

Business Intelligence

- SEMINAR WINTER SEMESTER 2012/2013 -

Introduction to ROC Analysis

- SEMINAR PAPER -

Submitted by: Aleksandar Dimitrievski

Advisor: Prof. Dr. Dirk Neumann

Table of Contents

1. Introduction	3
2. Receiver operating characteristic (ROC)	3
2.1 Basic concepts	3
2.2 ROC curves and the area under the curves.....	6
3. Constructing a ROC curve using R	8
3.1 Sample dataset.....	8
3.2 Constructing ROC curves using R	9
4. Conclusion.....	10
5. References	10

1. Introduction

Making predictions has become an essential part of every business enterprise and all academic fields (Gönen 2007). One critical aspect of evaluating and comparing these prediction models, algorithms or technologies is the assessment of their accuracy. Receiver operating characteristic curves address that problem. They show how accurate a model, algorithm or a technology is, depending on its past results. Formally, a receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers, set of categories from the dataset, based on their performance (Fawcett 2006).

The development of the ROC curves dates back from the Second World War where it was firstly used to determine if a radar signals are actual enemy objects or noise. The damage that can occur if an enemy aircraft is mistaken for a friendly or noise by mistake can have catastrophic consequences, but at the same time aircraft could not be sent to intercept all of the objects that appear on the radar. Therefore, the ROC curve was devised as graphical means to explore the trade-offs between these competing losses at various decision thresholds when a particular quantitative variable, y , is used to guide the decision (Brown and Davis 2006). In more recent history ROC curves are extensively used in medicine. In the field of medical diagnosis, receiver operating characteristic (ROC) curves have become an important tool for evaluating how accurate a test is in determining if a patient has a certain disease. Furthermore, the receiver operating characteristic curves have become increasingly important and common method used in other fields such as finance, atmospheric science and machine learning (Gönen 2007).

This brings us to the main focus on this paper, which is explaining the main theory behind ROC curves the way ROC curves are generated and their possible use cases. Moreover we will use an example dataset to plot ROC curves using R. Finally we will examine the possibility of using this concept in economic theory and practice, concluding with an overview of the whole paper.

2. Receiver Operating Characteristic (ROC) curves, and area under the curve

ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. They are widely applicable, regardless of the source of predictions (Gönen 2007). In order to explain the concept of ROC curves and the usefulness behind them, firstly we need to focus on some basic concepts.

2.1 Basic Concepts

A great number of models and tests give us dichotomous results. The examples are numerous, a test for a certain disease, weather forecast, credit scoring, news and rumors, etc. What all of these

examples have in common is the uncertainty in the accuracy of the results. Even though the dichotomous test gives us only two results often there is some error margin included. Therefore the Receiver Operating Characteristic (ROC) curves were developed. ROC curves determine the goodness of the test. They are a graphical plot which illustrates the performance of the test as its threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate) at various threshold settings (Fawcett 2006).

In order to explain the Receiver Operating Characteristic (ROC) curves in more detail we consider an experiment that gives us the following results. Instances that resulted in positive outcome form our experiment **P** and instances that resulted in negative outcome **N**. Considering the conditions of our instances (either positive or negative) we can create a 2x2 contingency or confusion matrix (see table 1) to present all the possible outcomes.

	Condition positive	Condition Negative
Outcome Positive	True Positive (TP)	False Positive (FP)
Outcome Negative	False Negative (FN)	True Negative (TN)

Table 1. Confusion Matrix.

Given the possible outcomes from Table 1, we can define two essential statistical measures of performance that also play a main role in creating the ROC curves. These two statistical measures are sensitivity and specificity.

Sensitivity (also called true positive rate TPR) shows the test's ability to identify positive results, it measures the percentage of actual positives which our experiment did identify as positive. Specificity on the other hand (also called the true negative rate) shows the test's ability to identify negative results. For example the percentage of enemy planes identified refers to the sensitivity and the percentage of friendly aircrafts identified refers to the specificity of the test. Formally sensitivity and specificity can be calculated as follows.

$$\text{Sensitivity (TPR)} = \text{TP} / \text{TP} + \text{FN} \tag{1}$$

$$\text{Specificity} = 1 - \text{FPR} = \text{TN} / (\text{FP} + \text{TN}) \tag{2}$$

In effect, sensitivity and specificity represent two kinds of accuracy: the first for actually positive cases and the second for actually negative cases (Metz 1978). Sensitivity and specificity depend on two factors, how well the test separates the two groups and what threshold we choose (Spitalnic 2004). Furthermore they are closely related to the false positive and negative error occurring in our

experiment. Changing the threshold affects all of the values in our confusion matrix. If we take a look at distribution from the positive and negative results in Figure 1 we can easily show how the threshold value affects the results in our confusion matrix. Starting with the first threshold $t1$ we can clearly see that this threshold gives us no false negative occurrences therefore the True Positive Rate (Sensitivity) will give us the maximum value. Moving the threshold along to the left will increase the number of false negative occurrences and in the same time decrease the number of false positive occurrences. Coming to the far left side with the threshold $t4$ which will result in no false positive occurrences i.e. the False Positive Rate will be zero and we will achieve 100% specificity.

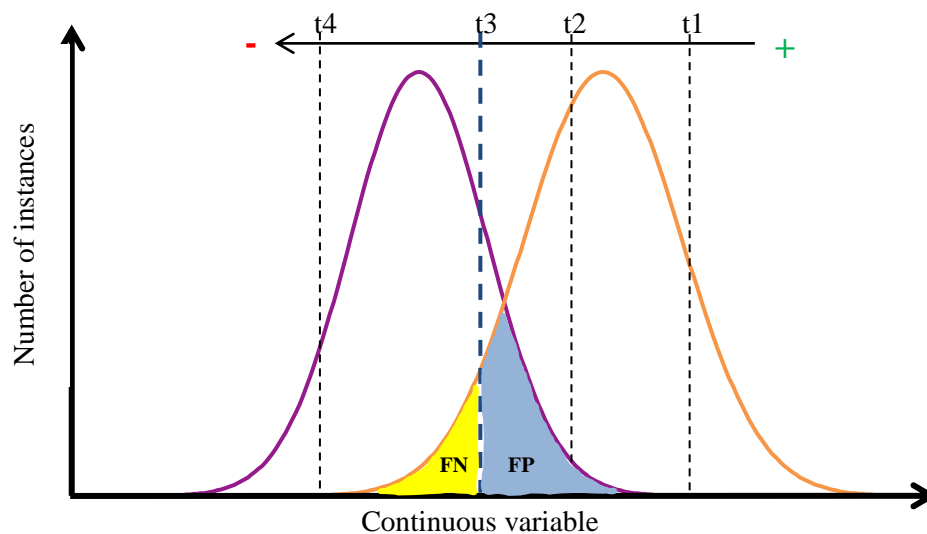


Figure 1. An illustration of the occurrence of positive (orange) and negative (purple) events and varying threshold.

If we change the threshold several times as described above we can obtain many pairs of sensitivity and specificity values. These pairs will show us how the test adjusts the identification of the true positive and true negative entries when we have a varying threshold. Furthermore if we focus our attention only to the positive values of our test we need to take a look at the true positive rate (sensitivity) and the false positive rate ($1 - \text{specificity}$). These parameters form paired values that adjust depending on the assigned threshold. The paired values that we get can easily be plotted as “x” and “y” coordinates on a graph. The axes on this graph range from zero to one because these are the limits of possible sensitivity and specificity values (Metz, Basic principles of ROC analysis 1978). The result from this coordinates is our Receiver Operating Characteristic (ROC) curve. Thus the ROC curve is simply a plot of the intuitive trade-off between sensitivity and specificity, with the horizontal axis flipped for historical reasons (Eng 2005).

2.2 ROC curves and the area under the curves

As we showed before the ROC curves are simply created by plotting the paired values of the true positive rate and the false positive rate that we get from considering many possible settings of the decision threshold. This gives us a curve that represents the goodness of the test (Figure 2). Historically the aim of the ROC curve was getting the positive values and therefore we take the false positive rate as the second variable.

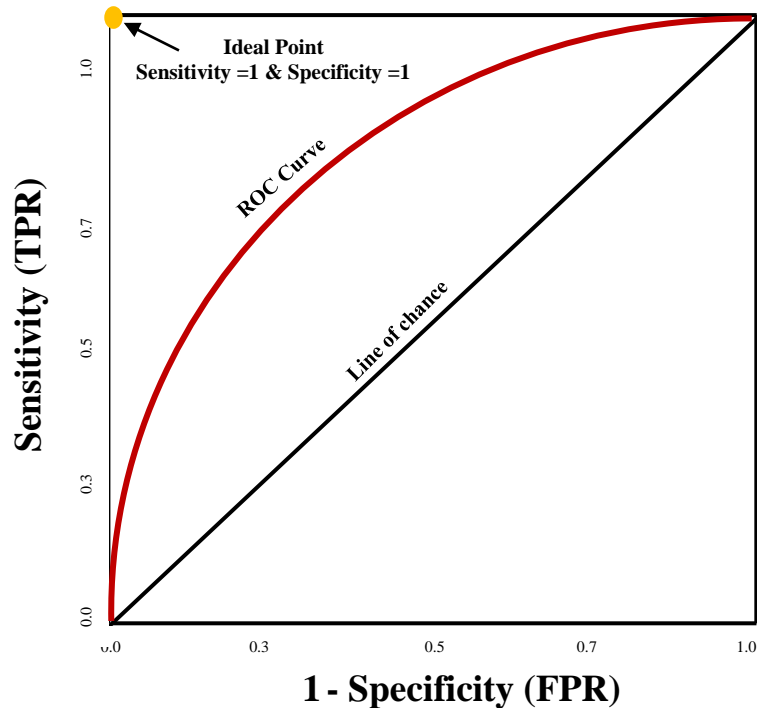


Figure 2. Sample Receiver Operating Characteristic (ROC) Curve.

If we look at Figure 2 we can see that the ROC curve starts at the (0, 0) coordinates. This means that with this threshold the test gives us value for specificity, i.e. the true negative rate, equal to 100% and a true positive rate, i.e. sensitivity equal to 0. Therefore the test has a very high threshold that results in only negative results. Moving along to the other end of the curve we get completely opposite results. Here the threshold value is too low, and this results in only positive results (TPR=1, FPR=0). The diagonal line connecting these two points represents test or model that yields randomly guessing results. This means that regardless of the thresholds chosen the specificity and sensitivity values will be 0.5, meaning that there is 50% chance for the model to reveal the true positive or true negative occurrences, therefore this line is called the line of chance. Thoughtful consideration of Figure 2 reveals that ROC curves higher or lower than the line of chance correspond to tests that show better predictive results. An ideal decision variable would have a ROC curve that passed through Sensitivity=1 and Specificity=1, which would correspond to a point in the top left corner of the ROC axis (Brown and Davis 2006). Finally, one can show on theoretical grounds that if the decision maker

uses available information in a proper way, the slope of the ROC curve must steadily decrease (i.e., it must become less steep) as one moves up and to the right on the curve (Metz, Basic principles of ROC analysis 1978).

Another commonly used indicator of the accuracy of the test is the area under the ROC curve (AUC). The area measures discrimination, that is, the ability of the test to correctly classify those with and without the disease (Tape 2004). The area under the curve has an important statistical property: it shows the probability that the observer will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative (Hanely and McNeil 1982). An area of 1 that shows a test that is 100% accurate because both the sensitivity and specificity are 1.0 so there are no false positives and no false negatives (Systat Software Inc 2006). As we reduce the area under the curve the accuracy of the test gets worse. If we look at Figure 3 we can see different ROC curves and their performance according to the AUC.

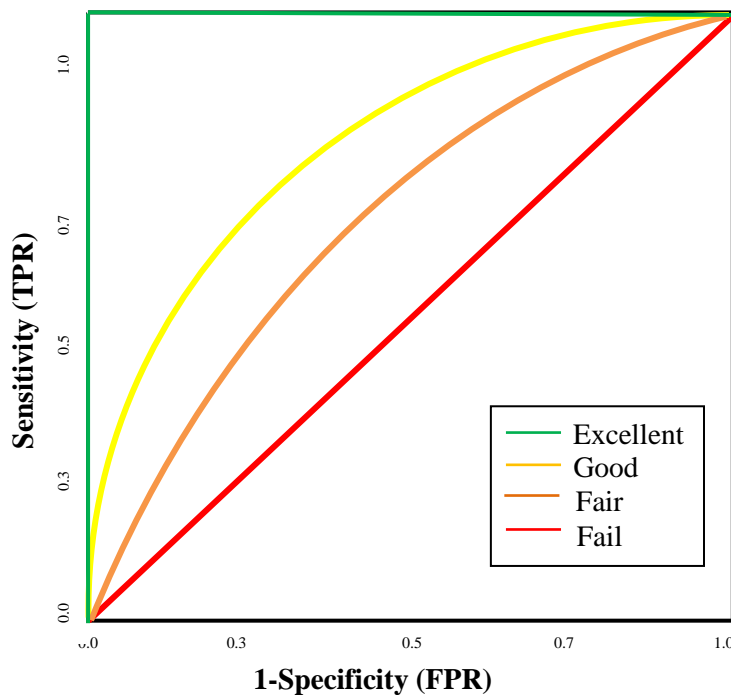


Figure 3. Different ROC Curves and their prediction accuracy.

The area under the curve also allows us to compare different tests. Since the AUC shows the average sensitivity over the entire range of possible specificities, or the average specificity over the entire range of possible sensitivities (Eng 2005), we can compare different tests not depending on the population size, test group size, etc. Formally the area under the ROC curve can be calculated with the following formula:

$$AUC = \int_{-\infty}^{\infty} TPR(T)FPR'(T)dT \quad (3)$$

The ROC curve provides a very simple graphical view of the trade-space that is possible for a given decision variable, and the general discriminating power of the variable for detecting positive/negative events independent of the event rates (Swets, Dawes and Monahan 2000). In the next section we will use an example dataset in order to construct a ROC curve using R.

3. Constructing a ROC curve using R

A typical ROC study begins by asking an observer to interpret a number of cases, some of which are positive and some negative, where the true classification unknown to observer (Eng 2005). In our example we take bank loan requests as cases that need to be interpreted. When bank decides whether or not it will approve a loan a lot of factors are taken into consideration. Ultimately the bank needs to decide if the person will be able to return the money in the determined timeframe. If we take the banks decision point of view we can easily realize that they are faced with yes/no decision that can yield four results. They can approve the loan and get the loan back in the determined timeframe together with some interest rate, or lose the loan money if the person is not able to pay. On the other hand the bank can decline the loan and save them from future loss if the person is not able to return the money, or decline the loan and lose the interest rate that they would have earned if the person was able to return the loan in the determined timeframe. In Table 2 we can see the results that can occur when a bank decides to approve or decline a bank loan.

	Able to return the money	Not able to return the money
Approve loan	Profit True Positive (TP)	Loss False Positive (FP)
Decline Loan	Loss False Negative (FN)	Profit True Negative (TN)

Table 2. Possible outcomes from the bank's decision about the loan.

3.1 Sample Dataset

Our dataset consist of 1000 entries from people who have requested a loan form a bank. The dataset originates from an open source web page (Frekonimics.com). In these 1000 entries we have 382 declined and 618 approved applications. Furthermore there are two variables for each entry, firstly we have a coefficient that corresponds to the size of the person's income, and secondly we have a coefficient that corresponds to the size of the loan that the person requested. In this hypothetical example the bank will consider only these two variables in order to decide whether or not to approve the requested loan. The bank decides upon giving a loan by considering the share of the loan form the person's income. In addition our example dataset includes information whether or not a person will be able to pay back the bank loan as an assumed dichotomous variable. The information about the ability

to pay back the loan is needed for calculating the goodness of our test, i.e. calculating the sensitivity and specificity values.

3.2 Constructing ROC curves using R

In this section we will provide some information about creating ROC curves in R and show the results from our test. For more detailed information upon the coding used to construct the curves in R please refer to the copy of the R code provided in addition of this paper. For this example we assume that the bank calculates the probability of getting the loan returned as a ratio between the loan and income index. Therefore we firstly calculate the third variable in our dataset, i.e. the ratio. This gives us the probability if the person will get the loan. Since the probability is calculated as the ratio between loan and income index, the higher the ratio the less chances the person has in receiving a loan. After calculating the probability we can easily continue creating the ROC curve by using the pROC package in R. In Figure 4 we can see the result from our test in R. As shown in Figure 4 the AUC is 0,847 meaning that the test yields satisfying results. Furthermore we can conclude that if the bank uses the ratio between loan size and income as their only decision factors of whether or not to grant a loan, they will be giving the loans to the persons who are able to return the money 84,7% of the time.

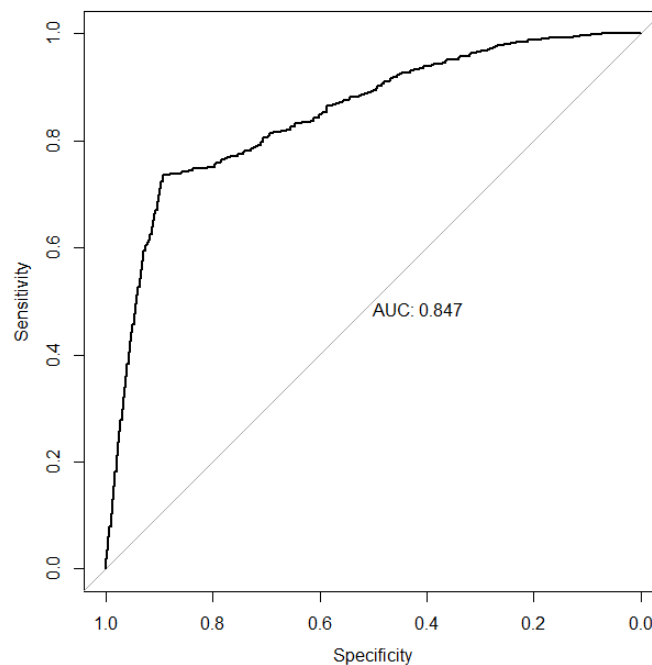


Figure 4. Constructed ROC curve and AUC from the test in R.

4. Conclusion

This paper should serve as a good starting point for further research around the topic of accuracy of dichotomous tests, and the use of statistical computing software for evaluating the goodness of the tests. With the increasing number of predictions generated on a daily basis, ROC curves can provide an easy and simple way of assessing their results. Given the simplicity of the ROC curves they can provide a decent test of the accuracy of the prediction. They are a useful tool for picturing and estimating tests. Furthermore, their simplicity is their biggest benefit, since results are very intuitive and straight forward. This opens the doors for ROC Curves and their usage in medicine learning, in the field of medical diagnosis, finance, atmospheric science, etc.

With our example we have shown how ROC curves are created and how can we use the results that our test yielded in order to evaluate the goodness of the test. Moreover we have shown how the additional add-ins for R simplify the process of constructing the ROC curves.

5. References

- Brown, Christopher D, and Herbert T Davis. "Receiver operating characteristics curves and related decision measures: A tutorial." *Chemometrics and Intelligent Laboratory Systems*, 2006: 24 – 38.
- Eng, John. "Receiver Operating Characteristic Analysis: A Primer." *Academic Radiology*, 2005: 909-916.
- Fawcett, Tom. "An introduction to ROC analysis." *Pattern Recognition Letters*, 2006: 861-874.
- Gönen, Mithat. *Analyzing Receiver Operating Characteristic Curves with SAS*. Heidelberg: SAS Institute, 2007.
- Hanely, James A, and Barbara J McNeil. "The meaning and use of the area under a receiver operating characteristic curve (ROC)." *Diagnostic Radiology*, 1982: 29-36.
- Investopedia. *Investopedia*. 2013. <http://www.investopedia.com/university/stocks/stocks4.asp> (accessed 1 18, 2014).
- Metz, Charles E. "Basic principles of ROC analysis." *Seminars in Nuclear Medicine*, 1978: 283–298.
- Metz, Charles E. "Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems." *Journal of the American College of Radiology*, 2006: 413–422.
- Spitalnic, S. "Test Properties 1: Sensitivity, Specificity and Predictive Values." *Hospital Physician*, 2004: 27-31.
- Swets, John A, Robyn Dawes, and John Monahan. "Psychological Science Can Improve Diagnostic Decisions." *Psychological Science in the Public Interest*, 2000: 1 - 26.
- Systat Software Inc. "Systat Software." *ROC Curves Analysis*. 2006. http://www.sigmaplot.com/products/sigmaplot/ROC_Curves_Analysis.pdf (accessed December 27, 2013).
- Tape, Thomas G. 2004. <http://gim.unmc.edu/dxtests/roc3.htm> (accessed 1 18, 2014).

- . "Interpreting Diagnostic Tests." *The area under ROC Curve*. 2004. <http://gim.unmc.edu/dxtests/roc3.htm> (accessed January 6, 2014).
- The R Foundation for Statistical Computing. *R Project*. 2013. <http://cran.r-project.org/> (accessed December 27, 2013).