IDEA GROUP PUBLISHING
ITB12298
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

This paper appears in the publication, *Biometric Image Discrimination Technologies*
**by** David Zhang, Xiaoyuan Jing, and Jian Yang © 2006, Idea Group Inc.

## Chapter III

# Linear Discriminant Analysis

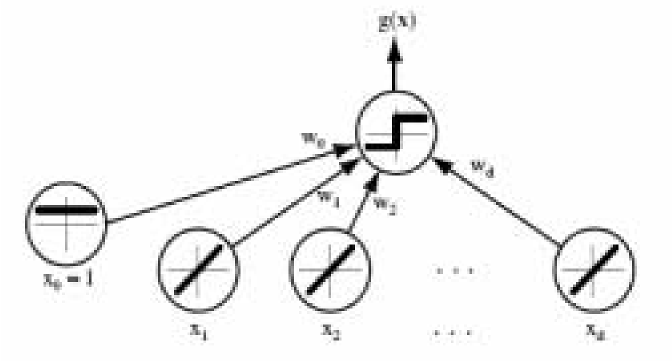
## ABSTRACT

*This chapter deals with issues related to linear discriminant analysis (LDA). In the introduction, we indicate some basic conceptions of LDA. Then, the definitions and notations related to LDA are discussed. Finally, the introduction to non-linear LDA and the chapter summary are given.*

## INTRODUCTION

Although PCA finds components useful for representing data, there is no reason to assume these components must be useful for discriminating between data in different classes. As was said in Duda, Hart and Stork (2000), if we pool all of the samples, the directions that are discarded by PCA might be exactly the directions needed for distinguishing between classes. For example, if we had data for the printed uppercase letters O and Q, PCA might discover the gross features that characterize Os and Qs, but might ignore the tail that distinguishes an O from a Q. Whereas PCA seeks directions that are efficient for representation, *discriminant analysis* seeks directions that are efficient for discrimination (McLachlan, 1992; Chen, Liao, Ko, Lin, & Yu, 2000; Hastie, Buja, & Tibshirani, 1995). In the previous chapter we introduced algebraic considerations for dimensionality reduction which preserve variance. We can see that variance preserving dimensionality reduction is equivalent to (1) de-correlating the training sample data, and (2) seeking the d-dimensional subspace of $\mathbb{R}^n$ that is the closest (in the least-squares

*Figure 3.1. 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 output unit sums all these products and emits a +1 if $\mathbf{w}^T\mathbf{x} + w_0 > 0$ or a 1 otherwise (Duda, Hart, & Stork, 2000)*

sense) possible to the original training sample. In this chapter, we extend the variance preserving approach for data representation for labeled data sets. In this section, we will focus on two-class sets and look for a separating hyperplane (Yang & Yang, 2001; Xu, Yang, & Jin, 2004; Etemad & Chellappa, 1997):

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{3.1}$$

such that $\mathbf{x}$ belongs to the first class if $g(\mathbf{x}) > 0$ and $\mathbf{x}$ belongs to the second class if $g(\mathbf{x}) < 0$. In the statistical literature, this type of function is called a linear discriminant function. The decision boundary is given by the set of points satisfying $g(\mathbf{x}) = 0$ which is a hyperplane. Fisher's LDA is a variance preserving approach for finding a linear discriminant function (Duda, Hart, & Stork, 2000; McLachlan, 1992; Chen, Liao, Ko, Lin, & Yu, 2000).

## The Two-Category Case

A discriminant function that is a linear combination of the components of $\mathbf{x}$ can be written as Equation 3.1, where $\mathbf{w}$ is the *weight vector* and $w_0$ the *bias* or *threshold weight*. A two-category threshold weight linear classifier implements the following decision rule: Decide $\omega_1$ if $g(\mathbf{x}) > 0$ and $\omega_2$ if $g(\mathbf{x}) < 0$. Thus, $\mathbf{x}$ is assigned to $\omega_1$ if the inner product $\mathbf{w}^T\mathbf{x}$ exceeds the threshold $-w_0$ and $\omega_2$ otherwise. If $g(\mathbf{x}) = 0$, $\mathbf{x}$ can ordinarily be assigned to either class, but in this chapter we shall leave the assignment undefined. Figure 3.1 shows a typical implementation, a clear example of the general structure of a pattern recognition system we saw (Duda, Hart, & Stork, 2000).

The equation $g(\mathbf{x}) = 0$ defines the decision surface that separates points assigned to $\omega_1$ from points assigned to $\omega_2$. When $g(\mathbf{x})$ is linear, this decision surface is a *hyperplane* (Burges, 1996; Evgeniou, Pontil, & Poggio, 1999; Ripley, 1994; Suykens & Vandewalle,

## Related Content