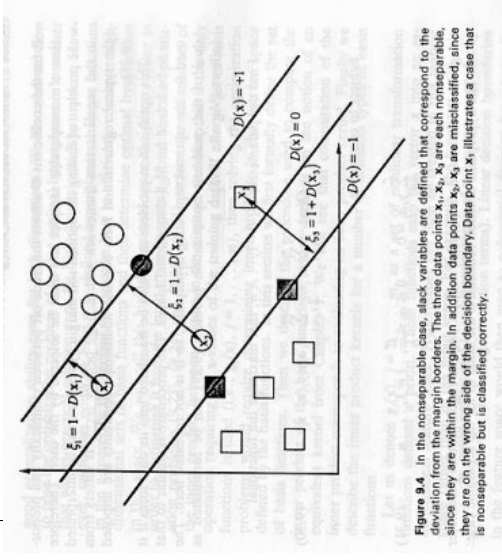


Figure 9.4 In the nonseparable case, slack variables are defined that correspond to the deviation from the margin borders. The three data points  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ ,  $\mathbf{x}_3$  are each nonseparable since they are within the margin. In addition data points  $\mathbf{x}_4$ ,  $\mathbf{x}_5$  are misclassified, since they are on the wrong side of the decision boundary. Data point  $\mathbf{x}_k$  illustrates a case that is nonseparable but is classified correctly.

22

## 9.1. Optimal Separating Hyperplane (14)

- Approximation of (9.26) is used

$$Q(\xi) = \sum_{i=1}^n \xi_i^p \quad (9.27)$$

- QP (when  $p=1$ )

$$\text{minimize}_{\mathbf{w}} \quad C \sum_{i=1}^n \xi_i + \frac{1}{2} \|\mathbf{w}\|^2$$

subject to

$$y_i[(\mathbf{w} \cdot \mathbf{x}_i) + w_0] \geq 1 - \xi_i$$

23

## 9.1. Optimal Separating Hyperplane (15)

- Dual Problem

$$\text{maximize}_{\alpha} Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^n y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \dots, n$$

- Resulting equation of s.h.

$$D(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + w_0^*$$

24

## 9.2. High-Dimensional Mapping and Inner Product Kernels (1)

- Complexity of optimal hyperplanes are dimensionality independent.
- Dual problem only needs the inner product between vectors in feature space.
- Nonlinear transformation function  $\mathbf{g}(\mathbf{x})=[g_1(\mathbf{x}), \dots, g_m(\mathbf{x})]$ .
  - ◆ Even for a small problem the feature space can be very large.

25

## 9.2. High-Dimensional Mapping and Inner Product Kernels (2)

- Example
  - ◆  $g_j(\mathbf{x}), j=1, \dots, m$  are polynomial terms of  $\mathbf{x}$  up to 3rd-order
  - ◆ Feature space has 16 dimension.

$$\begin{aligned}
 g_1(x_1, x_2) &= 1 \\
 g_2(x_1, x_2) &= x_1 \\
 g_3(x_1, x_2) &= x_1^2 \\
 g_4(x_1, x_2) &= x_1^3 \\
 g_5(x_1, x_2) &= x_2 \\
 g_6(x_1, x_2) &= x_2^2 \\
 g_7(x_1, x_2) &= x_2^3 \\
 g_8(x_1, x_2) &= x_1x_2 \\
 g_9(x_1, x_2) &= x_1^2x_2 \\
 g_{10}(x_1, x_2) &= x_1x_2^2 \\
 g_{11}(x_1, x_2) &= x_1^3x_2 \\
 g_{12}(x_1, x_2) &= x_1^2x_2^2 \\
 g_{13}(x_1, x_2) &= x_1x_2^3 \\
 g_{14}(x_1, x_2) &= x_1^3x_2^2 \\
 g_{15}(x_1, x_2) &= x_1^2x_2^3 \\
 g_{16}(x_1, x_2) &= x_1^3x_2^3
 \end{aligned}$$

26

## 9.2. High-Dimensional Mapping and Inner Product Kernels (3)

- Decision function
 
$$D(\mathbf{x}) = \sum_{j=1}^m w_j g_j(\mathbf{x})$$
- Dual form of decision function

$$D(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i H(\mathbf{x}_i, \mathbf{x})$$

- ◆ where

$$\begin{aligned}
 H(\mathbf{x}, \mathbf{x}') &= \mathbf{g}(\mathbf{x})^T \mathbf{g}(\mathbf{x}') \\
 &= \sum_{j=1}^m g_j(\mathbf{x}) g_j(\mathbf{x}')
 \end{aligned}$$

27

## 9.2. High-Dimensional Mapping and Inner Product Kernels (4)

- Any symmetric function  $H(\mathbf{x}, \mathbf{x}')$  satisfying the Mercer's condition can be used as an inner product.

$$\iint H(\mathbf{x}, \mathbf{x}') \varphi(\mathbf{x}) \varphi(\mathbf{x}') d\mathbf{x} d\mathbf{x}' > 0 \quad \text{for all } \varphi \neq 0, \int \varphi^2(\mathbf{x}) d\mathbf{x} < \infty$$

- ◆ Polynomials of degree  $q$ :

$$H(\mathbf{x}, \mathbf{x}') = [(\mathbf{x} \cdot \mathbf{x}') + 1]^q$$

- ◆ RBF with width  $\sigma$ :

$$H(\mathbf{x}, \mathbf{x}') = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right\}$$

28

## 9.2. High-Dimensional Mapping and Inner Product Kernels (4)

- ◆ Neural network with parameters  $v, a$  satisfying the Mercer's theorem:

$$H(\mathbf{x}, \mathbf{x}') = \tanh(v(\mathbf{x} \cdot \mathbf{x}') + a)$$

29

## 9.3. Support Vector Machine for Classification (1)

- Decision function for nonseparable data

$$D(\mathbf{x}) = \sum_{i=1}^n \alpha_i^* y_i H(\mathbf{x}_i, \mathbf{x})$$

- Dual problem

$$\text{maximize}_{\alpha} Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j H(\mathbf{x}_i \cdot \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^n y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \dots, n$$

30

### Example 9.1

- The exclusive-or (XOR) problem
- The inner product kernel for polynomial

$$H(\mathbf{x}, \mathbf{x}') = [(\mathbf{x} \cdot \mathbf{x}') + 1]^2$$

- The set of basis function
- Solve the dual problem when  $C = \infty$

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

31

### Example 9.1

Index $i$	$\mathbf{x}$	$y$
1	(1, 1)	1
2	(1, -1)	-1
3	(-1, -1)	1
4	(-1, 1)	-1

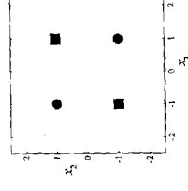


Figure 9.5 The exclusive-or data set. The problem is not linearly separable in the input space.

32



### Example 9.1

$$\text{maximize } Q(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} \sum_{i,j=1}^4 \alpha_i \alpha_j y_i y_j h_{ij}$$

subject to

$$\sum_{i=1}^4 y_i \alpha_i = \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0$$

$$0 \leq \alpha_1$$

$$0 \leq \alpha_2$$

$$0 \leq \alpha_3$$

$$0 \leq \alpha_4$$

### Example 9.1

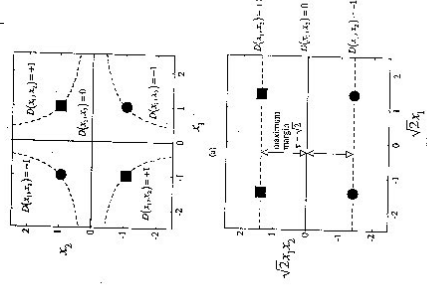


Figure 9.6 Decision function determined by the support vector machine with 5 failure points. (a) In the two-dimensional input space, the decision function is nonlinear. (b) In the six-dimensional feature space, the decision function is linear with maximum margin.

### Example 9.1

- Inner product model

$$H = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

- The solution to this optimization problem

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.125$$

- The decision function in the inner product representation

$$D(\mathbf{x}) = \sum_{i=1}^4 \alpha_i^* y_i H(\mathbf{x}_i, \mathbf{x}) = (0.125) \sum_{i=1}^4 y_i [(\mathbf{x}_i \cdot \mathbf{x}) + 1]^2,$$

## 9.4. Support Vector Machine for Regression (1)

- A function linear in parameters is used to approximate the regression in the feature space.

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^m w_j g_j(\mathbf{x})$$

- A special loss function (*Vapnik's loss function*)

$$L_{\epsilon}(y, f(\mathbf{x}, \mathbf{w})) = \begin{cases} e & \text{if } |y - f(\mathbf{x}, \mathbf{w})| \leq \epsilon \\ |y - f(\mathbf{x}, \mathbf{w})| & \text{otherwise} \end{cases}$$

- More relaxed assumption about noise than  $L_2$  loss function.
- $\epsilon$  controls the width of the insensitive zone.

## 9.4. Support Vector Machine for Regression (2)

- Quadratic Problem

$$\underset{\mathbf{w}}{\text{minimize}} \quad \frac{C}{n} \left( \sum_{i=1}^m \xi_i + \sum_{i=1}^m \xi'_i \right) + \frac{1}{2} (\mathbf{w}^T \cdot \mathbf{w})$$

subject to

$$y_i - \sum_{j=1}^m w_j g_j(\mathbf{x}_i) \leq e + \xi'_i$$

$$\sum_{i=1}^m w_j g_j(\mathbf{x}_i) - y_i \leq e + \xi_i$$

$$\xi'_i \geq 0$$

$$\xi_i \geq 0$$

37

## 9.4. Support Vector Machine for Regression (3)

- Dual Problem

$$\underset{\alpha, \beta}{\text{maximize}} \quad Q(\alpha, \beta) = -e \sum_{i=1}^n (\alpha_i + \beta_i) + \sum_{i=1}^n y_i (\alpha_i - \beta_i)$$

$$- \frac{1}{2} \sum_{i, j=1}^n (\alpha_i - \beta_i) (\alpha_j - \beta_j) H(\mathbf{x}_i, \mathbf{x}_j)$$

subject to

$$\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i, \quad 0 \leq \alpha_i \leq \frac{C}{n}, \quad 0 \leq \beta_i \leq \frac{C}{n}, \quad i = 1, \dots, n$$

- The resulting regression function

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i^* - \beta_i^*) H(\mathbf{x}_i, \mathbf{x})$$

38

## 9.5. Summary

- SVM's four principles.
  - ◆ Direct solution rather than indirect via density estimation
  - ◆ Dimension independent complexity control
  - ◆ Nonlinear feature selection
    - Directly incorporated in parameter optimization.
  - ◆ Implementation of an inductive principle

39