CpE 646 Pattern Recognition and Classification

Prof. Hong Man

Department of Electrical and Computer Engineering
Stevens Institute of Technology

STEVENS Institute of Technology

Visual Information Environment Laboratory

Chapter 5: Linear Discriminant Functions

Chapter 5 (Section 5.1 – 5.5, 5.11):

- Introduction
- Linear Discriminant Functions and Decisions Surfaces
- Generalized Linear Discriminant Functions
- The Two-Category Linear Separable Case
- Minimizing the Perceptron Criterion Function
- Support Vector Machine

Introduction

- In chapter 3, the underlying probability densities were known (or given)
- The training sample was used to estimate the parameters of these probability densities (ML, MAP estimations)
- In this chapter, we only know the proper forms for the *discriminant functions*, the training samples will be used to estimate the parameters of the discriminant function.
- We focus on linear discriminant functions, which are either linear in the components of $\boldsymbol{x}$, or linear in some functions of $\boldsymbol{x}$.
- They may not be optimal, but they are very simple to use

Linear Discriminant Functions and Decisions Surfaces

- Definition of a linear discriminant function
  - It is a function that is a linear combination of the components of $\boldsymbol{x}$

$$g(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + w_0 \qquad (1)$$

    where $\boldsymbol{w}$ is the weight vector and w_0 the bias
  - In the case of c categories, there will be c such discriminant functions.

Linear Discriminant Functions and Decisions Surfaces

- A two-category classifier with a discriminant function of the form (1) uses the following rule:
  - Decide ω_1 if $g(\mathbf{x}) > 0$ and ω_2 if $g(\mathbf{x}) < 0$

    It is equivalent to

    Decide ω_1 if $\mathbf{w}^t\mathbf{x} > -w_0$ and ω_2 otherwise
  - If $g(\mathbf{x}) = 0 \Rightarrow \mathbf{x}$ is assigned to either class
  - The equation $g(\mathbf{x}) = 0$ defines the decision surface that separates points assigned to the category ω_1 from points assigned to the category ω_2
  - When $g(\mathbf{x})$ is linear, the decision surface is a hyperplane denoted as H.

Linear Discriminant Functions and Decisions Surfaces

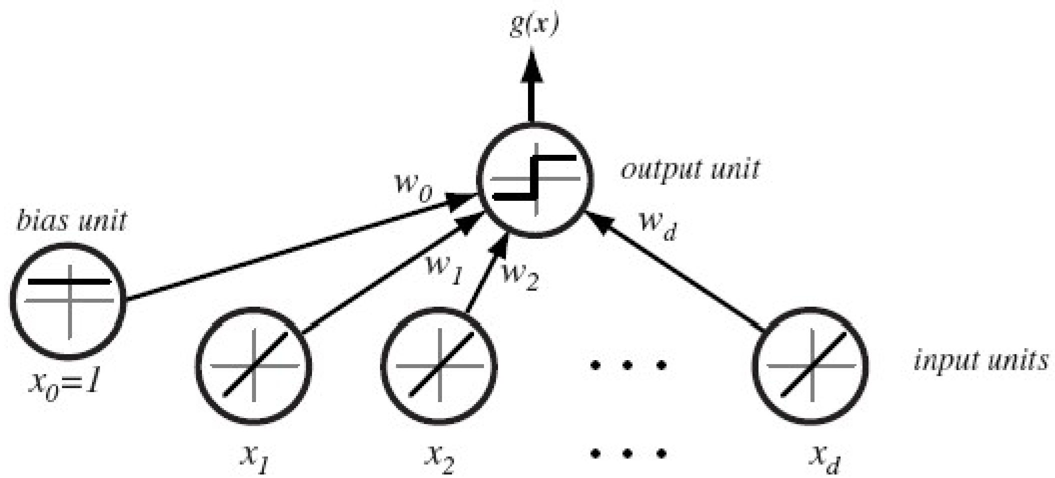

FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show in each unit its effective input-output function. Thus each of the d input units is linear, emitting exactly the value of its corresponding feature value. The single bias unit unit always emits the constant value 1.0. The single output unit emits a +1 if $\mathbf{w}^t\mathbf{x} + w_0 > 0$ or a −1 otherwise. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Discriminant Functions and Decisions Surfaces

– $g(\boldsymbol{x})$ also provides an algebraic measure of the distance from $\boldsymbol{x}$ to the hyperplane

$$x = x_p + \frac{r \cdot w}{\|w\|}$$

where $\boldsymbol{x_p}$ is the normal projection of $\boldsymbol{x}$ onto H, and r is the distance between $\boldsymbol{x}$ and H.

since $g(x_p) = w^t x_p + \omega_0 = 0$ and $w^t \cdot w = \|w\|^2$

$$g(x) = w^t x + \omega_0 = w^t \left(x_p + \frac{r \cdot w}{\|w\|} \right) + \omega_0 = r \| w \|,$$

or $\quad r = \dfrac{g(x)}{\|w\|}$

Linear Discriminant Functions and Decisions Surfaces

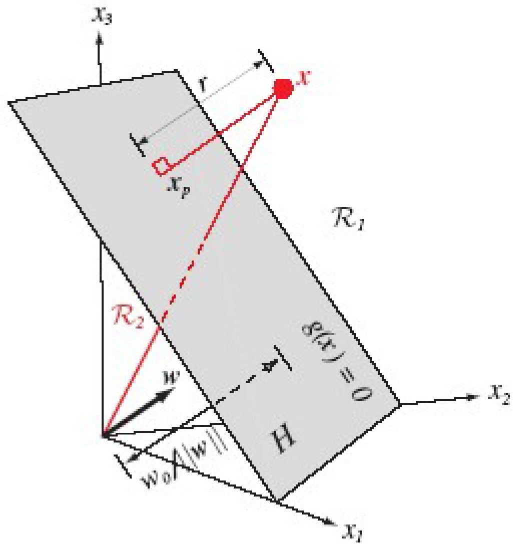

FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Discriminant Functions and Decisions Surfaces

– In conclusion, a linear discriminant function divides the feature space by a hyperplane decision surface
– The orientation of the surface is determined by the normal vector $\boldsymbol{w}$ and the location of the surface is determined by the bias w_0

Linear Discriminant Functions and Decisions Surfaces

- The multi-category case
  - We define c linear discriminant functions

$$g_i(x) = w_i^t x + w_{i0} \qquad i = 1, \ldots, c$$

    - assign x to ω_i if $g_i(x) > g_j(x)$ $\forall\, j \neq i$;
    - in case of ties, the classification is undefined
  - In this case, the classifier is a "linear machine"
  - A linear machine divides the feature space into c decision regions, with $g_i(\boldsymbol{x})$ being the largest discriminant if $\boldsymbol{x}$ is in the region R_i

Linear Discriminant Functions and Decisions Surfaces

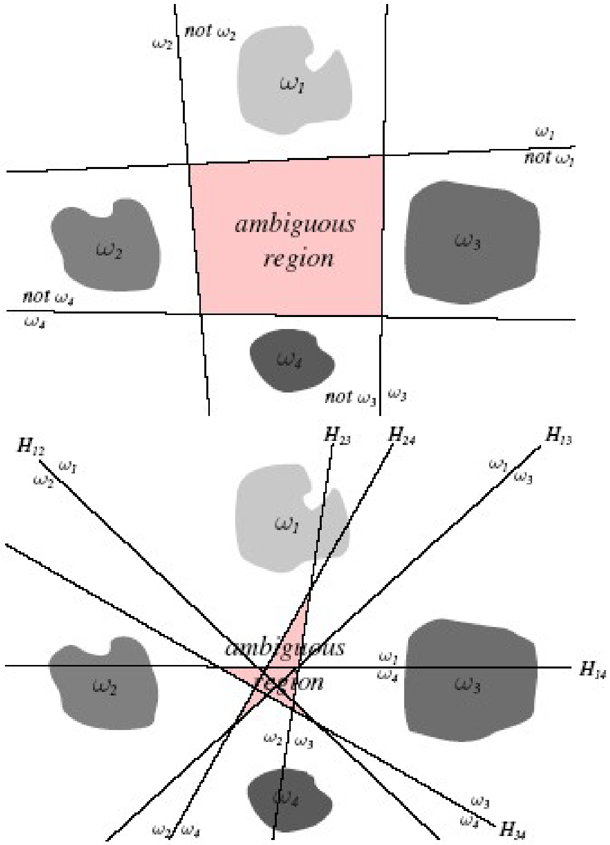

FIGURE 5.3. Linear decision boundaries for a four-class problem. The top figure shows ω_i/not ω_i dichotomies while the bottom figure shows ω_i/ω_j dichotomies and the corresponding decision boundaries H_{ij}. The pink regions have ambiguous category assignments. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Discriminant Functions and Decisions Surfaces

- For a two contiguous regions R_i and R_j, the boundary that separates them is a portion of hyperplane H_{ij} defined by:

$$g_i(x) = g_j(x) \iff (w_i - w_j)^t x + (w_{i0} - w_{j0}) = 0$$

- $(w_i - w_j)$ is normal to H_{ij} and

$$d(x, H_{ij}) = \frac{g_i - g_j}{\|w_i - w_j\|}$$

Linear Discriminant Functions and Decisions Surfaces

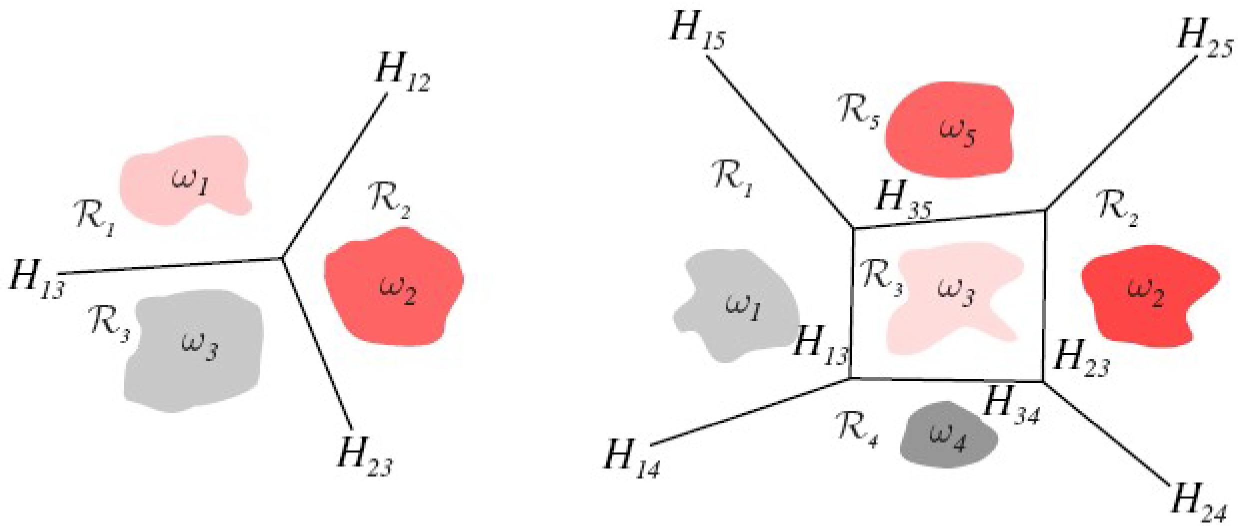

FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Linear Discriminant Functions and Decisions Surfaces

- With the linear machine , it is not the weight vectors, but their differences $||w_i\text{-}w_j||$ that matters.
- It is easy to show that the decision regions for a linear machine are convex, this restriction limits the flexibility and accuracy of the classifier
  - Every decision region is singly connected.
  - This is suitable for conditional densities $p(\boldsymbol{x}|\omega_i)$ are unimodal

Generalized Linear Discriminant Functions

- Decision boundaries which separate between classes may not always be linear
- The complexity of the boundaries may sometimes request the use of highly non-linear surfaces
- A popular approach to generalize the concept of linear decision functions is to consider a generalized decision function as:
  - $g(\boldsymbol{x}) = w_1 f_1(\boldsymbol{x}) + w_2 f_2(\boldsymbol{x}) + \ldots + w_N f_N(\boldsymbol{x}) + w_{N+1}$
    where $f_i(\boldsymbol{x})$, for $1 \leq i \leq N$, are functions of the pattern $\boldsymbol{x}$, where $\boldsymbol{x} \in R^d$ (Euclidean Space)

Generalized Linear Discriminant Functions

- The quadratic discriminant function

$$g(x) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j \qquad (2)$$

  - This function has $(d+1)+d(d+1)/2 = (d+1)(d+2)/2$ terms as well as the weighting coefficients, so it has more flexibility for complicated separating surface.
  - This function produces a hyperquadratic surface
  - The weight matrix is defined as $\boldsymbol{W}=[w_{ij}]$. It is symmetric.

Generalized Linear Discriminant Functions

– Example: a quadratic decision functions for a 2-dimensional feature space

$$g(x) = w_0 + w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 + w_4 x_1 + w_5 x_2$$

where $w = (w_0, w_1, w_2, ..., w_5)^T$

and $\hat{x} = (1, x_1^2, x_1 x_2, x_2^2, x_1, x_2)^T$

Generalized Linear Discriminant Functions

– The commonly used quadratic decision function can be represented as the general d-dimensional quadratic surface:

$$g(x) = x^t\boldsymbol{A}x + x^t b + c$$

where the matrix $\boldsymbol{A} = (a_{ij})$, the vector $b = (b_1, b_2, \ldots, b_n)^t$ and c, depends on the weights w_{ii}, w_{ij}, w_i of equation

Generalized Linear Discriminant Functions

- If $\boldsymbol{A}$ is positive definite then the decision function is a hyperellipsoid with axes in the directions of the eigenvectors of $\boldsymbol{A}$
- If $\boldsymbol{A} = \boldsymbol{I}_d$ (Identity), the decision function is simply the d-dimensional hypersphere
- If $\boldsymbol{A}$ is negative definite, the decision function describes a hyperboloid
- In conclusion: it is only the matrix $\boldsymbol{A}$ which determines the shape and characteristics of the decision function

Generalized Linear Discriminant Functions

– Example: Let R^3 be the original pattern space and let the decision function associated with the pattern classes ω_1 and ω_2 be

$$g(x) = 2x_1^2 + x_3^2 + x_2x_3 + 4x_1 - 2x_2 + 1$$

for which $g(x) > 0$ if $x \in \omega_1$ and $g(x) < 0$ if $x \in \omega_2$

- Rewrite $g(x)$ as $g(x) = x^T A x + x^T b + c$
- Determine the class of each of the following pattern vectors

  [1,1,1], [1,10,0], [0,0.5,0]

Generalized Linear Discriminant Functions

$$g(x) = 2x_1^2 + x_3^2 + x_2 x_3 + 4x_1 - 2x_2 + 1$$

$$g(x) = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 4 \\ -2 \\ 0 \end{bmatrix} + 1$$

$$g(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T) = 7, \quad g(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T) = -13, \quad g(\begin{bmatrix} 0 & 0.5 & 0 \end{bmatrix}^T) = 0$$

Generalized Linear Discriminant Functions

- If more terms such as $w_{ijk}x_ix_jx_k$ are added into the discriminant function, we obtain the class of polynomial discrinimant functions.
- All these can be expressed in terms of the generalized linear discrinimant function

$$g(x) = \sum_{i=1}^{\hat{d}} a_i y_i(x) \quad \text{or} \quad g(x) = a^t y$$

where $\boldsymbol{a}$ is a $\hat{d}$-dimensional weight vector, and $\hat{d}$ functions $y_i(\boldsymbol{x})$, called φ functions, can be arbitrary functions of $\boldsymbol{x}$.

Generalized Linear Discriminant Functions

– The linear discriminant function can also be expressed in the form of generalized linear discriminant function

$$\text{given} \quad g(x) = w_0 + \sum_{i=1}^{d} w_i x_i$$

$$\text{let } y = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \quad \text{and} \quad a = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$$\text{we have} \quad g(x) = a^t y$$

Generalized Linear Discriminant Functions

- By selecting these φ functions carefully, and let d be large enough, we can approximate any desired discriminant functions.
- The resulting discriminant function is not linear to $\boldsymbol{x}$, but is linear to $\boldsymbol{y}$
- The $\hat{d}$ functions $y_i(\boldsymbol{x})$ map d-dimensional x-space to $\hat{d}$ -dimensional y-space. Therefore this mapping reduces the problem to one of finding a homogeneous linear discriminant function.

Generalized Linear Discriminant Functions

– Example: let the quadratic discriminant function be

$$g(x)=a_1+a_2x+a_3x^2$$

the original x-pace is 1-dimensional, and the 3-dimensional vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y}=[1, x, x^2]^t$$

- Varying x in 1-D will cause $\boldsymbol{y}$ to trace out a curve in 3-D

Generalized Linear Discriminant Functions

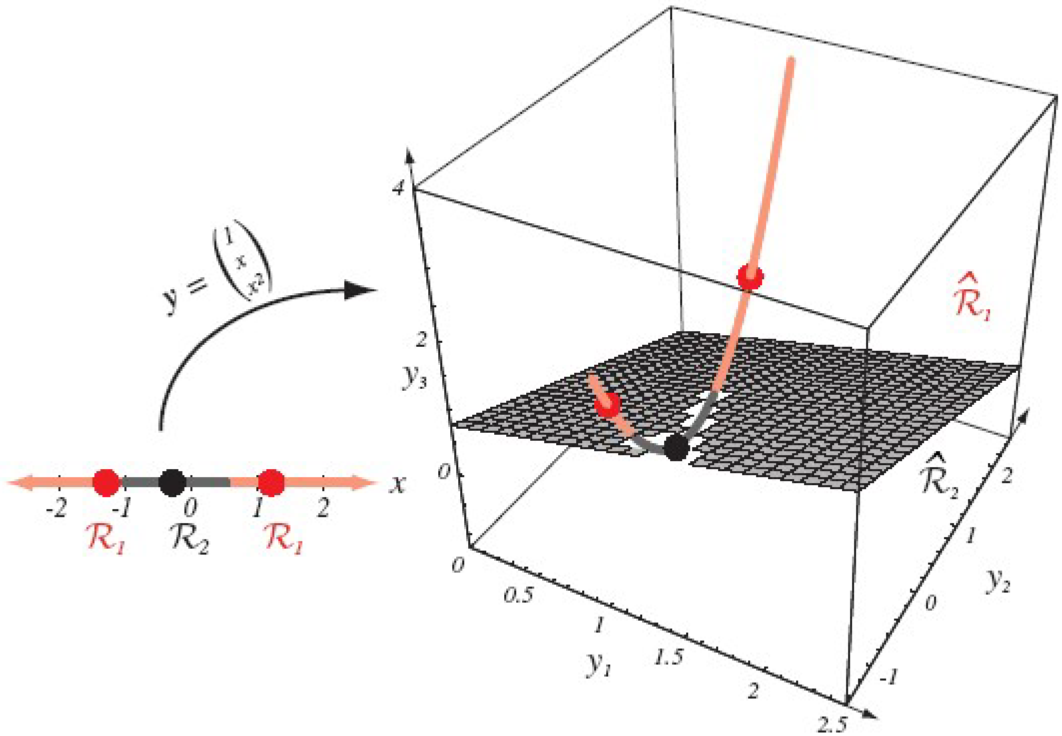

FIGURE 5.5. The mapping $\mathbf{y} = (1, x, x^2)^t$ takes a line and transforms it to a parabola in three dimensions. A plane splits the resulting $\mathbf{y}$-space into regions corresponding to two categories, and this in turn gives a nonsimply connected decision region in the one-dimensional x-space. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Generalized Linear Discriminant Functions

- Disconnected regions may become connected in high dimension
- If x is from $p(x)$, the density in y-space, $\hat{p}(y)$, will be degenerate, being zero everywhere except on the curve, where it is infinite. This is common problem when $\hat{d} > d$, or mapping from a lower dimensional space to a higher dimensional space.

Generalized Linear Discriminant Functions

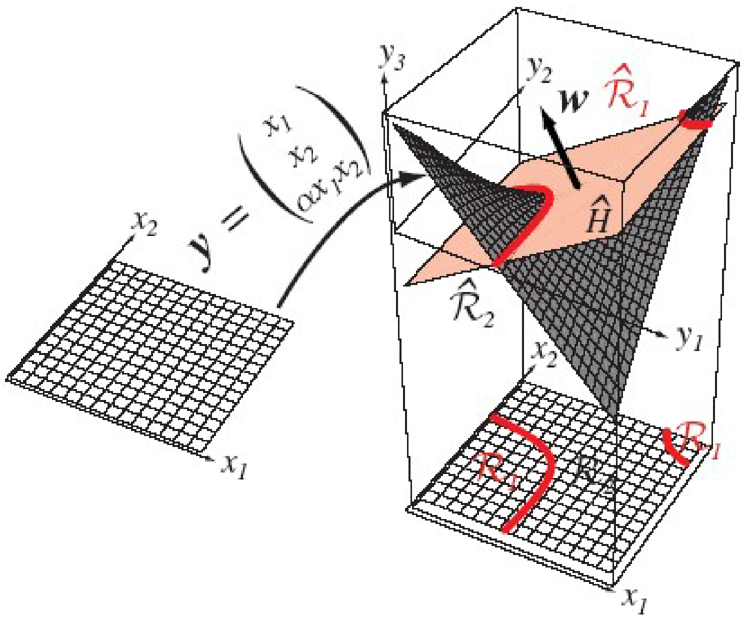

FIGURE 5.6. The two-dimensional input space $\mathbf{x}$ is mapped through a polynomial function f to $\mathbf{y}$. Here the mapping is $y_1 = x_1$, $y_2 = x_2$ and $y_3 \propto x_1x_2$. A linear discriminant in this transformed space is a hyperplane, which cuts the surface. Points to the positive side of the hyperplane $\hat{H}$ correspond to category ω_1, and those beneath it correspond to category ω_2. Here, in terms of the $\mathbf{x}$ space, $\mathcal{R}_1$ is a not simply connected. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Generalized Linear Discriminant Functions

– The major disadvantage of the generalized linear discriminant functions is the curse of dimensionality
  - For d=50, $\hat{d} = \dfrac{(d+1)(d+2)}{2} = 1326$

The Two-Category Linearly Separable Case

- In a 2-category case, given a linear discriminant function $g(\boldsymbol{x})=\boldsymbol{a}^t\boldsymbol{y}$, if there is a weight vector that can correctly classify all the samples, these samples are called linearly separable.
- In this case, if $\boldsymbol{a}^t\boldsymbol{y}_i>0$, $\boldsymbol{y}_i$ is labeled ω_1, and if $\boldsymbol{a}^t\boldsymbol{y}_i<0$, $\boldsymbol{y}_i$ is labeled ω_2.
- Or we can replace all samples labeled ω_2 by their negatives, and then we are looking for a weight vector $\boldsymbol{a}$ such that $\boldsymbol{a}^t\boldsymbol{y}_i>0$ for all the samples. This weight vector is called a separating vector.

The Two-Category Linearly Separable Case

- To find this weight vector $\boldsymbol{a}$, we consider it as a point in a weight space (with all possible weight vectors).
- Each sample $\boldsymbol{y}_i$ will impose a constraint on the possible location of this $\boldsymbol{a}$. The equation $\boldsymbol{a}^t\boldsymbol{y}_i=0$ defines a hyperplane
  - In data space $\boldsymbol{a}$ is vector perpendicular to the separating hyperplane $\boldsymbol{a}^t\boldsymbol{y}_i=0$.
  - The region satisfies $\boldsymbol{a}^t\boldsymbol{y}_i>0$ is a half-space on one side of the hyperplane $\boldsymbol{a}^t\boldsymbol{y}_i=0$.
- If a solution vector exists, it will come from a solution region, which is an intersection of all half-spaces that each satisfies $\boldsymbol{a}^t\boldsymbol{y}_i>0$ for a particular $\boldsymbol{y}_i$.

The Two-Category Linearly Separable Case

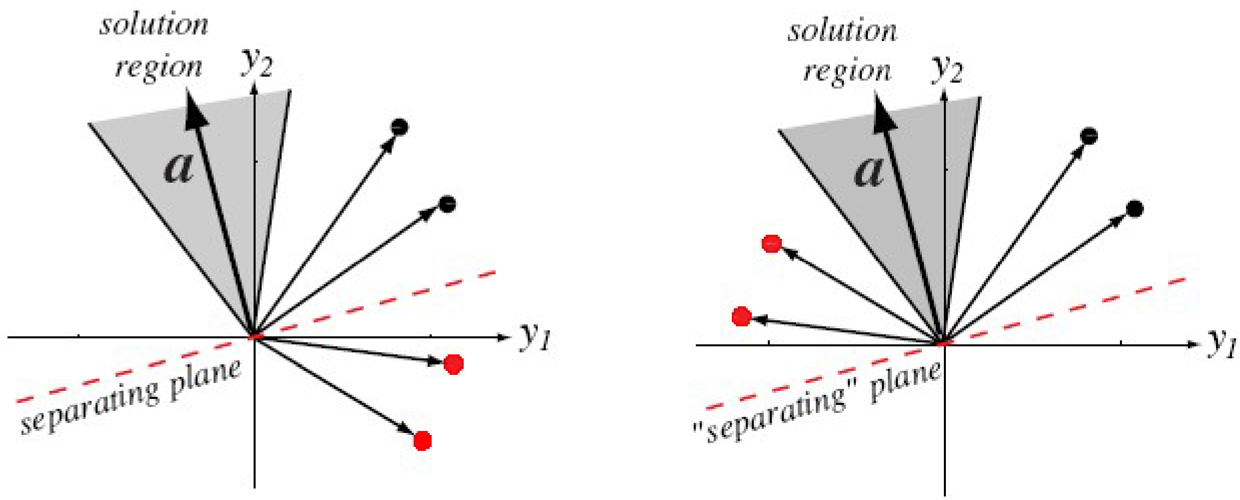

FIGURE 5.8. Four training samples (black for ω_1, red for ω_2) and the solution region in feature space. The figure on the left shows the raw data; the solution vectors leads to a plane that separates the patterns from the two categories. In the figure on the right, the red points have been "normalized"—that is, changed in sign. Now the solution vector leads to a plane that places all "normalized" points on the same side. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Two-Category Linearly Separable Case

- So solution vector is usually not unique to a finite set of samples.
- Additional constraints can be imposed to fine a unique solution
  - One possibility is to seek a unit-length weight vector that maximize the minimum distance from the sample to the separating plane.
  - Another possibility is to seek a minimum-length weight vector that satisfies $\boldsymbol{a}^t\boldsymbol{y}_i \geq b$ for all i and some positive constant b, called margin. This will move the new boundaries from the old boundaries by $b/\|\boldsymbol{y}_i\|$ for all i

The Two-Category Linearly Separable Case

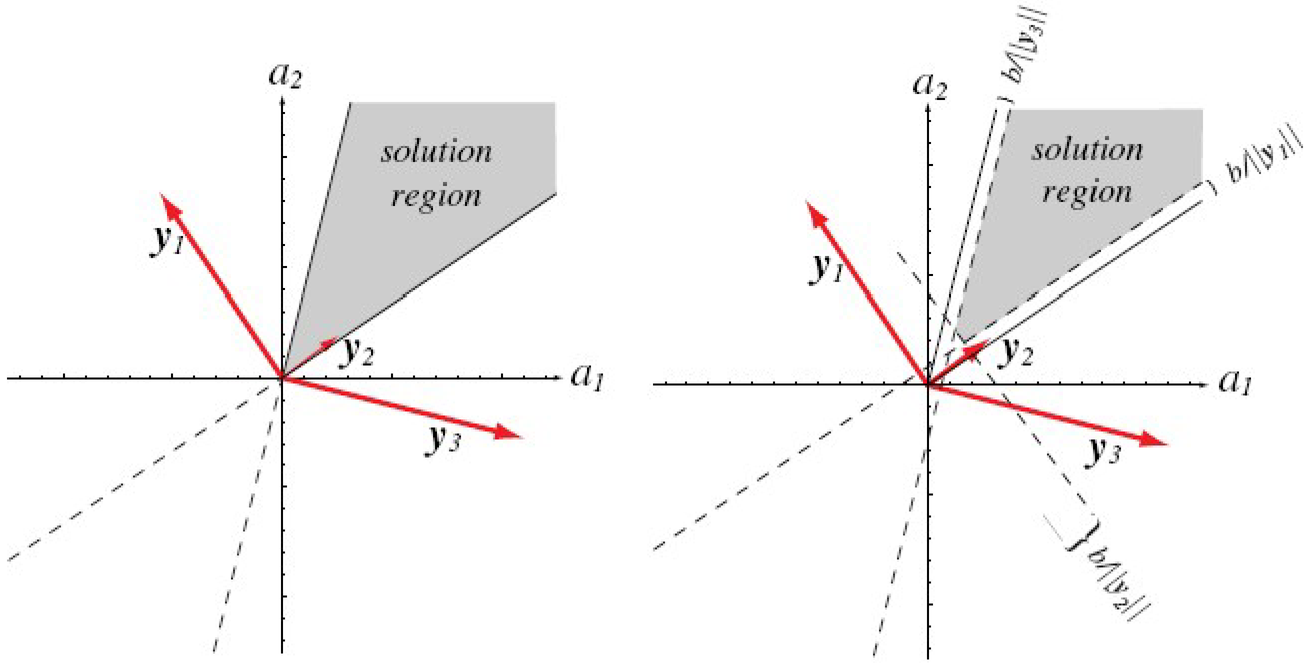

FIGURE 5.9. The effect of the margin on the solution region. At the left is the case of no margin ($b = 0$) equivalent to a case such as shown at the left in Fig. 5.8. At the right is the case $b > 0$, shrinking the solution region by margins $b/\|\mathbf{y}_i\|$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Two-Category Linearly Separable Case

- Gradient Descent Procedure
  - The goal is to find a solution to the set of linear inequalities $\boldsymbol{a}^t\boldsymbol{y}_i>0$
  - Define a criterion function $J(\boldsymbol{a})$ such that when $\boldsymbol{a}$ is a solution vector, this function is minimized.
  - Minimizing this scale function can be done through a gradient descent procedure

The Two-Category Linearly Separable Case

– Basic gradient descent
  - At iteration i=0, $\boldsymbol{a}(0)$ is a arbitrary weight vector
  - At iteration i=k, calculate

$$\boldsymbol{a}(k+1)=\boldsymbol{a}(k)-\eta(k)\nabla J(a(k))$$

    i.e. moving $\boldsymbol{a}(k+1)$ from $\boldsymbol{a}(k)$ along the negative of the gradient, $\eta(k)$ is step size or learning rate
  - Continue until $|\eta(k)\nabla J(\boldsymbol{a}(k))| < \theta$, where θ is a threshold.

The Two-Category Linearly Separable Case

– Newton descent

- The iteration is based on

$$\boldsymbol{a}(k+1)=\boldsymbol{a}(k)-\boldsymbol{H}^{-1}\nabla J$$

  where $\boldsymbol{H}$ is the Hessian matrix of second partial derivatives $\partial^2 J/\partial a_i \partial a_j$ evaluated at $\boldsymbol{a}(k)$
- Usually Newton descent will give a greater improvement per step than the simple gradient descent algorithm.
- It is not applicable if the Hessian matrix is singular

The Two-Category Linearly Separable Case

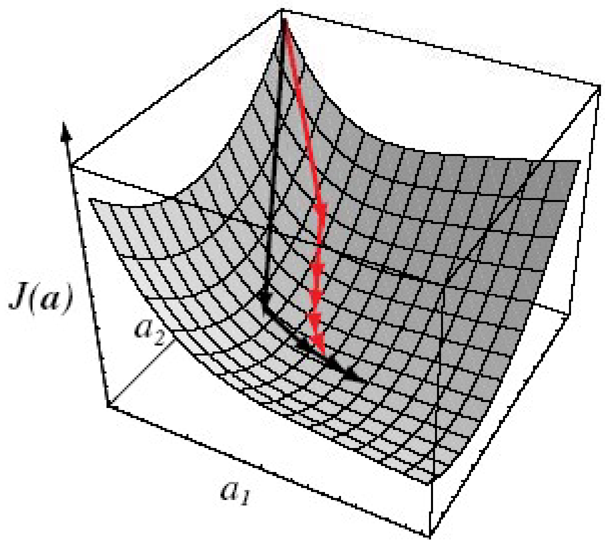

FIGURE 5.10. The sequence of weight vectors given by a simple gradient descent method (red) and by Newton's (second order) algorithm (black). Newton's method typically leads to greater improvement per step, even when using optimal learning rates for both methods. However the added computational burden of inverting the Hessian matrix used in Newton's method is not always justified, and simple gradient descent may suffice. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Two-Category Linearly Separable Case

- Criterion Functions
  - The Perceptron criterion function

$$J_p(a) = \sum_{y \in Y} (-a^t y) \text{ and then } \nabla J_p = \sum_{y \in Y} (-y)$$

where $Y(\boldsymbol{a})$ is the set of samples misclassified by ***a.***

- The update rule becomes

$$a(k+1) = a(k) + \eta(k) \sum_{y \in Y} y$$

- This is called Batch Perceptron

The Two-Category Linearly Separable Case

- Some other criterion functions

$$J_q(a) = \sum_{y \in Y} (-a^t y)^2$$

$$J_r(a) = \frac{1}{2} \sum_{y \in Y'} \frac{(a^t y - b)^2}{\| y \|^2}$$

where Y' is the set of samples for which $a^t y \le b$

$$J_s(a) = \| Ya - b \|^2 = \sum_{i=1}^{n} (a^t y_i - b_i)^2$$

Support Vector Machine

- Support vector machines (SVMs) are linear machines with margins
- SVMs rely on nonlinear function (kernel) $\varphi(\cdot)$ to map the data into a sufficiently high dimension, in which two categories can always be separated by a hyperplane.
- Given n samples, k=1, 2, …, n. Let z_k=+1 or -1 for sample x_k in ω_1 or ω_2 respectively. We have

$$\boldsymbol{y_k} = \varphi(\boldsymbol{x_k})$$

$$g(\boldsymbol{y}) = \boldsymbol{a}^t\boldsymbol{y}$$

the separating hyperplane ensures

$$z_k g(\boldsymbol{y_k}) \geq 1, \quad k=1, \ldots, n.$$

Support Vector Machine

- The goal in training a SVM is to find the separating hyperplane with the largest margin.
- The distance from the hyperplane to a pattern $\boldsymbol{y}$ is $|g(\boldsymbol{y})|/\|\boldsymbol{a}\|$, assume a positive margin exists, then

$$\frac{z_k g(y_k)}{\| a \|} \geq b, \qquad k = 1, \ldots, n.$$

- We will seek $\boldsymbol{a}$ that maximizes b. To ensure uniqueness of the solution, we impose a constraint $b\|\boldsymbol{a}\|=1$. This is equivalent to minimizing $\|\boldsymbol{a}\|^2$

Support Vector Machine

- The support vectors are the transformed training samples for which $z_k g(\boldsymbol{y_k})=1$, they are equally close to the hyperplane, and they are the most difficult patterns to classify

Support Vector Machine

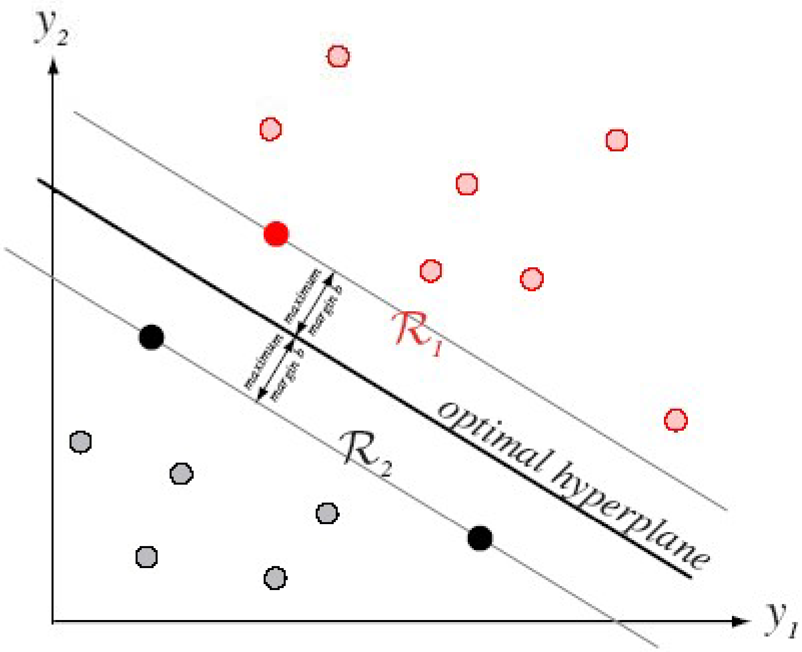

FIGURE 5.19. Training a support vector machine consists of finding the optimal hyperplane, that is, the one with the maximum distance from the nearest training patterns. The support vectors are those (nearest) patterns, a distance *b* from the hyperplane. The three support vectors are shown as solid dots. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Support Vector Machine

- SVM training
  - Choice of φ(·) requires domain knowledge. Other common choices include polynomials, Gaussians, radial basis functions (RBF)
  - Define $$L(a,\alpha) = \frac{1}{2}\| a \|^2 - \sum_{k=1}^{n} \alpha_k [z_k a^t y_k - 1]$$

    where α_k are Lagrange multipliers.
  - We will minimize $L(\cdot)$ w.r.t. ***a***, and maximize it w.r.t. $\alpha_k \geq 0$.