DISCRIMINANT ANALYSIS BETWEEN TWO POPULATIONS USING THE MAXIMUM FUNCTION

T. Pham-Gia
Universite De Moncton, Canada
phamgit@umoncton.ca

The maximum function can be used very efficiently in discriminant analysis and hypothesis testing, and can significantly improve the comprehension of the approach presented.

A PROBLEM ENCOUNTERED IN TEACHING DISCRIMINANT ANALYSIS

Discriminant analysis is a statistical technique designed to find a relation between variables, to be used to distinguish between distinct groups (or discriminate one group from another). This relation is to be established from a set of training data. Usually, the same relation also serves to classify a new data set into one of the groups.

We consider here the Bayesian approach, based on two, or several, distinct statistical distributions which represent the groups. In classical discriminant analysis with the normal model, as taught at universities and colleges, we consider the linear or quadratic discriminant function, and look for the values of its parameters that would fit the data best. However, the presentation of this approach, limited to data R^p space, does not easily allow a clear comprehension of the arguments by students, and emphasis is put on the linear case, which is simpler. In this article, we will use the function

$$\mathbf{g}_{\max}(\mathbf{x}) = \max\{qf_1(\mathbf{x}), (1-q) f_2(\mathbf{x})\}$$

and its graphs to improve this presentation. Students will be able to see where the linear or quadratic boundaries come from, and their very logical uses.

DISCRIMINATION BETWEEN TWO NORMAL POPULATIONS: CLASSICAL APPROACH

Let $\boldsymbol{X}_\mathbf{i} \sim \mathbf{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i =1,2, be two independent normal (column) vectors in R^p, representing populations π_1 and π_2. In classification, we consider the two normal multivariate densities, together with their prior probabilities q and 1 - q.

In practice, estimates of the population mean vectors, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and the variance-covariance matrices, Σ_1 and Σ_2, are obtained from independent samples of size n_1 and n_2, from the two populations, using their sample equivalents, $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, and $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.

Equal Covariance Matrices: $\Sigma_1 = \Sigma_2 = \Sigma$

The decision rule is as follows: For a new observation $\boldsymbol{x}_0$, allocate it to π_1 if $ld(\mathbf{x}_0)>0$, where the linear discriminant function is

$$ld(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\boldsymbol{S}^*)^{-1}\boldsymbol{x} - A \quad\text{.......(1)}$$

where

$$A = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'(\boldsymbol{S}^*)^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + \ln((1-q)/q)$$

and $\boldsymbol{S}^*$ is the estimate of Σ obtained by pooling $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. The function $ld(\boldsymbol{x})$ is represented by the straight line D_1.

Different Covariance Matrices: $\Sigma_1 \neq \Sigma_2$

We allocate $\boldsymbol{x}_0$ to π_1 if the value of the quadratic discriminant function at $\boldsymbol{x}_0$,

$$qd(\boldsymbol{x}_0) = \left[\bar{\boldsymbol{x}}_1'(\boldsymbol{S}_1)^{-1} - \bar{\boldsymbol{x}}_2'(\boldsymbol{S}_2)^{-1}\right]x_0 - \frac{1}{2}\boldsymbol{x}_o'\left[(\boldsymbol{S}_1)^{-1} - (\boldsymbol{S}_2)^{-1}\right]\boldsymbol{x}_0 - k > \ln((1-q)/q) \quad\text{.....(2)}$$

and to π_2, otherwise, where

$$k = \frac{1}{2}\left[\ln(|\boldsymbol{S}_1|/|\boldsymbol{S}_2|) + (\bar{\boldsymbol{x}}_1'(\boldsymbol{S}_1)^{-1}\bar{x}_1 - \bar{\boldsymbol{x}}_2'(\boldsymbol{S}_2)^{-1}\bar{\boldsymbol{x}}_2)\right]$$

The function $qd(x)$ is represented by a curve of 2^{nd} degree (parabola, hyperbola or ellipse).

PROPOSED METHOD

The above approach is solely based on mathematical equations. We now use graphs too.

We define $g_{max}(\mathbf{x})$ as above, and in R^{p+1}, draw the graph $\{\mathbf{x}, g_{max}(\mathbf{x})\}$. We adopt the decision rule: For $\mathbf{x}_0$, allocate it to π_1 if $g_{max}(\mathbf{x}) = qf_1(\mathbf{x})$, and to π_2 otherwise. We can easily prove the equivalence with the classical method. Furthermore, this function enjoys several properties associated with the Bayes error, P_e, which is the minimum error in classification.

EXAMPLE: DISCRIMINATION BETWEEN CANADIAN AND ALASKAN SALMON

A good example to illustrate this process is the following (Johnson and Wichern (1998, p.659)). We wish to classify a salmon to either the Alaskan or the Canadian species, based on observations X and Y, where X is the diameter of its scale rings for the first year of freshwater growth, and Y is similarly defined for the first year of marine growth. We have 2 samples of 50 for each species. The two populations are supposed normal and $q = 0.50$.

Linear Discriminant Function

Supposing a common covariance matrix (estimated by the mean of $\mathbf{S}_1$ and $\mathbf{S}_2$ computed in the next section), we have:

$$S^* = \begin{bmatrix} 17.26^2 & -28.96 \\ -28.96 & 28.67^2 \end{bmatrix}$$

while $\mathbf{x}_1 = (98.38, 429.66)$ and $\mathbf{x}_2 = (137.46, 366.62)$.

By (1), the linear boundary D_1, given by $ld(\mathbf{x})$, is $y = 2.4336x + 111.1705$.

In R^3 the two populations are represented by two normal surfaces with similar shape. They intersect each other along a space curve G_1, whose projection on the (x,y)-plane is the above straight line D_1 (Figure 1). The graph of g_{max} is easily obtained from these surfaces.

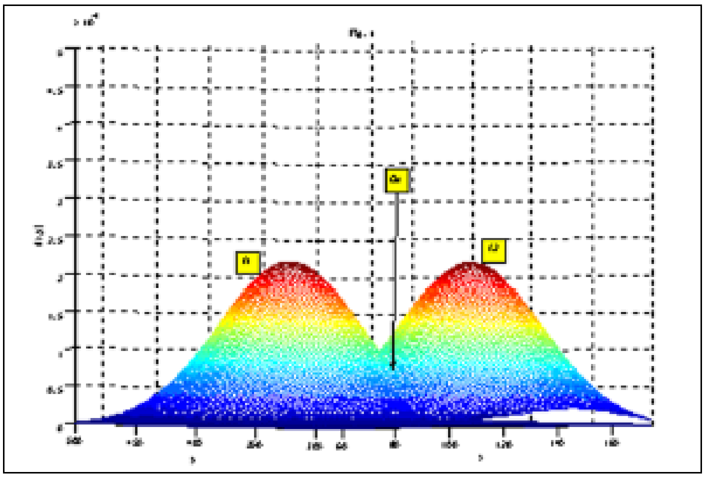

Figure 1

Quadratic Discriminant Function

The different covariance matrices are estimated by

$$\mathbf{S}_1 = \begin{bmatrix} 16.43^2 & -191.43 \\ -191.43 & 37.404^2 \end{bmatrix} \quad \text{and} \quad \mathbf{S}_2 = \begin{bmatrix} 18.06^2 & 133.50 \\ 133.50 & 29.89^2 \end{bmatrix}$$

By (2), the boundaries D_2 and D_2' are given by the quadratic discriminant function $qd(\mathbf{x})$, which, in turn, gives a curve with two branches, with equation:

$$y = -65.427 + 2.6428 \pm 1.2471 \sqrt{[1.0371 \times 10^5 - 1.4846 \times 10^3 x + 6.0673 \times 10 x^2]}.$$

The densities are now represented by two normal surfaces with different shapes, intersecting along 2 space curves G_2 and G_2', whose projections on the (x,y)-plane are the two curves D_2 and D_2' mentioned above (Figure 2 and Figure 3).

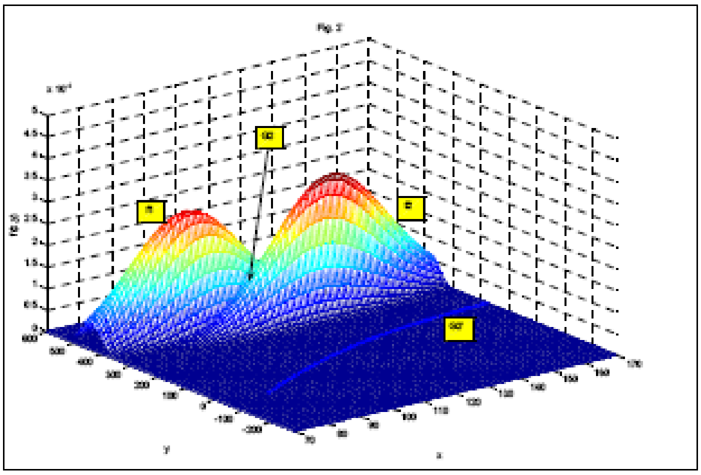

Figure 2

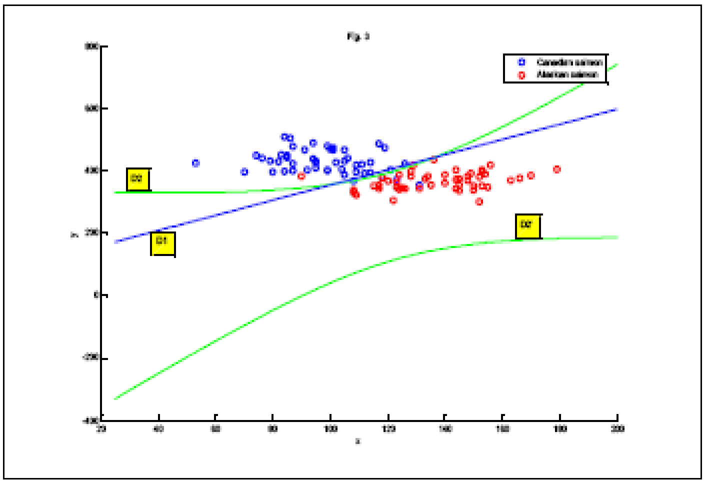

Figure 3

CONCLUSION

The use of the $g_{max}(\boldsymbol{x})$ function, and of its graph in R^3 (when $\boldsymbol{X}$ is two-dimensional), provides a presentation of the arguments behind discriminant analysis, using visual effects as well as equations, and is more effective, as the reactions of our students, from different domains, seem to show.

REFERENCES

Glick, N. (1973). Separation and probability of correct classification among two or more distributions. *Annals of the Institute of Statistical Mathematics*, 25, 373-382.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis* (4th edition). Saddle River, NJ: Prentice-Hall.

Pham-Gia, T. and Turkkan, N. (to appear). The Bayes error in classification, a Bayesian approach. *Statistical Methods and Applications*