IJIRR

International Journal of Information Research and Review
Vol. 05, Issue, 03, pp.5270-5278, March, 2018

REVIEW ARTICLE

SOME ROC MODELS WITH SKEWED DISTRIBUTIONS

[1]Dr. Ch. Prasuna and [2]Dr. B.Prabhakara Rao

[1]Department of Statistics, Vikrama Simhapuri University, Kakuturu, Nellore (Dist), India
[2]Department of Political Sciences, Vikrama Simhapuri University, Kakuturu, Nellore (Dist), India

ARTICLE INFO

Article History:

Received 10th December, 2017
Received in revised form
22nd January, 2018
Accepted 04th February, 2018
Published online 30th March, 2018

Keywords:

ROC, AUC,
Skewed distributions

ABSTRACT

Parametric ROC Curve is a mathematical model for assessing the performance of a binary classifier assuming that the data in both classes follow a known distribution. Construction of ROC model with skewed distributions has been the interest of researchers in the recent times. In this thesis we have focused on developing theoretical ROC model when both the distributions are skewed. We have used exponential and generalized exponential distributions to arrive at the new model. In this paper we have going to some important skewed distributions and their use in ROC modeling are discussed. We the skewed distribution and derive a new method of determining the AUC by utilizing the interval estimates of Scale and location parameters the distribution.

INTRODUCTION

Skewed distribution is a distribution that has lack of symmetry (Szekely, Mori (2001)). A distribution is said to be *skewed* when the data points if

- In other words, the right and the left side of the distribution are shaped differently from each other.
- Mean, Median and Mode fall at different points, i.e., Mean ≠ Median ≠ Mode.
- Quartiles are not equidistant from median, and
- Cluster more toward one side of the scale than the other, creating a curve that is not symmetrical.

There are several continuous distributions which are skewed. Many skewed distribution is characterized by location, scale and shape parameters. However distributions like exponential and gamma are often described by a single parameter. It is interesting note that discrete theoretical distributions are always skewed. Some important continuous skewed distributions are presented in Table 1. Skewed distributions have potential applications in the field of reliability and survival analysis. In medicine some neurological parameters are known to follow skewed distributions. We focused on developing ROC models with test score following a) Exponential b) Generalized exponential and c) Log normal distribution in both groups.

ROC Models with Skewed distributions

Construction of ROC model with skewed distributions has been the interest of researchers in the recent times. In this thesis we have focused on developing theoretical ROC model when both the distributions are skewed. We have used exponential and generalized exponential distributions to arrive at the new model. In the following section an ROC Models with exponential distribution are discussed.

Bi-exponential ROC Model

Suppose the distributions of the test values follow exponential distribution in both D and H groups with means λ_D, λ_H respectively.

Corresponding author: **Dr. Ch.Prasuna,**
Department of Statistics, Vikrama Simhapuri University, Kakuturu, Nellore (Dist), India.

The probability density function of X in the D group is

Table 1. Some important continuous Skewed distributions

S.No	Distribution	Density function
1.	Exponential	$\frac{1}{\lambda}e^{-x/\lambda}$
2.	Gamma	$\frac{e^{-x}x^{\lambda-1}}{\Gamma\lambda}$
3.	Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^k}$
4.	Cauchy	$\frac{1}{\pi\sigma\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]}$
5.	Half normal	$\frac{\sqrt{2}}{\sigma\sqrt{\pi}}e^{\left(\frac{-x^2}{2\sigma^2}\right)}$
6.	Log normal	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\left(\frac{(\ln x-\mu)^2}{2\sigma^2}\right)}$
7.	Pareto	$\frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$
8.	Rayley	$\frac{x}{\sigma^2}e^{x^2/2\sigma^2}$
9.	Laplace exponential	$\frac{1}{2b}e^{\left(-\frac{\lvert x-\mu\rvert}{b}\right)}$
10.	Beta	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$
11.	Folded normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-\mu)^2}{2\sigma^2}}+\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x+\mu)^2}{2\sigma^2}}$
12.	Gumbel	$\frac{1}{\beta}e^{-(z+e^{-z})}$
13.	Logistic	$\frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$

$f(x) = \frac{1}{\lambda_D}e^{-x/\lambda_D}$, $\lambda_D > 0$, $x > 0$
$= 0$ otherwise

and in the H group the probability density function of y is

$f(y) = \frac{1}{\lambda_H}e^{-y/\lambda_H}$, $\lambda_H > 0$, $x > 0$
$= 0$ otherwise

Under this model the false positive rate with cutoff value 'c' is given by

$x(t) = P(X > t \mid H)$
$= 1 - P(X \leq t \mid H)$
$= 1 - \left(1 - e^{-t/\lambda_H}\right)$
$= e^{-t/\lambda_H}$
$\Rightarrow \ln(x(t)) = -t/\lambda_H$

This gives

$t = -\ln(x(t))\,\lambda_H$ ……(1)

Now the ROC Curve is given by

$y(t) = P(X > t \mid D)$
$= 1 - p(X \leq t \mid D)$
$= 1 - \left(1 - e^{-t/\lambda_D}\right)$
$= e^{-t/\lambda_D}$

From (1) it follows that

$y(t) = e^{\frac{\ln(x(t))\lambda_H}{\lambda_D}}$

$\Rightarrow y(t) = x(t)^{\frac{\lambda_H}{\lambda_D}}$

Then the Bi-exponential ROC Model is $y(t) = x(t)^{\beta}$, where $\beta = \frac{\lambda_H}{\lambda_D}$

The AUC is obtained as $\int_0^1 y(t)\, dFP$ and this reduces to (Vishnu vardhan, Sudesh Pundir and Sameera (2012))

$$AUC = \frac{\lambda_D}{\lambda_H + \lambda_D} \qquad \text{.....(2)}$$

Properties of bi-exponential distribution (Krzanowski and Hand (2009))

Property-1 bi-exponential ROC Curve is monotonically increasing.

Proof: Let us consider two false positive values P_1 and P_2 such that $P_1 < P_2$ be a strictly increasing function.
Since $P_1 < P_2$ which implies that $P_1^{\frac{\lambda_H}{\lambda_D}} < P_2^{\frac{\lambda_H}{\lambda_D}} \Rightarrow ROC(P_1) \leq ROC(P_2)$

Hence, the *bi-exponential ROC (model) Curve is monotonically increasing*.

Property-2 The bi-exponential ROC Curve is invariant under strictly increasing transformation.

Proof: Let 'S' denote the set of scores with $S \subset \Re$ and h(.) is strictly increasing function. Let a, b ϵ S and a < b, then by using the strictly increasing function, we can write h(a) < h(b).

The transformed random variables U and V from the respective healthy and diseased classes are

$P(U \leq t) = P[h(U) \leq h(t)]$

and $P(V \leq t) = P[h(V) \leq h(t)]$

Let us consider the points (x*(t), y*(t)) on the ROC Curve for the transformed scores

$x^*(t) = P\{h(U) > h(t) \mid H\}$
$= 1 - P\{h(U) \leq h(t)\}$
$= 1 - P(U \leq t) = x(t)$
$y^*(t) = P\{h(V) > h(t) \mid D\}$
$= 1 - P\{h(V) \leq h(t)\}$
$= 1 - P(V \leq t) = y(t)$

Thus, *the bi-exponential ROC Curve is invariant under monotonic transformation* of the test score.

Property-3 The slope of the ROC at the point with threshold value *c* is given by

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

Proof: First we observed that

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

$= e^{-c/\lambda_D}$

So that

$$\frac{dy}{dc} = -1/\lambda_D\, e^{-c/\lambda_D}$$

Thus

$$\frac{dy}{dx} = \frac{dy}{dc}\frac{dc}{dx} = -1/\lambda_D\, e^{-c/\lambda_D}\frac{dc}{dx}$$

Moreover,

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

so that

$$\frac{dx}{dc} = -1/\lambda_H\, e^{-c/\lambda_H}$$

Also

$$\frac{dc}{dx} = \frac{-\lambda_H}{e^{-c/\lambda_H}}$$

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

In the following section an ROC Models with Generalized exponential distribution are discussed.

Bi-Generalized exponential ROC Model

Let X and Y be the random variables indicating the test value in the H & D groups with underlying GE density functions.

$$g(x;\alpha_H,\lambda_H,\mu_H) = \frac{1}{\lambda_H} e^{-\left(\frac{x-\mu_H}{\lambda_H}\right)}\ \alpha_H = 1\ \&\ \lambda_H,\ \mu_H > 0,\ x > 0$$

$$g(y;\alpha_D,\lambda_D,\mu_D) = \frac{1}{\lambda_H} e^{-\left(\frac{x-\mu_D}{\lambda_D}\right)}\ \alpha_D = 1\ \&\ \lambda_D,\ \mu_D > 0,\ x > 0$$

Since the data on both X and Y is continuous, each data point serves as a possible cutoff denoted by t. At each cutoff t we find the false positive rate x (t) either by counting number of cases of interest or by using a formula (in case of theoretical model). When the location parameters is set zero then $\mu_H = \mu_D = 0$. Again with $\alpha_H = \alpha_D = 1$ we get the case of one parameter bi-exponential distribution model.

Under this model the false positive rate with threshold t is given by

x (t) = P (X > t | H)

= 1 - P (X ≤ t | H)

$$= 1 - \left[1 - e^{\frac{-(t-\mu_H)}{\lambda_H}}\right]$$

$$x(t) = e^{\frac{-(t-\mu_H)}{\lambda_H}}$$

$$\Rightarrow \ln(x(t)) = \frac{-(t-\mu_H)}{\lambda_H}$$

This gives

$t = \mu_H - \lambda_H * \ln(x(t))$ ……..(3)

Now the ROC curve is given by

Y(t) = P(X > t | D)

= 1 - P(X ≤ t | D)

$$= 1 - \left[1 - e^{\frac{-(t-\mu_D)}{\lambda_D}}\right]$$

$Y(t) = e^{\frac{-(t-\mu_D)}{\lambda_D}}$

From (2.3.3) it follows that

$Y(t) = e^{-\left(\mu_H - \lambda_H \ln(x(t)) - \mu_D\right)/\lambda_D}$

$$=>Y(t)=x(t)^{\beta} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \quad \ldots\ldots(4)$$

Then the Bi GE ROC model is of the form $Y(t) = x(t)^{\beta} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)}$ where $\beta = \frac{\lambda_H}{\lambda_D}$, and μ_H , μ_D , λ_H and λ_D are the location and scale of the test result in the two groups.

The AUC is obtained as $\int_0^1 y(t)\, dFP$ and this reduces to

$$AUC = \frac{\lambda_D}{\lambda_D+\lambda_H} e^{\left(\frac{\mu_D-\mu_H}{\lambda_D}\right)} \quad \ldots\ldots(5)$$

Properties of bi-generalized exponential distribution

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 bi- generalized exponential ROC Curve is invariant under strictly increasing transformation.

Proof: Let 'S' denote the set of scores with $S \subset \Re$ and h(.) is strictly increasing function. Let a, b $\in$ S and a < b, then by using the strictly increasing function, we can write h(a) < h(b).

The transformed random variables U and V from the respective healthy and diseased classes are

$P(U \le t)= P[h(U) \le h(t)]$

and $P(V \le t)=P[h(V)\le h(t)]$

Let us consider the points (x*(t), y*(t)) on the ROC Curve for the transformed scores

$x^*(t)=P\{h(U) > h(t) \mid H\}$
$=1-P\{h(U) \le h(t)\}$
$=1-P(U \le t)=x(t)$

$y^*(t)=P\{h(V) > h(t) \mid D\}$
$=1-P\{h(V) \le h(t)\}$
$=1-P(V \le t) = y(t)$

Thus, the bi- generalized exponential ROC curve *is invariant under monotonic transformation* of the test score.

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 we observed that

$y(t) = P(S > c \mid D) = 1-\int_{-\infty}^{t} P(s|D)ds$
$= e^{-\left(\frac{c-\mu_D}{\lambda_D}\right)}$

So that

$$\frac{dy}{dc} = -1/\lambda_D \, e^{-\left(\frac{c-\mu_D}{\lambda_D}\right)}$$

Thus

$$\frac{dy}{dx} = \frac{dy}{dc}\frac{dc}{dx} = -1/\lambda_D \, e^{-\left(\frac{c-\mu_D}{\lambda_D}\right)} \frac{dc}{dx}$$

Moreover,

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

so that

$$\frac{dx}{dc} = -1/\lambda_H \, e^{-\left(\frac{c-\mu_H}{\lambda_H}\right)}$$

Also

$$\frac{dc}{dx} = \frac{-\lambda_H}{e^{-\left(\frac{c-\mu_H}{\lambda_H}\right)}}$$

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

In the following section an ROC Models with log normal distribution are discussed.

Bi- lognormal ROC model

let us assume that the test scores X and Y are independent from the healthy and diseased population and X follows log normal distribution with parameters μ_H, σ_H^2 and X follows log normal distribution with parameters μ_D, σ_D^2.

The cumulative distribution function is defined by

$$F(x) = \Phi\left(\frac{lnx-\mu}{\sigma}\right) \qquad \text{……(6)}$$

where Φ is the standard normal distribution. Under this model the false positive rate with threshold t is given by (Amala ,Pundir (2012))

$x(t) = P(X > t \mid H)$
$= 1 - P(X \leq t \mid H)$
$= 1- \Phi\left(\frac{l-\mu_H}{\sigma_H}\right)$
$1-x(t) = \Phi\left(\frac{lnt-\mu_H}{\sigma_H}\right)$
$\Phi(\ln(t) - \mu_H) = (1-x(t))\,\sigma_H$
$\Rightarrow \ln(t) = \mu_H + \Phi^{-1}(1-x(t))\,\sigma_H$

This gives

$t = \exp(\mu_H + \Phi^{-1}(1-x(t))\,\sigma_H)$ (7)

Now the ROC curve is given by

$y(t) = P(X > t \mid D)$
$= 1 - P(X \leq t \mid D)$
$= 1- \Phi\left(\frac{\ln(t)-\mu_D}{\sigma_D}\right)$

Form (2.3.7) it follows that

$$y(t) = 1- \Phi\left(\frac{\mu_H+\Phi^{-1}(1-x(t))\sigma_H-\mu_D}{\sigma_D}\right)$$

$$= 1- \Phi\left(\frac{\mu_H-\mu_D}{\sigma_D}+\frac{\Phi^{-1}(1-x(t))\sigma_H}{\sigma_D}\right)$$

$$y(t) = 1- \Phi\left(a+\Phi^{-1}(1-x(t))b\right)$$

where

$$a=\frac{\mu_D-\mu_H}{\sigma_D} \text{ and } b=\frac{\sigma_H}{\sigma_D}.$$

Then AUC can be defends AUC = P (X > Y) = P (X – Y > 0)

Let lny = Y and lnx = X for diseased and healthy score respectively. Since it is easier to work with normal as compared to log normal distribution.

AUC = P (Y > X) = P (lny > lnx)

= P (y > x)

We know that $Z_x=\frac{c-\mu_H}{\sigma_H}$ follows N (0, 1) and $Z_y=\frac{c-\mu_D}{\sigma_D}$ follows N (0, 1).

If X ~ N (μ_H, σ_H^2) and Y ~ N (μ_H, σ_H^2) then X- Y ~ N (μ_H - μ_D, σ_H^2+ σ_D^2). Hence if Z denotes a standard normal random variate,

$$\text{AUC} = P\left(Z > 0 - \left(\frac{\mu_H-\mu_D}{\sqrt{\sigma_H^2+\sigma_D^2}}\right)\right)$$

$$= 1-\Phi\left(\frac{-\mu_H-\mu_D}{\sqrt{\sigma_H^2+\sigma_D^2}}\right)$$

$$=\Phi\left(\frac{\mu_H-\mu_D}{\sqrt{\sigma_H^2+\sigma_D^2}}\right)$$

Dividing numerator and denominator by σ_x, then

$$\text{AUC} = \Phi\left(\frac{\frac{\mu_H-\mu_D}{\sigma_H}}{\frac{\sqrt{\sigma_H^2+\sigma_D^2}}{\sigma_H}}\right)$$

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

where $a=\frac{\mu_H-\mu_D}{\sigma_H}$ and $b=\frac{\sigma_D}{\sigma_H}$.

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.

Properties of bi-lognormal distribution

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 bi- lognormal ROC Curve is invariant under strictly increasing transformation.

Proof: Let 'S' denote the set of scores with $S \subset$ □ and h(.) is strictly increasing function. Let a, b $\in$ S and a < b, then by using the strictly increasing function, we can write h(a) < h(b).

The transformed random variables U and V from the respective healthy and diseased classes are

$P(U \leq t)=P[h(U) \leq h(t)]$
and $P(V \leq t)=P[h(V) \leq h(t)]$

Let us consider the points (x*(t), y*(t)) on the ROC Curve for the transformed scores

$x^*(t)=P\{h(U) > h(t) \mid H\}$
$=1-P\{h(U) \leq h(t)\}$
$=1-P(U \leq t)=x(t)$
$y^*(t)=P\{h(V) > h(t) \mid D\}$
$=1-P\{h(V) \leq h(t)\}$
$=1-P(V \leq t) = y(t)$

Thus, the bi- log normal ROC curve is invariant under monotonic transformation of the test score.

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* | 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* | H).

Also

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

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

Thus for the above skewed distributions the ROC curve and its AUC can be evaluated without reference to statistical tables or functions.

REFERENCES

Amala, R. and Sudesh Pundir. 2012, Statistical Inference on AUC from A Bi-Lognormal ROC Model for Continuous Data, *International Journal of Engineering Science and Innovative Technology,* (IJESIT) Volume 1, Issue 2, November 2012, ISSN: 2319-5967.

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

Szekely, G. J. and Mori, T. F. 2001. "A characteristic measure of asymmetry and its application for testing diagonal symmetry", Communications in Statistics – Theory and Methods 30/8&9, 1633–1639.

Vishnu Vardhan, R., Sudesh Pundir and Sameera 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.

****