International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 282-285 www.ijaem.net ISSN: 2395-5252

Bi-Normal Reciver Opareting Characteristic Model

Dr. Ch. Prasuna
Department of Statistics
Vikrama Simhapuri University
Assistant Professor on Contract

Date of Submission: 15-09-2023 Date of Acceptance: 25-09-2023

ABSTRACT:
The ROC curve is a graphical method that summarizes how well a binary classifier can discriminate between two populations, often called the "negative" population (individuals who do not have a disease or characteristic) and the "positive" population (individuals who do have it). There is a theoretical model, called the bi-normal model that describes the fundamental features in binary classification. The model assumes a set of scores that are normally distributed for each population, and the mean of the scores for the negative population is less than the mean of scores for the positive population.
KEYWORDS: ROC curve, Bi-normal Curve

I. RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

The word ROC analysis had its origin in Statistical Decision Theory as well as in Signal Detection Theory (SDT) and was used during II World War for the analysis of radar images (Green and Swets (1966), Bamber (1975), Egan(1975)). In these studies, the objective is mainly to distinguish between the two possible outcomes of a dichotomous event like signal/no-signal or diseased/healthy. The ROC curve is a graphical representation of the performance of a test or marker. It is a plot of TPF (Sensitivity) against FPF (1-Specificity) and lies in the unit square. Given a marker, at each possible cutoff value, the TPF and FPF are calculated and plotted as the ROC curve.

The ROC curve was first introduced in the biomedical area by Lusted (1960) for medical imaging applications but it became a much popular statistical tool after the publication of Swets and Pickkets (1982). Two excellent reviews of ROC methodology applied in the biomedical area are given by Zhou et al.(2002) and Pepe (2003).

Much of the work in the area of ROC curves was reported by Green and Swets (1966). Metz (1978) stated that ROC analysis is useful to determine the discriminating ability of a diagnostic test. In later years, eventually ROC analysis made its way into other areas of medicine.

The prominent uses of ROC curve analysis are listed below.

1) Finding optimal cutoff point of a test
2) Evaluating the discriminatory ability of a test to correctly classify the subjects.
3) Comparing the efficacy of two or more tests for assessing the same disease
4) Comparing two or more observers measuring the same test

II. THE BI-NORMAL ROC MODEL

The bi-normal model is commonly adopted parametric model. It is based on the assumption that the test values d_i and h_j in the D and H groups are normally distributed i.e di ~ N (μ_D, S_D^2) and hj ~ $N(\mu_H, S_H^2)$ respectively. Suppose X and Y denote the diagnostic marker measurements, called test score S, on a continuous scale for the D and H populations respectively such that X ~ N (μ_x, σ_x^2) and Y ~ N (μ_y, σ_y^2). The bi-normal model gives an expression for the TPR as a function of the FPR expressed in terms of cumulative normal probability.

Theorem: Let X ~ N (μ_x, σ_x^2) and Y ~ N (μ_y, σ_y^2) such that $\mu_x > \mu_y$. Given the threshold c, the bi-normal ROC model is given by

$y(x) = \Phi[a + b\Phi^{-1}\{ x(c)\}]$

where Φ is the cumulative standard normal distribution and a, b are constants.

Proof: Since X ~ N (μ_x, σ_x^2) and Y ~ N (μ_y, σ_y^2) it follows that

$\frac{(S-\mu x)}{\sigma x}$ is a standard normal deviate of X and $\frac{(S-\mu y)}{\sigma y}$ is the standard normal deviate of Y.

According to the classifier threshold c, the FPR is a function of c and denoted by x(c)

$x(c) = P(S > c \mid Y)$

$= P\left(\frac{S-\mu y}{\sigma y} > \frac{C-\mu y}{\sigma y}\right)$

$x(c) = P\left(Z > \frac{C-\mu y}{\sigma y}\right)$

where Z is a standard normal deviate of Y.

By symmetry property of the normal distribution, the FPR x(c) can be written as

$x(c) = 1-P\left(Z \le \frac{c-\mu y}{\sigma y}\right)$

$= P\left(Z \le \frac{\mu y-c}{\sigma y}\right)$

$= \Phi\left(\frac{\mu y-c}{\sigma y}\right)$

$\Rightarrow \quad x(c) = \Phi(Z)$

where $\Phi(\cdot)$ is the normal cumulative distribution function (cdf).

If Z_x is the value of the Z at this cdf, then

$Z_x = \Phi^{-1}[x(C)]$

Now $\Phi^{-1}[x(c)] = Z_x$ or Z

$\Phi^{-1}[x(c)] = \frac{\mu y-c}{\sigma y}$

and consider

$Z_x = \frac{\mu y-c}{\sigma y}$

$\mu_y - c = Z_x \sigma_y$

$\Rightarrow \quad c = \mu_y - Z_x \sigma_y$

(1)

The TPR, y at the given FPR, x is denoted by y(x) and given by

$y(x) = P(S > c \mid X)$

$= P\left(\frac{S-\mu x}{\sigma x} > \frac{c-\mu x}{\sigma x}\right)$

$\Rightarrow \quad y(x) = P\left(Z > \frac{c-\mu x}{\sigma x}\right)$

Due to the symmetry property of the normal distribution y (x) can be written as

$y(x) = 1-P\left(Z \le \frac{c-\mu x}{\sigma x}\right)$

$= P\left(Z \le \frac{\mu x-c}{\sigma x}\right)$

So $y(x) = \Phi\left(\frac{\mu x-c}{\sigma x}\right)$

From (1.5.1),

$y(x) = \Phi\left(\frac{\mu x-\mu y+Zx*\sigma y}{\sigma x}\right)$

$= \Phi\left(\frac{\mu x-\mu y}{\sigma x} + \left(\frac{\sigma y}{\sigma x}\right)Zx\right)$

$\Rightarrow \quad y(x) = \Phi(a + bZx)$, where $a = \frac{\mu x-\mu y}{\sigma x}$ and $b = \frac{\sigma y}{\sigma x}$.

Hence the ROC curve is of the form

$y(x) = \Phi(a + bZx)$

$\Rightarrow \quad \Phi^{-1}[y(x)] = a + b\,\Phi^{-1}[x(c)]$

Hence the proof.

The constants a and b are estimated by the method of maximum likelihood. Obviously both a and b are non-negative since it is assumed that $\mu_x > \mu_y$.

Several authors have explored interesting characteristics of the bi-normal model which are outlined below.

- Green and Swets (1966), Metz et al. (1998) have shown the bi-normal model for ROC visually as TPR (c) = $\Phi(a + b\,\Phi^{-1}(FPR(c)))$; $c \in R$
- Also the bi-normality assumption implies a perfect linear relationship between TPR and FPR on deviate axes because
  $\Phi^{-1}(TPR(c)) = (a + b\,\Phi^{-1}(FPR(c)))$; $c \in R$
- Lloyd (1998) considered a special case to interpret the meaning of the parameters a and b in the bi-normal model. According to him If $F = N(0, \sigma^2)$, $G = N(\delta, \sigma^2)$ then a = 0 and b = 1. The intercept a is also equal to the square root of the Mahalonobis distance between F and G distribution functions of the H and D populations respectively. Hence the parameters a and b qualify the standardized separation and the ratio of standard deviation of the two random variables X and Y.

III. PROPERTIES OF BI-NORMAL ROC CURVE

Let us define the ROC in more familiar mathematical notation as the curve y= h(x), where y is the true positive rate TPR and x is the false positive rate FPR defined at a threshold c.

Three main properties of the bi-normal ROC Curve are briefly explained below.

Property-1

y = h(x) is a monotone increasing function in the positive quadrant, lying between y = 0 at x = 0 and y = 1 at x = 1.

Proof: The scores are arranged in such a way that both x (t) and y(t) increase and decrease together as c varies. Moreover, $\lim_{c\to\infty} x(c) = \lim_{c\to\infty} y(c) = 0$ and $\lim_{t\to-\infty} x(c) = \lim_{t\to-\infty} y(c) = 1$, which establishes the result.

Property-2

The ROC curve is unaltered if the classification scores undergo a strictly increasing transformation.

Proof: Suppose that U = Φ (S) is a strictly increasing transformation. Then there exists two values S_1 and S_2 such that $U_1 = \Phi(S_1)$ and $U_2 = \Phi(S_2)$.

It follows that $S_1 > S_2 \leftrightarrow U_1 > U_2$

Consider the point on the ROC curve for S at threshold value c, and let $v = \Phi(c)$. Then it follows that

$P(U > v \mid D) = P(\Phi(S) > \Phi(c) \mid D) = P(S > c \mid D)$

and

$P(U > v \mid H) = P(\Phi(S) > \Phi(c) \mid H) = P(S > c \mid H)$

so that the same point exists on the ROC curve for U.

The same argument holds in the reverse way and it hence the two curves are identical.

Property 3

The slope of the ROC at the point with threshold value c is is given by

$$\frac{dy}{dx} = \frac{P(c/D)}{P(c/H)}$$

Proof: First note that

$$y(t) = P(S > c|D) = 1 - \int_{-\infty}^{t} P(s|D)ds,$$

So that

$$\frac{dy}{dc} = -P(c \mid D).$$

Thus

$$\frac{dy}{dx} = \frac{dy}{dc}\frac{dc}{dx} = -P(c \mid D)\frac{dc}{dx}$$

Moreover,

$$X(c) = P(S > c \mid N) = 1 - \int_{-\infty}^{t} P(s|H)ds,$$

so that

$$\frac{dx}{dc} = -P(c \mid H).$$

Also

$$\frac{dc}{dx} = 1 \Big/ \frac{dx}{dc}$$

Therefore $\frac{dy}{dx} = -P(c \mid D) / -P(c \mid H)$ and the result follows.

IV. AREA UNDER THE BI-NORMAL ROC CURVE

In parametric approach, area under the bi-normal ROC curve is the summary index of the performance of the diagnostic test denoted by AUC.

In general AUC ranges from (0.5, 1). So the upper bound is 1.0, while for the case of random allocation AUC is the area under the chance diagonal so the lower bound is 0.5. AUC can be used to know the accuracy or efficiency of a test. The accuracy of a diagnostic test is the traditional academic point system is explained below.

If AUC = 0.5 means there is no discrimination by the test. An area of 1 represents a perfect test.

Mathematically the AUC is defined as

$$AUC = \int_0^1 y(x)dx$$

Hanley and McNeil (1982), Bradley (1997), Farragi and Reiser (2002) have studied area under the bi-normal ROC curve.

One of the very useful consequences of bi-normal ROC model is that its AUC can be derived very easily, and has a very simple form.

If X and Y are the scores allotted to randomly and independently chosen individuals from D and H populations respectively.

then AUC can be defined as

$AUC = P(X > Y)$

$AUC = P(X - Y > 0)$

If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ then $X - Y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$. Hence if Z denotes a standard normal random variate,

$$AUC = P\left(Z > 0 - \left(\frac{\mu x - \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)\right)$$

$$= 1 - \Phi\left(\frac{-\mu x + \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)$$

$$= \Phi\left(\frac{\mu x - \mu y}{\sqrt{\sigma x2 + \sigma y2}}\right)$$

Dividing numerator and denominator by σ_x, then

$$AUC = \Phi\left(\frac{\frac{\mu x - \mu y}{\sigma x}}{\frac{\sqrt{\sigma x2 + \sigma y2}}{\sigma x}}\right)$$

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b2}}\right)$$

where $a = \frac{\mu x - \mu y}{\sigma x}$ and $b = \frac{\sigma y}{\sigma x}$.

Thus bi-normal AUC is simply the cumulative standard normal probability and can be easily evaluated using statistical tables or with the Excel function NORMSDIST().

V. CONCLUSION

In summary, the bi-normal ROC curve illustrates fundamental features of the binary classification problem. Typically, you use a statistical model to generate scores for the negative and positive populations. The bi-normal model assumes that the scores are normally distributed and that the mean of the negative scores is less than the mean of the positive scores. With that assumption, it is easy to use the normal CDF function to compute the FPR and TPR for any value of a threshold parameter. We can graph the FPR and TPR as functions of the threshold parameter, or you can create an ROC curve, which is a parametric curve that displays both rates as the parameter varies. The bi-normal model is a useful theoretical model and is more applicable than you might think. If the variables in the classification problem are multivariate normal, then any linear classifier results in normally distributed scores. In addition, Krzandowski and Hand (2009, p. 34-35), state that the ROC curve is unchanged by any monotonic increasing transformation of scores, which means that the bi-normal model applies to any set of scores that can be transformed to normality.

REFERENCES

[1]. R.Vishnu Vardhan, et al., (2012), Estimating of Area under the ROC Curve Using Exponential and Weibull distributions, Bonfring International Journal of Data Mining , ISSN 2277 ,Vol-2, 52- 56.

[2]. Krzanowski W. D and J.Hand (2009), ROC Curves for Continuous data, Monographs on statistics and Applied Probability, CRS Press, Taylor and Francis Group, LLC.

[3]. Lasko, T.A. et al (2005), The use of receiver operating characteristics curves in biomedical informatics, Journal of Biomedical Informatics, 38: 404 – 415.

[4]. Dodd, L.E. and Pepe, M.S. (2003), Semiparametric regression for the area under the receiver operating characteristic curve. Journal of the American Statistical Association, 98, 409-417.

[5]. David Farragi and Benjamin Reiser (2002), Estimation of the area under the ROC curve, Statistic in Medicine, 21: 3093 - 3106.

[6]. Lloyd, C.J. (1998), The Use of smoothed ROC curves to summarize and compare diagnostic systems, Journal of American Statistical Association, 93: 1356 – 1364.

[7]. Bradley (1997), The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, 30, 1145-1159.

[8]. James A Hanley, Barbara J Mc Neil (1982), A Meaning and Use of the area under a Receiver Operating Characteristic (ROC) curves, Radiology, 143: 29 - 36.

[9]. Metz C.E (1978), Basic Principles of ROC analysis, Seminars in Nuclear Medicine, 8: 283 – 298.

[10]. Lusted (1971), Signal detectability and medical decision-making, Science, 17:1217-1219.

[11]. Green, D.M. and Swets, J.A. (1966), Signal Detection theory and Psychophysics, Wiley, Newyork.

[12]. Betinec. M (2008), Testing the difference of the ROC Curves in Bi-exponential Model, Tatra Mountains Mathematical Publications, 39, 215-223.

[13]. Zhou and Lin (2007), Semiparametric maximum likelihood estimation for ROC curves of continuous- scale tests, http:// www. bepress.com/uwbiostat/paper325, University of Washington Working Series paper.

[14]. Zou, K.H and Shapiro, DE (1997), Smooth nonparametric Receiver operating characteristic (ROC) curves for continuous diagnostic tests, Statistics in Medicine, 16: 2143-2156.

[15]. Swets, J.A. and Pickett, R.M. (1982). Evaluation of diagnostic systems: Methods from signal detection theory. New York; Academic.

[16]. Colton t, Statistics in medicine. Boston: Little, Brown and Company, 1974: 168.

[17]. Egan, J.P. (1975). Signal detection theory and ROC Analysis. Academic Press, New York.