Evaluation of Discriminant Function Classification Performance

[1] Opara, P. N
(oparap201@gmail.com),
Department of Maths/Statistics
Imo State Polytechnics

[2] Prof, S. I Onyeagu
(si.onyeagu@unizik.edu.ng),

[3]Prof, G. A Osuji
(ga.osuji@unizik.edu.ng)
Department of Statistics –
Nnamdi Azikiwe University

https://doi.org/10.56201/ijasmt.vol.12.no10.2026pg38.43

Abstract
If $f_i(x)$ is multivariate normal with mean μ and covariance Σ, we can easily calculate these rates. When the parameter is not known a number of error rates may be defined. The function T (R, F) defines the error rates. The first argument is the presumed distribution of the observation that will be classified.

Keyword: *Multivariate, Mean and Covariance*

1. Introduction

One of the objectives of evaluating a discriminant function is to determine its performance in the classification of function observations. When the parameters are known, the error rate is

$$T(R, f) = q_1 \int R_2 f_1(x)dx + q_2 \int R_1 f_2(x)dx \qquad (3.57)$$

This is the error rate that would hold if we know the parameters of the distribution. The optimum classification region is

$$R_1 = \{X: D_1(X) = [X - 1/2(\mu_1 + \mu_2)]'\sum^{-1}(\mu_1 - \mu_2) > 0\} \qquad (3.58)$$

And

$$R_2 = \left\{X: D_1(X) = \left[X - \frac{1}{2(\mu_1 + \mu_2)}\right]' \sum^{-1}(\mu_1 - \mu_2) > 0\right\} \qquad (3.59)$$

The error rates within each group are

$$\left.\begin{aligned} P_1 &= \int R_2 f_1(x)dx \\ &= P[D_1(x) < 0 | x \in \pi_1] \\ &= \Phi(-\delta/2) \end{aligned}\right\} \qquad (3.60)$$

$$\left.\begin{aligned} P_2 &= \int R_2 f_1(x)dx \\ &= P[D_1(x) > 0 | x \in \pi_2] \\ &= \Phi(-\delta/2) \end{aligned}\right\} \qquad (3.61)$$

$$T(R,F) = q_1\Phi(-\delta/2) + q_2\Phi(\delta/2)$$
$$= 1/2\Phi(-\delta/2) + 1/2\Phi(\delta/2)$$
$$= \Phi\left(-\frac{\delta}{2}\right) \quad (3.62)$$

This is the error rate for assessing the expected performance of a particular discriminant function.

The actual error rate is

$$T(R, f) = q_1 \int R_2 f_1(x)dx + q_2 \int R_1 f_2(x)dx \quad (3.63)$$

Where

$$\hat{R}_1 = \{X: D_s(X) = [X - {}^1/_2 (\bar{X}_1 + \bar{X}_2)]' S^{-1}(\bar{X}_1 - \bar{X}_2) > 0\}$$

$$\left.\begin{aligned} P_1 &= \int_{\hat{R}_2} f_1(x)dx \\ &= P[D_s(x) < 0 | x \in \pi_1] \\ &= \Phi\left[\frac{-D_s(\mu_1)}{\sqrt{V_D}}\right] \end{aligned}\right\} \quad (3.64)$$

Where

$$D_s(\mu_1) = [\mu_1 - {}^1/_2 (\bar{X}_1 + \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2)]$$
$$V_D = (\bar{X}_1 - \bar{X}_2)' S^{-1} \Sigma S^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$\left.\begin{aligned} P_2 &= \int_{\hat{R}_1} f_2(x)dx \\ &= P[D_s(x) > 0 | x \in \pi_1] \\ &= \Phi\left[\frac{D_s(\mu_2)}{\sqrt{V_D}}\right] \end{aligned}\right\} \quad (3.65)$$

Where

$$D_s(\mu_2) = [\mu_2 - {}^1/_2 (\bar{X}_1 + \bar{X}_2)' S^{-1}(\bar{X}_1 - \bar{X}_2)]$$

$$T(\hat{R}, f) = {}^1/_2 \Phi\left[\frac{-D_s(\mu_1)}{\sqrt{V_D}}\right] + {}^1/_2 \Phi\left[\frac{D_s(\mu_2)}{\sqrt{V_D}}\right] \quad (3.66)$$

This formula will not be of much use unless the estimates of μ_1, μ_2 and Σ are used. The actual error rates are most needed when performing sampling experiments.

While the actual error rates are of interest in assessing the expected performance of a particular discriminant function, the expected error rates, obtained by considering

$$\bar{X}_1, \bar{X}_2 \text{ and S in } P_i = \Phi\left[(-1)^1 \frac{D_s(U_i)}{\sqrt{V_D}}\right] \; i = 1,2 \quad (3.67)$$

As random variables are more appropriate when considering the expected Performance of the classification rule based on randomly chosen samples of sizes n_1 and n_2 from π_1 and π_2 respectively.

The sample discriminant function for discriminating between two groups is $D_s(X) = [X - {}^1/_2 (\bar{X}_1 + \bar{X}_2)]' S^{-1}(\bar{X}_1 - \bar{X}_2)$ and its mean in the ith group is

$$D_s(U_i) = [U_i - {}^1/_2 (\bar{X}_1 + \bar{X}_2)]'^{S^{-1}(\bar{X}_1 - \bar{X}_2)} \quad (3.68)$$

and variance is

$$V_D = (\bar{X}_1 - \bar{X}_2)'^{S^{-1}} \Sigma S^{-1} (\bar{X}_1 - \bar{X}_2) \quad (3.69)$$

We consider estimates of P_1 and P_2, the conditional probabilities of misclassification by replacing $D_S(U_i)$ and V_D by their expected values over all samples of size n_1 and n_2. We need expectation of S^{-1} and $S^{-1}\Sigma S^{-1}$

Lachenbruch (1968) has shown that

$$E(S^{-1}) = \frac{n_1 + n_2 - 2}{n_1 + n_2 - k - 3}\Sigma^{-1} \Big\} = C_1\Sigma^{-1} \quad (3.70)$$

And

$$E(S^{-1}\Sigma S^{-1}) = \frac{(n_1 + n_2 - 3)(n_1 + n_2 - 2)^2\Sigma^{-1}}{(n_1 + n_2 - k - 2)(n_1 + n_2 - k - 3)(n_1 + n_2 - k - 5)} \Big\} = C_2\Sigma^{-1} \quad (3.71)$$

Therefore

$$E[D_s(U_i)] = trE[D_s(\mu_i)]$$
$$= trE\{(\bar{X}_1 - \bar{X}_2)[\mu_1 - (\bar{X}_1 + \bar{X}_2)]'S^{-1}\}$$
$$= \text{trE}\left\{\left[(\bar{X}_1 - \bar{X}_2)\mu_i - {}^1/_2\left(\bar{X}_1\bar{X}_1{}^1 + {}^1/_2\,\bar{X}_2\bar{X}_2{}^1\right)\right]C_1\Sigma^{-1}\right\}$$
$$= \text{tr}\left[(\mu_1 - \mu_2)\mu_1{}^1\Sigma^{-1} - {}^1/_2\left[\frac{n_1 - n_2}{n_1 n_2}\right] - {}^1/_2\,(\mu_1 + \mu_2)^1\Sigma^{-1}(\mu_1 - \mu_2)C_1\right]$$
$$= {}^1/_2\,(\mu_1 - \mu_2)^1\Sigma^{-1}(\mu_1 - \mu_2)(-1)^{i+1} - k\left[\frac{n_2 - n_1}{n_1 n_2}\right] \quad (3.72)$$

Where

$$\delta^2 = (\mu_1 - \mu_2)^1\Sigma^{-1}(\mu_1 - \mu_2)$$

Similarly, we have for

$$E(V_D) = E(\bar{X}_1 - \bar{X}_2)^1S^{-1}\Sigma^{-1}S^{-1}(\bar{X}_1 - \bar{X}_2)$$
$$= \text{trE}[(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)^1S^{-1}\Sigma^{-1}S^{-1}]$$
$$= \text{trE}[(\bar{X}_1 - \bar{X}_2)(\bar{X}_1 - \bar{X}_2)^1\Sigma^{-1}]C_2$$
$$= \text{tr}\{[(\mu_1 - \mu_2)(\mu_1 - \mu_2)^1 + ({}^1/_{n_1} + {}^1/_{n_2})\Sigma]\Sigma^{-1}\}C_2$$
$$= C_2\left[(\mu_1 - \mu_2)^1\Sigma^{-1}(\mu_1 - \mu_2) + k\left[\frac{n_2 - n_1}{n_1 n_2}\right]\right] \quad (3.73)$$

Thus

$$E[D_s(\mu_1)] = \frac{n_1 + n_2 - 2}{2(n_1+n_2-k-3)}\,\delta^2(-1) - \frac{k(n_2 - n_1)}{n_1 n_2}$$

$$E(V_D) = \left[\delta^2 + \frac{k(n_2 - n_1)}{n_1 n_2}\right]\frac{(n_1+n_2-3)\,(n_1+n_2-2)^2}{(n_1+n_2-k-2)\,(n_1+n_2-k-3)\,(n_1+n_2-k-5)} \quad (3.74)$$

Lachenbruch (1968) stated that although $D_s(X)$ *is* normally distributed conditionally On $\bar{X}_1$, $\bar{X}_2$ and S, it is not unconditionally normally distributed. For n_1 and n_2 sufficiently large, the conditional is very close to normal. The approximately values of P_1 and P_2 for samples of sizes n_1 and n_2 are

$$\overline{P_1} = \Phi\left\{\frac{E[-D_s(\mu_1)]}{\sqrt{V_D}}\right\} \quad (3.75)$$

and

$$\overline{P_2} = \Phi\left\{\frac{E[-D_s(\mu_2)]}{\sqrt{V_D}}\right\} \quad (3.76)$$

$\bar{P}_1$*and* $\bar{P}_2$ *are approximations to the expected* actual error rate over all samples of size $n_1$and n_2 . It might be noted that most of the methods suggested estimates, the expected actual error rate. The simplest estimator of the actual error rate

$$P_1 = \Phi\left[(-1)^1\left[\frac{D_s(\mu_1)}{\sqrt{V_D}}\right]\right]$$

i = 1, 2 is the D method obtained by replacing the parameters in it by their samples based estimators. This yields estimators

$$\hat{P}_1 = \hat{P}_2 = \Phi\left(-D/2\right) \qquad (3.77)$$

This estimator is clearly also obtained by estimating δ^2 by D^2 in the expression for the optimum error rate

$T(R,f) = \phi(-\delta/2)$

The re-substitution method has the advantages of being, to calculate and not requiring distributional assumptions. However, it is seriously biased and underestimates the expected actual error. Another estimate of the expected actual error rate based on the expected mean and variance of D_s (X) is the Ū-method. Lachenbruch (1968) used this approach to determine the necessary sample size to be within a given tolerance of its optimum error rate.

A simple estimator of the expected actual error rate is obtained by replacing δ^2 by its unbiased estimator

$$D^{*2} = \frac{n-k-1}{n} D^2 - \frac{K(n_1 + n_2)}{n_1 n_2} \qquad (3.78)$$

where

$n = n_1 + n_2 - 2$

This yields

$$P_1 = \Phi\left[-\frac{1}{2} \sqrt{\frac{(n-k-1)D^2}{n} - \frac{k(n_1+n_2)}{n_1 n_2}} \right] \qquad (3.79)$$

The apparent drawback of this estimator is that it can produce a negative estimate of the Mahalanobis distance. Costanza and Afifi (1979) have shown that this can be only occur if the two samples are so close together as to make any classification a dubious proposition and that the use of the estimator with zero replacing any negative distance is satisfactory.

Okamoto's expansion for the distribution of the two-group sample discriminant function Ds(X) also estimates the expected actual error rate. Fatti, Hawkins and Raath (1982) have also shown given differences between some of the estimators expected actual error rate.

Hills (1966) has shown that the expected actual error rate is less than the optimum Error rate, whereas the actual error rate is in fact greater than it. Thus we have

$$E[\Phi(-D/2)] < \Phi(-\delta/2) < \Phi(-D/2) \qquad (3.80)$$

Onyeagu and Adegboye (1996) compare the seven methods of estimating error rates by sampling experiments. These methods are the R, D, DS, O, OS, U and Ū methods. Since the classification rule is invariant under the linear transformation of the variates, the distribution of the discriminant function could be made to be independent of the covariance matrix and to depend only on δ^2, n_1, n_2 and K. They set $\sum = 1$, $u_{1=0, U_2=(\delta 0 \ldots\ldots 0)'}$ where δ varied from 0.5 to 2.0. Since the mean vector for each group was known, it was possible to determine the exact probability of misclassification for each sample discriminant function.

Onyeagu and Adegboye (1996) generated sample of size $n_1 and\ n_2$ from N_k $(0_k\ 1)$ $((\delta 0 \ldots . 0)', 1)$ respectively using the standard program NAG. This was done on a CYBER 72 computer system. This generation of data was done for some configuration of $n_1, n_2, \delta^2\ and\ k$. The experiment was repeated five times and the mean for each combinationδ of $n_1, n_2, \delta^2$and k recorded.

Onyeagu and Adegboye (1996) concluded that with small samples and small values of δ^2, the OS method is the best estimator of P(MC). For small number of variables, it is the D method that performed best. When $n_1\ and\ n_2$ are varied from 50 and 38 to 100 and 75, the OS, D and DS methods are the best estimators of the probability of misclassification P(MC). The OS and D are also best for increasing values of δ^2. Finally as k increases from 3 to 10 the best overall methods are the D DS methods. No one method was found to be uniformly best for all situations. In all combinations of $n_1, n_2, \delta^2\ and\ k$ considered the R

method performed worse that all the methods followed by U and O methods.

Onyeagu and Adegboye (1996) concluded that the R, U and O methods are in no way competing with the D, DS, OS and $\bar{U}$ *methods* and therefore should not be considered as good procedures for estimating the probability of misclassification P(MC).

Reference:

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fan, J., Guo, S., & Hao, N. (2012). Variance estimation using refitted cross-validationin ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B, 74*(1), 37–65.

Fan, J., & Lv, J. (2010). A Selective Overview of Variable Selection in High-Dimensional Feature Space. *Statistica Sinica, 20*(1), 101–148.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics, 9*(6), 1218–1228.

Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery andclass prediction by gene expression monitoring. *Science, 286*(5439), 531– 537.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical Report*, University of Toronto.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary

Lachenbruch, P. A. (1968). 0n expeated probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correction coefficient Biometrics 24(4), 823-834

Onyeagu and Adegboye (1996). The first course in multivariate statistical analysis, 2003(240-248).observations. *Annals of Statistics, 17*(3), 1217–1241.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. Springer Series in Statistics.

Shao, J., & Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics, 17*(3), 1176–1197.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29*(2), 614.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics, 14*(4), 1261–1295.