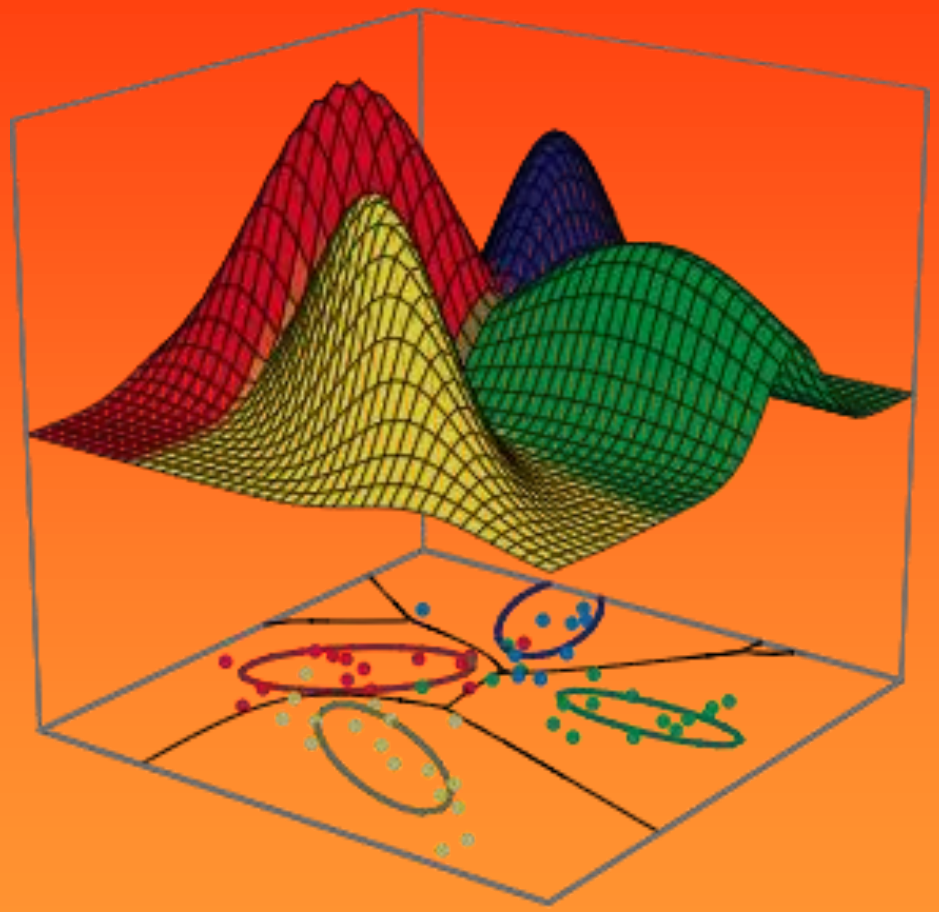

Pattern Classification

All materials in these slides were taken from *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher

Chapter 4, Section 4.5

K-Nearest Neighbor Classification

The nearest –neighbor classification rule

- Let $D_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n labeled prototypes
- Let $x' \in D_n$ be the closest prototype to a test point x; then, the nearest-neighbor rule for classifying x is to assign it the label associated with x'
- The nearest-neighbor rule leads to an error rate greater than the minimum possible: the Bayes rate
- If the number of prototypes is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Bayes rate (it can be demonstrated!)
- If $n \rightarrow \infty$, it is always possible to find a labeled x' sufficiently close so that:

$$P(\omega_i \mid x') \cong P(\omega_i \mid x)$$

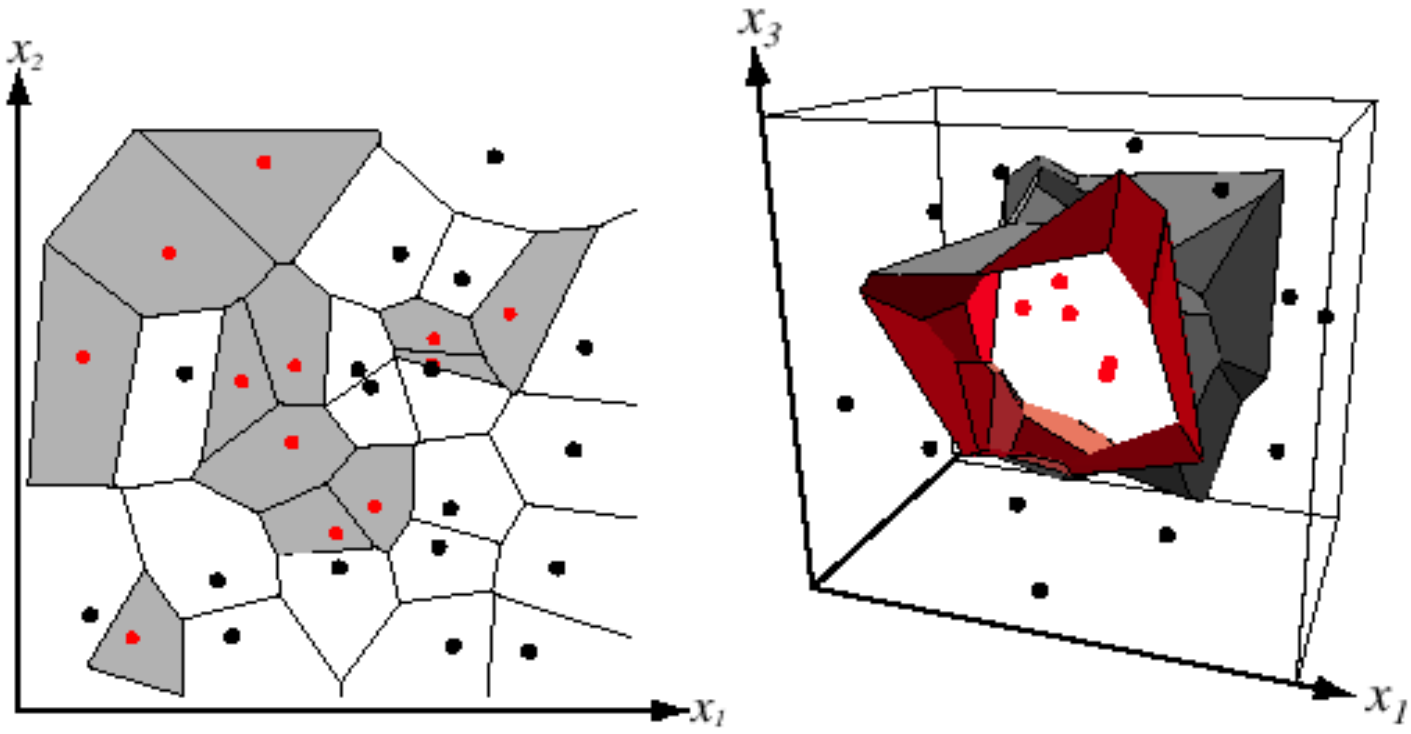

FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Each cell defines the region “closest” to the training point in that cell.

Example: more than two classes

$x = (0.68, 0.60)^t$

Prototypes	Labels	A-posteriori probabilities estimated
(0.50, 0.30)	ω_2 ω_3	0.25 $0.75 = P(\omega_m \mid x)$
(0.70, 0.65)	ω_5 ω_6	0.70 0.30

Decision: ω_5 is the label assigned to x

If $P(\omega_m \mid x) \cong 1$, then the nearest neighbor selection is almost always the same as the Bayes selection

The k – nearest-neighbor rule
(k >= 1, k usually an odd number)

- Goal: Classify x by assigning it the label most frequently represented among the k nearest samples and use a voting scheme

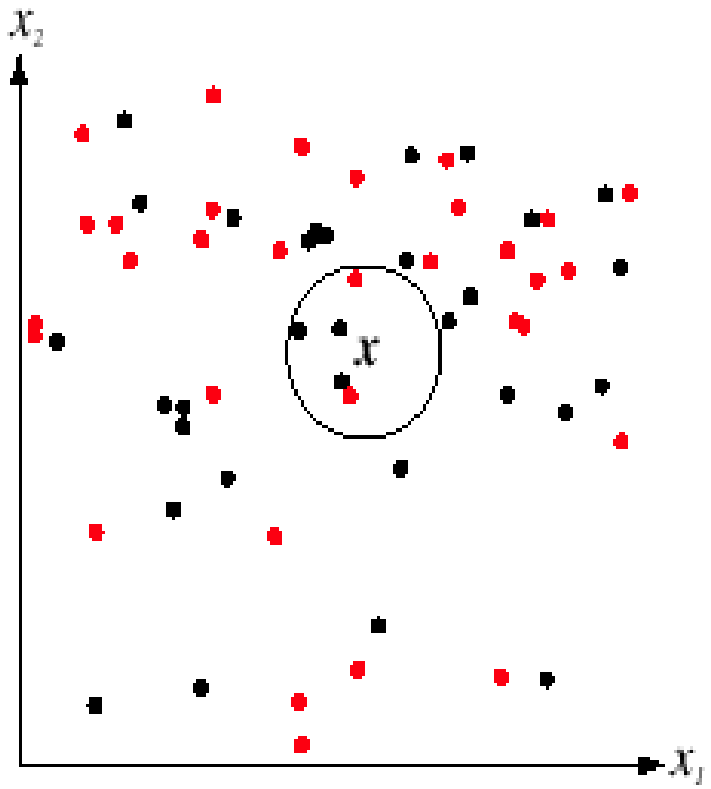

FIGURE 4.15. The k-nearest-neighbor query starts at the test point $\mathbf{x}$ and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point $\mathbf{x}$ would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Example:

k = 3 (odd value) and $x = (0.10, 0.25)^t$

Prototypes	Labels	Euclidean Distance
(0.15, 0.35)	ω_1	*0.112*
(0.10, 0.28)	ω_2	*0.030*
(0.09, 0.30)	ω_5	*0.051*
(0.12, 0.20)	ω_2	*0.054*

One voting scheme assigns the label ω_2 to x since ω_2 is the most frequently represented in the k=3 closest points

Chapter 5:
Linear Discriminant Functions
(Sections 5.1-5.3)

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 (assume we know) the proper forms for the discriminant functions: similar to non-parametric techniques
- They may not be optimal, but they are very simple to use
- They provide us with linear classifiers

Linear discriminant functions and decision surfaces

Definition

A function that is a linear combination of the components of x

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

where w is the weight vector and w_0 the bias

A two-category classifier with a discriminant function of the form (1) uses the following rule:

Decide ω_1 *if* $g(x) > 0$ and ω_2 if $g(x) < 0$

$\Leftrightarrow$ Decide ω_1 *if* $w^t x > -w_0$ and ω_2 otherwise

If $g(x) = 0 \Rightarrow x$ is assigned to either class (randomly)

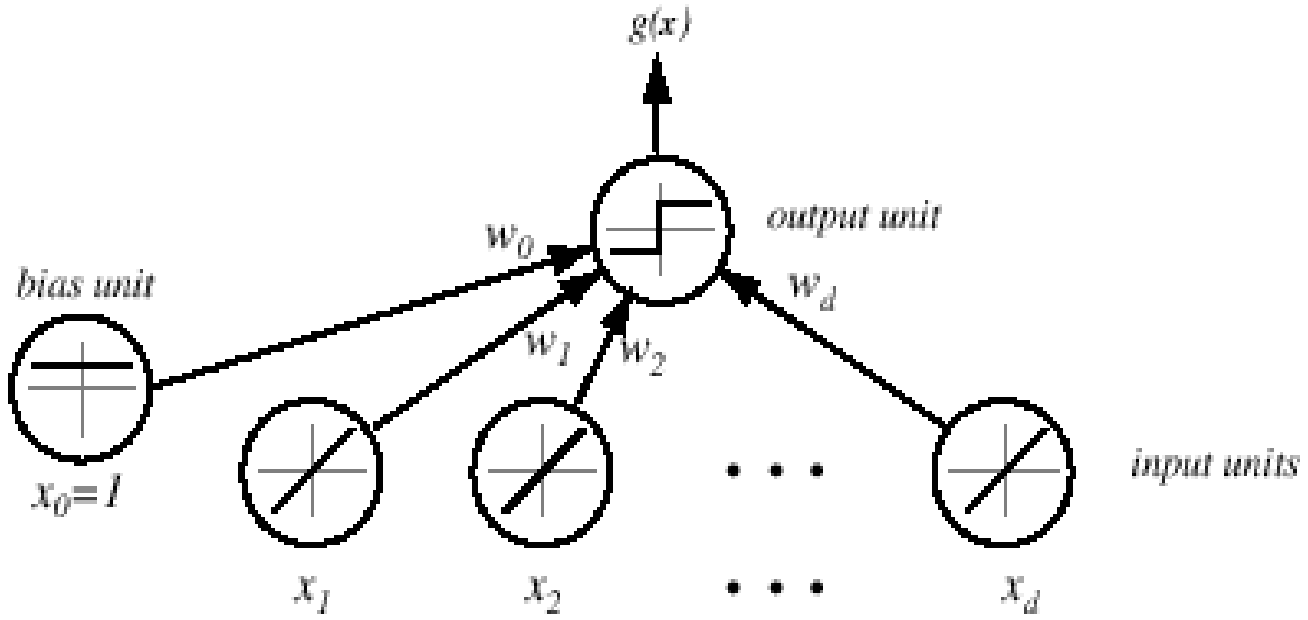

FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show 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.

- The equation $g(x) = 0$ defines the decision surface that separates points assigned to the category ω_1 from points assigned to the category ω_2
- When $g(x)$ is linear, the decision surface is a hyperplane
- Algebraic measure of the distance from x to the hyperplane (interesting result!)

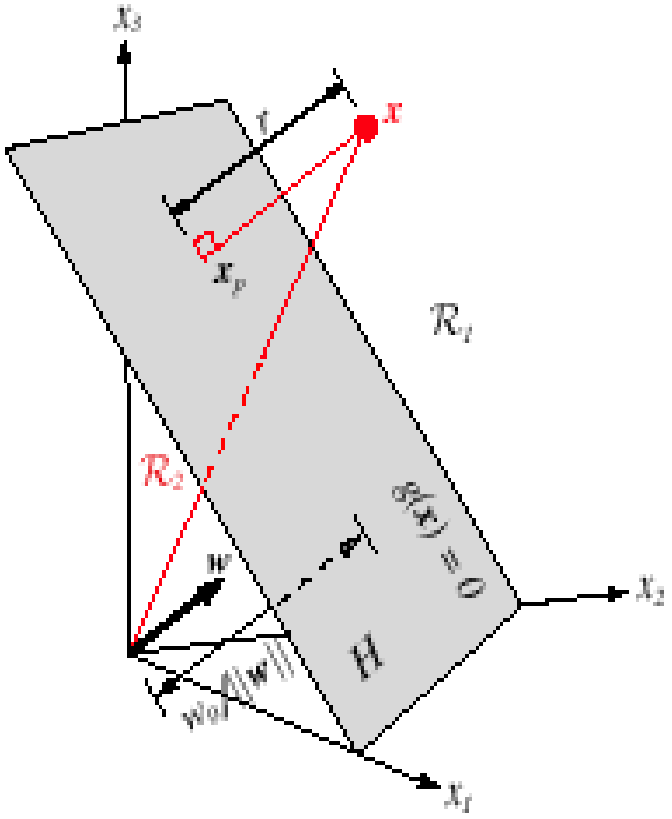

FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork,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

The multi-category case

- We define c linear discriminant functions

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

  and 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 $g_i(x)$ being the largest discriminant if x is in the region R_i
- For two contiguous regions R_i and R_j, the boundary that separates them is a portion of hyperplane H_{ij} defined by:

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

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

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

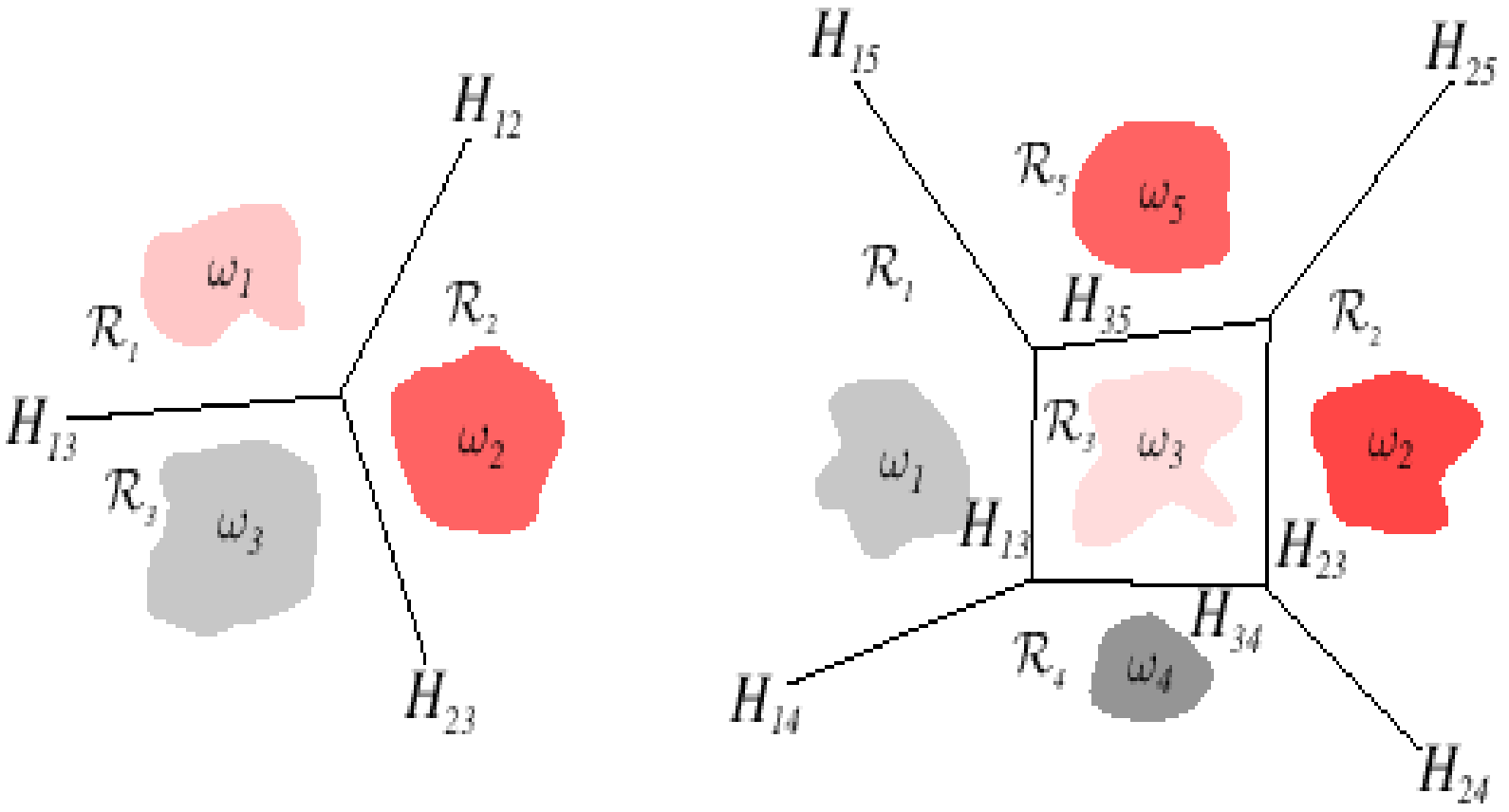

FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

One can show that the decision regions for a linear machine are convex. This restriction limits the flexibility and accuracy of the classifier