BASICS OF RECEIVER OPERATING CHARACTERISTIC CURVES

Dr.ch.prasuna

Assistant professor on contract
Department of statistics
Vikrama Simhapuri University

***Abstract*- The performance of a diagnostic test in the case of a binary predictor can be evaluated using the measures of sensitivity and specinfinity. However, in many instances, we encounter predictors that are measured on a continuous or ordinal scale. In such cases, it is desirable to assess performance of a diagnostic test over the range of possible cut points for the predictor variable. This is achieved by a receiver operating characteristic (ROC) curve that includes all the possible decision thresholds from a diagnostic test result. In this brief report, we discuss the salient features of the ROC curve, as well as discuss and interpret the area under the ROC curve, and its utility in comparing two different tests or predictor variables of interest.**

***Key Words*: Sensitivity, Specificity, ROC, AUC.**

1. Introduction

Classification problems frequently arise in medical diagnosis where it is required to classify a patient into one of the groups namely Healthy (H) and Diseased (D) basing on the data on a test conducted about the patient's parameters like Blood Glucose level, Blood Pressure etc. It is sometimes possible that a test shows positive (indicating the presence of disease) where as the true diagnosis may show a *negative* status. It means *test positive* need not imply a positive diagnosis and vice-versa. Since the classification is made into two categories, it is called ***binary classification***.

Most of the real problems are multivariate in nature. For instance, a patient is described in terms of several characteristics which include Anthropometric, pathological, physiological parameters. Such characteristics are called ***Biomarkers***. A biomarker is a feature that describes one of the groups (status). For instance body mass index (BMI) may be considered as a biomarker to classify persons into a state of hypertension or not. Certain markers are less precise or their measurement is difficult to certain patients. In such cases the clinical scientist looks for alternative individual markers or combination of markers by which a patient can be classified with minimum error.

Binary classification is a common requirement in medical context where a patient should be correctly classified into H or D groups basing on a test (marker). If the test result exceeds a cutoff value (called threshold) the patient is classified into D groups and treatment is started, otherwise declared healthy.

In reality every classification leads to error unless the classification rule is perfect. Consider a biomarker for which there is already a cutoff which is widely accepted, tested and offers error-free classification. Medical researchers call such a perfect method as ***Gold Standard***. For instance, the test IFN-γ is considered as a gold standard for detecting TB.

Using statistical tools, alternate biomarkers or alternate cutoff values (for the same marker) can be proposed by studying large data sets where classification is already done using a perfect method. Such alternate methods are important when the gold standard is difficult to implement or when search is required for new cutoff values.

It is also possible to combine several biomarkers into a statistical model to develop a new classification rule by using a portion of the available data (called *training data*). Such models are known as *predictive models.* The proposed model is tested for its correctness using *testing data*. This method is called ***supervised learning*** where the groups are well defined and known. When the groups are not pre-defined, clustering methods are used to identify groups and this is called ***unsupervised learning***. Several related concepts of learning models can be found in the field of data mining.

In general there are several methods of classification like Bayesian method, Logistic Regression, Disciriminant analysis, Support Vector Machine (SVM), Neural networks etc. Comparison of the efficiency of various methods is an interesting area of research. One needs such a classification rule which has the smallest percentage of misclassification.

Let the test result (discrete or continuous random variable) be denoted by X and x be a realization of X. Let *test positive* indicates a state of having disease (D) and test negative indicates a state of healthy (H). Define c as a threshold or cutoff value for classification. One rule is to cclassify a subject into D group if $x > c$ and H group otherwise. When the procedure is repeated on a number a cases and the test result is compared with the actual diagnosis, we get the following four counts.

a) True Positive (TP): Count of cases where both diagnosis and test are positive.
b) False Positive (FP) : Count of cases where the diagnosis is negative but test is positive
c) True Negative (TN) : Count of cases where both diagnosis and test are negative
d) False Negative (FN) : Count of cases where the diagnosis is positive but test is negative

These states are often shown as a matrix in which the entries are non-negative integers indicating the count of cases in the corresponding categories.

	Test result			
		Positive	Negative	Total
Diagnosis	Positive	TP	FN	P
	Negative	FP	TN	P_1
	Total	Q	Q_1	N

In an ideal situation one expects FP = FN = 0 but this never happens, unless the test is perfect in some sense. Some important indicators of the test performance are given below

Sensitivity (S_n)

It is the conditional probability of having a positive test among the patients who have a positive diagnosis (condition) and denoted by $S_n = P[X > c \mid D]$. This probability is estimated from sample data as

$$\hat{S}_n = \frac{\text{Number of True positives}}{\text{Number of True positives+Numberof False negatives}}$$

$$\hat{S}_n = \frac{TP}{TP+FN}$$

This is also known as True Positive Rate (TPR) or True Positive Fraction (TPF). For a given test, if the sensitivity is say 0.90 it means that in 90% of the cases where there is disease, the test shows positive. Hence we need a high sensitivity for a test.

Specificity (S_p)

It is the conditional probability of having a negative test among the patients who have a negative diagnosis (condition) and denoted by $S_p = P[X \leq c \mid D]$. This probability is estimated from sample data as

$$\hat{S}_p = \frac{\text{Number of True negatives}}{\text{Number of True negatives+Numberof False positives}}$$

$$\hat{S}_p = \frac{TN}{TN+FP}$$

This is also known as True Negative Rate (TNR) or True Negative Fraction (TNF). A specificity of say 0.80 means that in 80% of the cases where there is no disease the test also shows negative. Hence we need a high specificity for a test.

Both S_n and S_P values lie between 0 and 1 and correspond to one cutoff value. If the value of c increases, both S_n and S_P will change. A good diagnostic test is supposed to have high sensitivity with reasonably high specificity.

2 .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

2.1 ROC curve for rating data

When the test result X is a measurement, like the blood cholesterol level of a patient, and then would be a discrete random variable. Let $c_1, c_2,\ldots,c_k$ be k possible cutoff values with which the classification is made. At each c_i we get a pair (S_{ni} , ($1-S_{pi}$)) for i = 1,2,..,k. The plot of these k points produces a curve called the empirical ROC curve (Hanley and Neil (1982)), such a curve will not have a smooth form.

Let ROC (C_i, u) denotes the ROC value corresponding to u = $1-S_p$. Each cutoff Ci corresponds to a point ($1 - S_p$, S_n) on a ROC curve, i =1, 2,..,k. The jump at C_i becomes {ROC (C_i, u) – ROC (C_{i-1}, u)}. For a test with k (finite) number of possible cutoff values, we get a step function for the ROC.

In this paper we have focused on continuous random variables.

Hand and Krazanowski (2009) have summarized several important properties of ROC curves for continuous variables which are outlined below.

2.2 Properties of the ROC curve

1. Y = h(x) is the mathematical model of the ROC curve, where y denotes the true positive rate and x denotes the false positive rate. The curve is a monotonic increasing function in the positive quadrant, lying between y = 0 at x = 0 and y = 1 at x = 1
2. The ROC curve is unaltered if the classification scores undergo a strictly increasing transformation.
3. The slope of the ROC curve at threshold value c is given by

$$\frac{dy}{dx} = \frac{P(X > c \mid D)}{P(X > c \mid H)}$$

In the following section we present the method of constructing the ROC curve.

2.3 Constructing the ROC curve

Consider at test result which is a continuous random variable (X) and let the observed sample values from n-subjects be denoted by $x_1, x_2, \ldots, x_n$. Suppose there are k possible cutoff values for decision making, denoted by $c_1, c_2, \ldots, c_k$. For the i^{th} cutoff c_i we get a pair $\{S_{ni}, (1-S_{pi})\}$. The ROC curve is a plot of S_{ni} against $(1-S_{pi})$ for i = 1,2,….,k. The ROC curve lies in the unit square with diagonal line represented by the points for which $S_{ni} = (1-S_{pi})$.

The shape of the ROC curve is shown in Figure

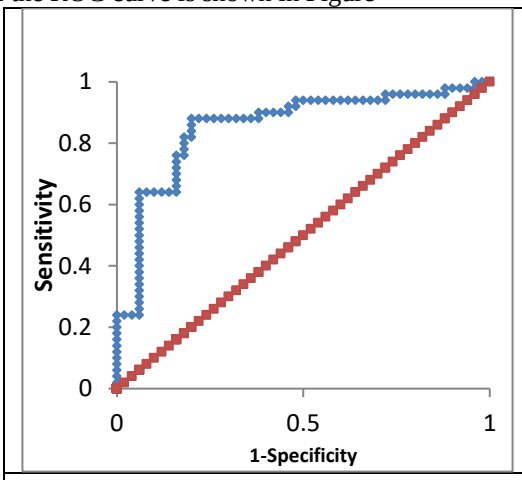

Figure 1(a) : ROC Curve for discrete data

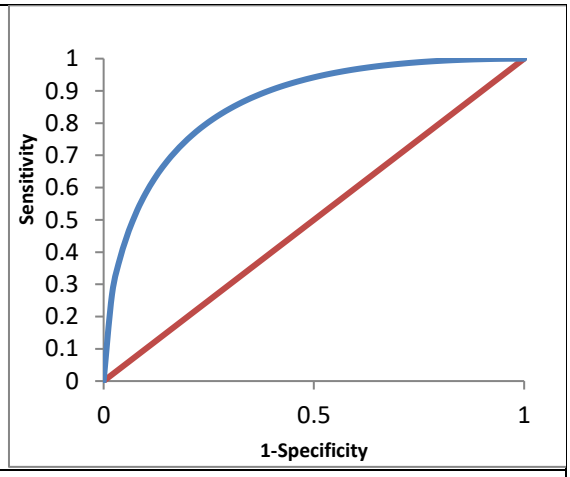

Figure 1(b) : ROC Curve for continuous data

The curve shown in Figure-1(a) is a step function with finite number of cutoff values. For continuous variables there will be infinite possible values with in an interval and it is difficult to specify what a cutoff means. In such cases each possible data values will be considered as a possible cutoff so that the ROC curve becomes nearly smooth as shown in Figure-1(b).

It is possible to develop theoretical models for ROC curves by using probability distribution of the test variable in the two groups. Farraggi and Reiser (2002), lasko et.al(2005), pepe et.al(2008) have discussed some important ROC models and their properties. In the following section a review of theoretical ROC models are discussed.

2.4 Estimation of ROC curves

Let X and Y be two continuous random variables representing the test score in D and H groups respectively. In more general terms, let F and G represent the distribution of score in the D and H groups respectively. The ROC Curve has the general form y(t) = h{x(t)} (Metz (1978)) where h(.) is a monotonic increasing function of x(t) where x(t) is the FPR at t. Since X and Y are continuous, every data value becomes a possible cutoff which is parameterized as t. Then x(t) = P (X > t | H) is the False Positive Rate (FPR) with reference t. Again y(t) = P(Y > t | D) is the True positive rate (TPR). Using the distribution F(.) and G(.) for the D and H groups respectively, we can write x(t) = 1-G(t) and

$y(t) = 1 - F\{G^{-1}(1-x(t)\}$ (1)

so that the ROC is a plot of (x(t), y(t)), $0 \le (x(t), y(t)) \le 1$. This is called parametric form of ROC curve. More details on the parametric form of ROC curves can be had from Krzanowski and Hand (2009). In case of ordinal data we get a jagged ROC Curve.

The estimation of the model depends on the type of the distribution used in the model (1). Since there are only two distributions involved we can consider the same family (like normal, exponential) with different parameters and such models offer interesting mathematical properties of the model. Examples are bi-normal, bi-exponential ROC models.

The estimation of the process of the model needs the use of inverse of the distribution function G (.). In the case of bi-normal model y (t) will have a simple linear structure as discussed in section. Since ROC curve only a graphical representation of a classifier we need a numerical summary of the ROC curve, which is called the Area Under the Curve (AUC).

3. Area under the curve (AUC)

The accuracy of a diagnostic test can be explained by using the AUC of an ROC curve. AUC describes the ability of the test to discriminate between D and H. The line (0,0) to (1,1) on the ROC Curve is the diagonal line above which the AUC is 0.5, which means that the test has only 50% chance of discriminating between D and H categories. The AUC lies between 0.5 and 1. Tests for which AUC > 0.5 are only considered and one expects that a good test has AUC close to 1. Test for which AUC < 0.5 need not be considered at all. Higher the AUC better will be the diagnostic test. The AUC has a nice interpretation as the average sensitivity at all possible specificities.

4. Summary

Studies designed to measure the performance of diagnostic tests are important for patient care and health care costs. ROC curves are a useful tool in the assessment of the performance of a diagnostic test over the range of possible values of a predictor variable. The area under an ROC curve provides a measure of discrimination and allows investigators to compare the performance of two or more diagnostic test.

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.