Universe Scientific Publishing

Application of Fisher's Discriminant Method and Bayes' Discriminant Method Based on “R Language Analysis” as an Example

Manli Zhang, Guangli Xu, Hui Zhang

Taishan College, Tai'an 271000, China.

Abstract: Discriminant analysis is a statistical discriminant and grouping technique, which uses the sample data of the research object to find a discriminant rule, which is called a discriminant function, which is used to explain the difference between two or more groups of groups and can classify a new product. The data determines which of the known types the new product belongs to. This paper discusses the discriminant analysis of Fisher's discriminant method and Bayes' discriminant method, and shows the application of the two methods in real life and the difference between them on the basis of R language.

Keywords: Fisher's Test; Bayes' Test

1. Fisher's test

The basic idea of Fisher's discriminant method [1] is to reduce the dimensionality of the original data, that is, to project the original data into a lower-dimensional space, and to seek a direction that can separate each class as much as possible, that is, to seek to make the mean of each type to the total mean. The distance is the largest, but the difference of each class is as small as possible, that is, a discriminant rule that the sum of squared deviations within the group is as small as possible[1]. The advantage of Fisher's discriminant method is that it does not need to know the distribution type of the population in advance, but it is usually assumed that there is the same covariance matrix. Let's take the Fisher's discriminant of two groups as an example, as shown in Figure 1:

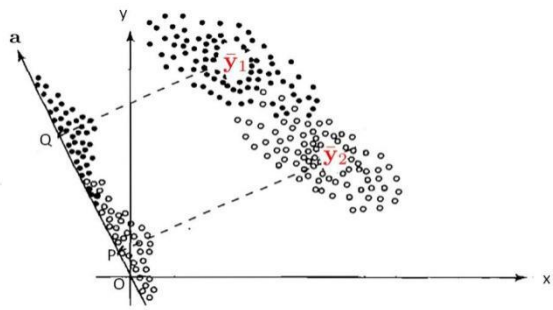

Figure 1 Comparison Chart of Fisher's Discriminant Methods

When the two groups are in the form shown in the figure above, no matter whether the two groups are projected on the x-axis or the y-axis, the two groups cannot be well separated. We need to find a direction a so that the two groups can be projected After that, it can be separated to the greatest extent possible. From the figure, we can see that the vector a is the direction that needs to be determined. If the direction is determined, the judgment criteria will be clear.

2. Bayes' discriminant

Fisher's discriminant method is simple and practical to calculate, but it also has two shortcomings [1]: one is that the discriminant method has nothing to do with the probability of the overall occurrence; This is unreasonable in practice. For

example, when we judge whether a company is an ST stock company, we usually acquiesce that it is not an ST, that is, it has a certain prior probability. Fisher's discriminant method does not consider the prior probability. The cost of not diagnosing a terminal illness in real life is significantly greater than misdiagnosing a patient with a terminal illness, so the cost of misdiagnosis should be taken into account at this time. Fisher's discriminant method does not consider the cost of misjudgment. Bayes is a judgment that considers both the prior probability and the loss of misjudgment, that is, according to the overall prior probability, the average loss of misjudgment is minimized. This is more practical in practice.

3. Example application of two discriminant methods

Discriminatory methods are widely used [2], such as animal and plant classification, medical disease diagnosis, community classification, meteorology, commodity classification, admissions applicant classification, occupational ability classification, and human archaeological chronology and ethnic classification. . We take the application of the above two discrimination methods in the admission of students as an example. The background is that the admissions office of a business school uses the indicators composed of X1 undergraduate grade point average (GPA) and X2 business management graduate admission test (GMAT) scores to help They decide whether an applicant is admitted to the college's graduate program. And these data are divided into three groups G1: admitted group, G2: not admitted group, G3: pending group. First of all, we make a scatter plot of GPA and GMAT. According to the scatter plot, it can be seen intuitively that the three groups have good clustering and differentiation effects, that is, the second group of sample point data with lower GPA and GMAT (not admitted group); GPA The sample points with intermediate GPA and GMAT belong to the third group (undetermined group); the sample points with higher GPA and GMAT belong to the first group (admitted group). Using the function lda() in the MASS package in R language to analyze, we need to use two discriminant functions to separate them Z1=-5.008*gpa-0.0086*gmat[2]; Z2=1.8767*gpa-0.01445*gmat These two The contribution rates of the individual discriminant functions are 0.9676 and 0.0327, respectively. We can verify this contribution rate through the histogram of each group of the discriminant function. Here, only the histogram corresponding to the first discriminant function is shown in Figure 2.

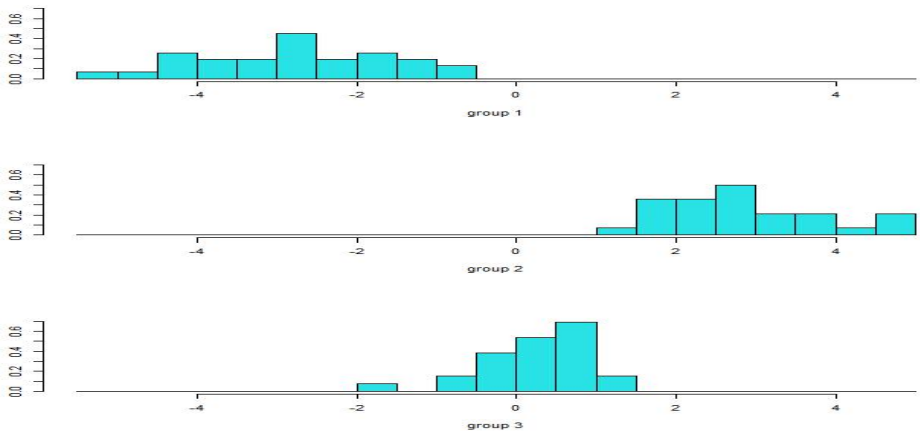

Figure 2 Corresponding to the histogram of the first discriminant function

From Figure 2, we can see that the histogram of the first group of discriminant functions shows that the first group and the second group have no overlap, and the second group and the third group have some overlap. We assume that the data obey the bivariate normal distribution and obtain the Bayes discriminant function through the function lda() of the R language and assume that the prior probabilities are 0.3, 0.6, and 0.1. The Bayes discriminant function is Z1=-5.052*gpa-0.0082*gmat; Z2=1.7579*gpa- 0.0146*gmat

The false positive rate of the two methods is 0.082, so the discriminant effect is still good.

4. Conclusion

This paper introduces the theory and application of Fisher's discriminant analysis and Bayes' discriminant analysis.

Fisher's discriminant analysis is a method based entirely on data, which does not require distribution assumptions but needs to have the same covariance; Bayes' discriminant analysis requires explicit distribution assumptions that are not It needs to have the same covariance and can consider the prior probability and misjudgment loss. In the specific case analysis, the statistical analyst can choose the appropriate method according to the situation of the data.

References

[1] Wu XZ. Multivariate Statistical Analysis [M]. Beijing: Renmin University of China Press, 2019.

[2] Wang BH. Multivariate Statistical Analysis and R Language Modeling [M]. Guangzhou: Jinan University Press, 2010.

Fund Project: Taishan College Teaching Reform and Research Project.

About the Author：

1. Manli Zhang(1985.9-),Female,Han Nationality,Shandong Tai'an people,Lecturer, master student, research direction: applied statistics.

2. GuangLi Xu (1981.10-),Female,Han Nationality,Shandong Tai'an, PhD candidate, research direction: resource and environmental economy.

3. Hui Zhang(1994.6-),Female,Han Nationality,Shandong Tai'an native, teaching assistant, master student, research direction: applied statistics.