Comments On Validation Methods For Two Group Classification Models Widely Accepted In Credit Scoring Or Response Analysis

Timothy H. Lee,
Equifax Decision Systems,
Equifax Credit Information Services, Inc.,
Internal Mail Code 42S,
P.O. Box 740006, Atlanta, Ga 30374-0006

The two group classification methods are popular approaches for the separation of one group from the other. For these purposes either parametric or non-parametric classification approaches are used. In many cases a scoring algorithm is derived and the score distribution serves as a basis of the decision making. Generally, validation of a model is to assure the model has reasonable separation power when it is applied to a different data set not used for the development of the model, i.e., holdout data set. In the credit scoring case, Regulation B of Equal Credit Opportunity Act requires the scoring algorithm be revalidated frequently enough to ensure that it continues to meet statistical standards. In addition, in case of comparison of more than one model, it is necessary to quantify model performance in some way. Two sample Kolmogorov-Smirnov test statistic, Kullback-Leibler Number, and Mahalanobis Distance, etc. are popular ways of quantifying model performance. In this study, such popular methods are discussed along with the advantages and disadvantages of each method using a simulated data set and a suggestion of an improved, intuitive, and simple quantifying method for model performance is made

KEY WORDS: Kullback-Leibler number, Two sample Kolmogorov-Smirnov Test, Logistic Regression, Discriminant Function

1. INTRODUCTION.

Two group classification analysis is a very popular approach in industries such as credit granting or target marketing. For instance, in the credit industry, credit grantors want to predict the creditworthiness of applicants. By the two group classification approach, more creditworthy applicants are separated from less creditworthy applicants. In the process of discrimination, a scoring algorithm is derived based on the known data and the algorithm is applied to applicants to score them. Without a doubt, a good scoring algorithm has better separation power than others. Of course, the ultimate performance of the model should be measured by the profitability. The profitability, however, is hard to be measured objectively. Besides, there are various uncontrollable econo-socio, consumer behavior related, or business related factors that affect profitability. In this paper, we would like to focus our attention on the separation power and separation pattern of a classification of models - especially on the quantification of model performance.

The measurement of model performance is important to see if the algorithms are discriminating adequately or to determine if other models do a better job of rank ordering. If a model is to screen a potentially better creditworthy applicant for credit extension, Regulation B of the Equal Credit Opportunity Act (ECOA) requires frequent validation of the model performance. It, however, does not point to any specific statistical method. The regulation simply states, 'The scoring system must be periodically revalidated by the use of appropriate statistical principles and methodology ...' Besides the regulatory reasons, there are many other reasons to quantify the performance of models.

In most applications, the models are for two group classification. The model provides a basis to assign an object to either of the two populations, p_1 or p_2. In the process of classification, multivariate observations $\mathbf{x}$ for each object were transformed to univariate observation y such that the y's derived from populations p_1 and p_2 were separated as much as possible. In the industry, each element in $\mathbf{x}$ is

demographic, socio-econo, or credit bureau factor pertaining to each individual and computed y is called score. The score per each individual is a base for the classification.

Next, we would like to review the statistical validation approaches widely used in the industry and discuss related issues. Finally an alternative measure will be proposed.

2. MEASUREMENT OF SEPARATION.

It has been an issue among analysts using two group classification methods including logistic regression, discriminant function or regression with a binary dependent variable what statistical method will be used to measure the performance of the model. Since most applications are two group classification, the model performance is measured by the accuracy of separation of a group from the other. If a non-parametric classification method is used, the classification error rate would be considered as a measure. In many cases, a parametric or semi-parametric approach is used for classification and score for each individual is computed as a basis for class assignment. In such a case, model performance should not be measured simply by the classification error. The separation pattern of a model should be taken into consideration because it may affect stability of decision.

The score distribution generated for each group differs from each other if the scoring algorithm separates. The degree of difference between the score distributions does not necessarily measure the performance of a model. We will visit this issue in the discussion session again.

Two most commonly used statistical methods for the validation of a model or for the comparison of model performance are i) Two Sample Kolmogorov Smirnov (K-S hereafter) test or ii) computation of the Kullback Leibler Entropy (Divergence) on the score distributions generated by the scoring algorithm. The scoring algorithm, sometimes called a scoring model, is an equation or a rule for assignment derived using any two group classification method such as Discriminant Function, Logistic regression, or other parametric or non-parametric classification technique, etc.

2.1 Two Sample K-S Test

The test was proposed by Smirnov, N. V. (1939) for the test of the hypothesis that any two samples are from the same population. It is to test H_0: $F(x) = G(x)$ for all x against the general alternative H_1 when the two samples, $X_1,...,X_m$ and $Y_1,...,Y_n$ are independent random samples from continuous distributions with c.d.f.'s $F(x)$ and $G(y)$. The test rejects H_0 if and only if the observed value of

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| \text{ for all } x,$$

where $F_m(x)$ and $G_n(x)$ denote the empirical distributions corresponding to $F(x)$ and $G(x)$, is greater than any threshold value determined at a proper significance level.

The threshold value is to be determined from the table or, when m and n are greater than 80, approximated by

$$z_a ¥ [(m+n)/(m ¥ n)]^{0.5},$$

where z_a is determined by proper significance level.

It is conventional, somehow, in the industry, that the $D_{m,n}$ is used as a measure of model performance. In other words, the test statistic value for the testing of equality of the two distributions is used for the measure of separation power of a model. We will revisit this issue in the following sections.

2.2 Divergence

As Soofi (1994) pointed out in his recent paper about capturing the intangible concept of information, many statisticians are familiar with the theory of discrimination information. Moreover, quantifying information in some statistical problems has been the highlight among statisticians in the industry. Since most often the purpose of the model is to separate one group from the other, the interest of the analysts is in the entropy of discrimination information. Shannon (1948) developed information entropy for quantifying the expected uncertainty associated with an outcome from a sample from a population that has distribution f. His formula for the entropy was

$$H(x) = -E[f(x) \log f(x)].$$

Kullback and Leibler (1951) generalized above entropy into relative entropy,

$$H(f,g) = \int \log [f(x)/g(x)]\, dF(x), \quad \ldots\ldots (1)$$

where f and g are probability distributions for p_1 and p_2, respectively.

Expression (1) is known as Kullback Leibler entropy, directed divergence, or the relative information of class 1 with respect to class 2. The entropy is not a symmetric function. Jeffrey (1946) considered a symmetric version of this function as a measure of divergence between two distributions with densities f and g,

$$D = H(f,g) + H(g,f) \quad \ldots\ldots\ldots\ldots (2)$$

The quantity is called Kullback Leibler Cross Entropy, Information Number, or Divergence. Right hand side of (2) can be rewritten as

$$\int \log [f(x)/g(x)]\, dF(x) - \int \log [f(x)/g(x)]\, dG(x)$$

$$= E\,[L(x)|p_1] - E\,[L(x)|p_2], \quad \ldots\ldots\ldots\ldots (3)$$

where $L(x) = \log [f(x)/g(x)]$.

The divergence is expressed as the difference in means of the two L(x)'s on p_1 and p_2, respectively.

Therrien (1989) showed that the divergence is equal to Mahalanobis distance between the two means when the data has Gaussian distribution and the two covariance matrices are equal. If the measure of divergence is applied to score distribution to see how well the two score distributions differentiate each other, the divergence can be written, under the normality and equal variance and covariance assumption,

$$D = (m_1 - m_2)^2/[(s_1^2 + s_2^2)/2], \quad \ldots\ldots\ldots\ldots (4)$$

where m_1, s_1^2, m_2, and s_2^2 are means and variances of the score distributions for p_1 and p_2, respectively.

In the above we reviewed two commonly used approaches for model validation in the industry. In the following sections we will discuss advantages and disadvantages of the approaches taking examples and a potentially superior alternative approach will be proposed.

3. EXAMPLE.

In the past major model developers in the credit industry have debated regarding selection of the validation methods and they tried to show that their approach was superior to their competitors'. Strange enough, each of major developers employed one approach.

In this section we will assume several cases of separation pattern and compare the changes of the two approaches, K-S test statistic vs. divergence. We will assume some different patterns of separation depending on the skewness conditions of the two scoring distributions for each group as in the following:

1) Two score distribution curves are normally distributed. (See Figure -1.)

2) Two score distribution curves are inwardly skewed. (See Figure - 2.)

3) Two score distribution curves are outwardly skewed. (See Figure - 3.)

4) One score distribution is nested by the other. (See Figure - 4.)

Figure -1. Two Normal Distributions

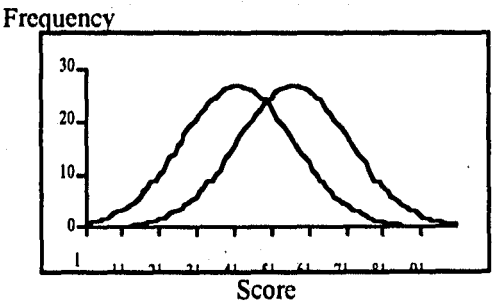


Figure - 2. Two Inwardly Skewed Distributions

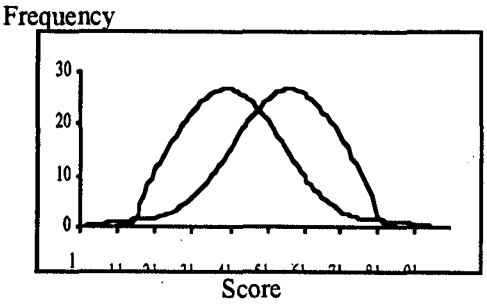


Figure - 3. Two Outwardly Skewed Distributions

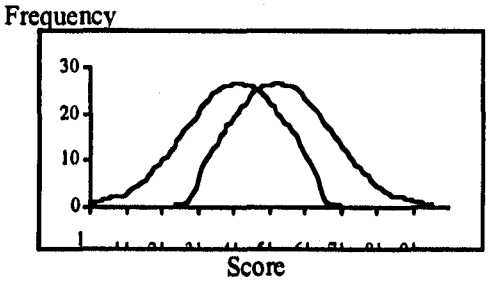
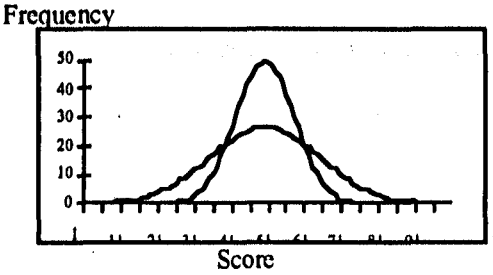


Figure - 4. One Distribution is nested by the other



In case 1) both divergence and K-S test statistic value are used correctly. In case 2) divergence will be measured too conservatively, while in case 3), divergence will give too optimistic measure. In case 4) K-S test statistic value will be little too optimistic but divergence will measure more accurately. Even though most cases are close to the case1), it is a natural desire for the analysts to use a method that measures the separation power of a model properly taking into consideration the separation pattern. In the following section a different idea from the previously mentioned methods will be presented.

4. COEFFICIENT OF SEPARATION.

As mentioned in the previous section, divergence seems to be affected by the skewness of the score distribution, while the Two sample K-S test is not proper as a measure of separation in the case when one distribution is included in the other, even though the test statistic can be a good measure for differentiation of one distribution from the other. To alleviate such problems of the two common measures the following approach is proposed:

1) Create a cumulative empirical score distribution for each group (e.g., creditworthy versus non-creditworthy).

2) Per each observed score point (or interval) read the two cumulative empirical probability as x and y coordinate.

3) Plot the coordinates on the unit square. Then, the trace of the points form a curve reflecting the pattern of separation as in the Figure - 5.

4) Find the area between the curve created in 3) and the 45 degree line (no separation line). If the curve is partially above and below the 45 degree line, find absolute difference of the area below the 45 degree line and that above the line. Such a case is observed when one distribution curve is included inside of the other. (See Figure - 5.)

5) The absolute difference of the areas computed in 4) is divided by the area of the triangle under the 45 degree line. This value will be used as a measure of separation.

Above Procedure is similar to plotting Lorentz curve or ROC (Receiver Operating

Characteristic) curve except finding the difference of the two areas. Simple calculus method such as Trapezoidal approximation of curve would be good enough to estimate the areas. This method (call it Coefficient of separation or C-S) is compared with the two commonly used methods, Two sample K-S test and divergence in Table - 1.

Figure -5. Separation Curve on Unit Square

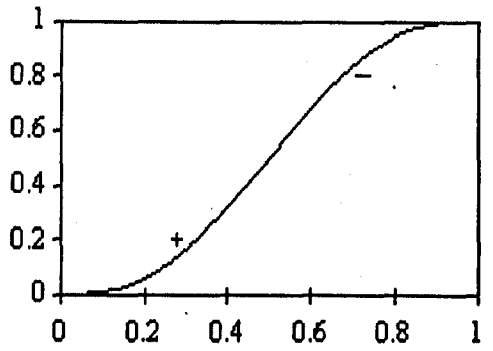


Table - 1. Comparison of model performance measures

Cases	Measures		
	K-S	Divergence	C-S
1. No skewness	38.16	0.99	51.84
2. Skewed inwardly	37.36	0.75	48.32
3. Skewed outwardly	37.65	1.30	55.83
4. One includes the other	14.72	0.00	0.00

5. REMARKS.

In this article we considered three different methods for model validation. It is often observed in the credit industry that selection of a validation method depends on the modeling method. For example, if the modeling approach is parametric or semi-parametric, Two sample K-S test is very often used. If the model is derived by iterative search method maximizing Information number, the measure for model performance is usually divergence. In most cases each of the three method works properly. Extreme cases such as mentioned ahead are very rare, even unrealistic. Such cases, however, can be artificially created by some transformation such as Logistic function. Two sample K-S test statistic value is not affected by any one to one transformation. The divergence, however, is affected when the skewness is changed. The coefficient of separation, compared to the other two methods, seems to be reasonable in most cases as a measure for model performance because it reflects separation pattern of a model.

ACKNOWLEDGMENT

The author thanks Dr. Donald Searls of University of Northern Colorado for his helpful comments and Dr. Ming Zhang of Equifax decision Systems for his careful review of the paper. The author also thanks to Dr. John Pohlmann of Southern Illinois University for helpful suggestions.

REFERENCES

Kullback, S., and Leibler R. A. (1951), "On Information and Sufficiency", *The Annals of Mathematical Statistics*, 22, 986-1005

Smirnov, N. V. (1939), "On the estimation of the discrepancy between empirical curves of distribution for two independent samples.". (Russian) Bull. Moscow Univ. 2, 3-16

Soofi, Ehsan S. (1994), "Capturing the Intangible Concept of Information", *Journal of the American Statistical Association*, 89, 1243-1254

Therrien, Charles W. (1989), *Decision, Estimation, and Classification*, New York: John Wiley