Geometric Discriminant Analysis (part II)

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

In these slides we discuss the approach originally proposed by Fisher. He formulated the classfication problem in a geometric way. He sought to find the linear combination of the predictors such that the between-group variance was maximized relative to the within-group variance.

In other words, he wanted to find the combination of the predictors that gave maximum separation between the centroids of the data while at the same time minimizing the variation within each group of data.

Main Problem

How to find a representation of the objects which provides the best separation between groups (description emphasis)?

How to find the rules for assigning the objects to their groups (prediction emphasis)?

Geometric Predictive Discriminant Analysis

Predictive Idea

Each observation x_i is classed in the group G_k for which the distance to the center $\mathbf{g_k}$ is minimal, the distance being calculated using the Mahalanobis metric $\mathbf{W^{-1}}$

$$\begin{aligned} d^2(\mathbf{x_i}, \mathbf{g_k}) &= (\mathbf{x_i} - \mathbf{g_k})^\top \mathbf{W^{-1}} (\mathbf{x_i} - \mathbf{g_k}) \\ &= \mathbf{x_i}^\top \mathbf{W^{-1}} \mathbf{x_i} - 2\mathbf{g_k}^\top \mathbf{W^{-1}} \mathbf{x_i} + \mathbf{g_k}^\top \mathbf{W^{-1}} \mathbf{g_k} \end{aligned}$$

Discriminant Functions

Note that minimizing $d^2(\mathbf{x_i}, \mathbf{g_k})$ is equivalent to maximizing $2\mathbf{g_k^T W^{-1} x_i} + \mathbf{g_k^T W^{-1} g_k}$.

Let $\alpha_k = \mathbf{g_k^T W^{-1} g_k}$. Note that this a constant that does not depend on $\mathbf{x_i}$.

Thus, for each of the k groups we have a **discriminant linear function** that is found after inversion of the matrix $\mathbf{W}$

Mahalanobis D^2

Classification is based on the concept of distance from a point to a centroid.

An important quantity is the so-called Mahalanobis D^2, which is the square of the distance between the two centroids:

$$D^2 = d^2(\mathbf{g_1}, \mathbf{g_2}) = (\mathbf{g_1} - \mathbf{g_2})^\top \mathbf{W}^{-1} (\mathbf{g_1} - \mathbf{g_2})$$

Mahalanobis D^2

The Mahalanobis D^2 measures the (square) distance between the two groups to be discriminated, and thus it also measures the quality of the discrimination: a higher value means better discrimination.

Geometric Classification Rule

The rule of classification of geometric discriminant analysis is a linear rule (which is why we speak of *linear discriminant analysis*).

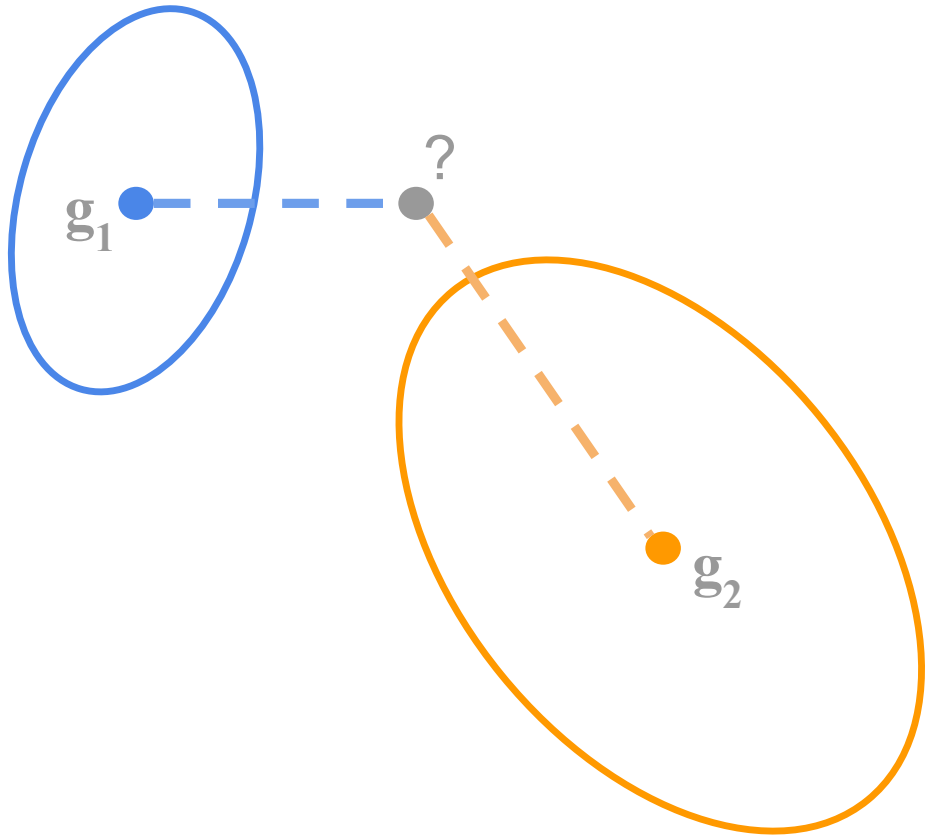
We assign each object to the group it is nearest to, using the $\mathbf{W}^{-1}$ metric to calculate the distance of the object from the centroid of the group, in other words by carrying out an oblique projection of $\mathbf{x_i}$ on the discriminant axis.

Geometric Classification Rule

Warning

The geometric rule of classification should not be used if the two groups have different *a priori* probabilities or variances.

Limitations of Geometric Classification



To which group should we assign the new object?

References

- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Oxford University Press.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*. Wiley.
- **Multivariate Analysis** by Maurice Tatsuoka (1988). *Chapter 7: Discriminant Analysis and Canonical Correlation*.
- **Practical Biostatistical Methods** by Steve Selvin (1995) *Chapter 6: Linear Discriminant Analysis*. Duxbury Press.

References

- **The use of multiple measurements in taxonomic problems** by R.A. Fisher (1936). *Annals of Eugenics, 7, 179-188*.
- **On the generalized distance in statistics** by P.C. Mahalanobis (1936). *Proceedings of the National Institute of Science, India, 12, 49-55*.
- **Discriminant Analysis** by Tatsuoka and Tiedeman (1954). *Review of Educational Research, 25, 402-420*.

References (French Literature)

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique explicative appliquee** by Nakache and Confais (2003). *Chapter 1: Analyse discriminante sur variables quantitatives*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.