Lab Report on ROC Curve and Linear Regression (March 2020)

Jagdish Khadka(073/BEX/313), Krishna Bahadur Ojha(073/BEX/317), Rajad Shakya(073/BEX/332)
Student Member, IEEE

***Abstract*—This report describes the lab session of Data mining on the study of Gaussian distribution, Receiver Operating Characteristics (ROC) Curve, Area Under the Curve (AUC) and Linear as well as Polynomial Regression using python programming language.**

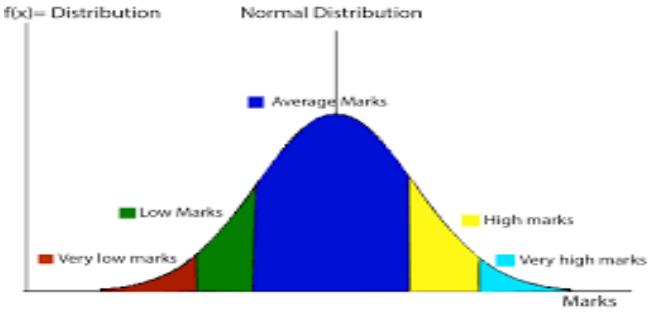
***Index Terms*— Gaussian distribution, Receiver Operating Characteristics (ROC) Curve, Area Under the Curve (AUC), Regression**

I. Introduction

DATA mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from data. In other words, it is the exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns. This lab session dealt with few of the parts of data mining which is described in the later sections.

II. Gaussian Distribution

A Gaussian or normal distribution is a bell-shaped frequency distribution curve. Most of the data values in a normal distribution tend to cluster around the mean. The further a data point is from the mean, the less likely it is to occur. There are many things, such as intelligence, height, and blood pressure that naturally follow a normal distribution.



The empirical rule tells what percentage of the data falls within a certain number of standard deviations from the mean:

• 68% of the data falls within one standard deviation of the mean.

• 95% of the data falls within two standard deviations of the mean.

• 99.7% of the data falls within three standard deviations of the mean.

The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean; the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the mean; the normal distribution will be flatter and wider.

Its properties are as follows:

1. The mean, mode and median are all equal.
2. The curve is symmetric at the center (i.e. around the mean, μ).
3. Exactly half of the values are to the left of center and exactly half the values are to the right.
4. The total area under the curve is 1.
5. Each normal distribution has a different mean and standard deviation that make it look a little different from the rest, yet they all have the same bell shape.
6. The standard deviation is the distance from the center to the saddle point (the place where the curve changes from an "upside-down-bowl" shape to a "right-side-up-bowl" shape.

III. ROC Curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

1. True Positive Rate
2. False Positive Rate

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{1}$$

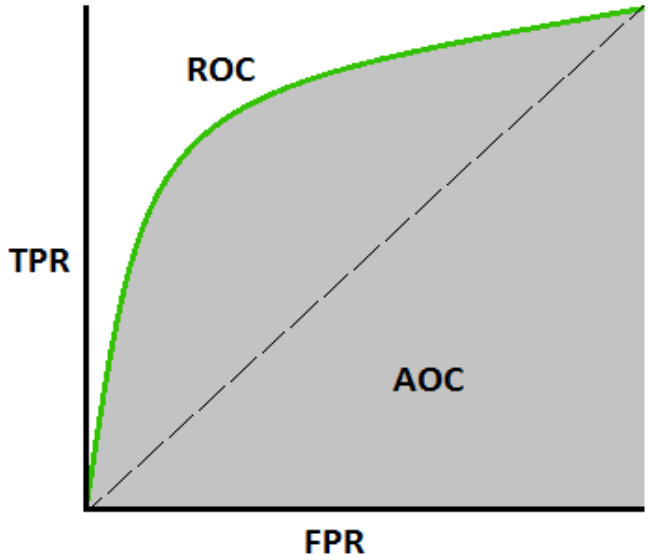
False Positive Rate (FPR) is defined as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2}$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

Receiver Operating Characteristics (ROC) graphs are a useful technique for organizing classifiers and visualizing their performance. ROC graphs are commonly used in medical decision making, and in recent years have been increasingly adopted in the machine learning and data mining research communities.



They are able to provide a richer measure of classification performance than accuracy or error rate can, and they have advantages over other evaluation measures such as precision-recall graphs and lift curves. However, as with any evaluation metric, using them wisely requires knowing their characteristics and limitations.

IV. Linear Regression

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent and independent variable. In other words, it is the process of using the relationship between variables to find the best fit line or the regression equation that can be used to make predictions.

There are multiple benefits of using regression analysis. They are as follows:

a. It indicates the significant relationships between dependent variable and independent variable.
b. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called Ordinary Least Squares. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a quadratic equation:

$$y = a + b * x^2 \quad (2)$$

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

V. Methodology

A. Libraries Used

1) Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. It is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

2) Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

3) Pandas

'Pandas' is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

4) Sklearn

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

B. Functions Used

1) Pandas DataFrame.dropna

Sometimes csv file has null values, which are later displayed as NaN in Data Frame. Pandas *dropna()* method allows the user to analyze and drop Rows/Columns with Null values in different ways. In order to drop rows with at least one Null value, data frame is read and all rows with any Null values are dropped. The size of old and new data frames is compared to see how many rows had at least 1 Null value.

In case it is not used, due to the presence of Null value, the

program encounters an error during runtime as while it is expecting an identifiable arithmetic value, it gets NaN. In the Cars dataset, there are some missing values which makes it necessary to use this function.

2) sklearn.metrics.auc

This function computes the Area under the Curve (AUC) using the trapezoidal rule.

C. Calculation of AUC

The AUC was calculated using the Trapezoidal method by the use of aforementioned function. The trapezoidal rule is a numerical integration method to be used to approximate the integral or the area under a curve. The integration of [a, b] from a functional form is divided into n equal pieces, called a trapezoid. Each subinterval is approximated by the integrand of a constant value. Two common methods are:

1) Linear Trapezoidal Method

The linear trapezoidal method uses linear interpolation between data points to calculate the AUC. This method is required by the OGD and FDA, and is the standard for bioequivalence trials. For a given time interval (t1 – t2), the AUC can be calculated as follows:

$$AUC = \frac{1}{2}(C_1 + C_2)(t_2 - t_1) \tag{35}$$

In essence the first two terms calculate the average concentration over the time interval. The last piece (t1 – t2) is the duration of time. So the linear method takes the average concentration (using linear methods) and applies it to the entire time interval. When you sum all of the intervals together, you will arrive at the total exposure from the first time point to the last. If you then divide the total AUC by the total time elapsed, you will arrive at the "average" concentration of drug in the body over the total time interval.

2) Logarithmic Trapezoidal Method

The logarithmic trapezoidal method uses logarithmic interpolation between data points to calculate the AUC. This method is more accurate when concentrations are decreasing because drug elimination is exponential (which makes it linear on a logarithmic scale). For a given time interval (t1 – t2), the AUC can be calculated as follows:

$$AUC = \frac{C_1 - C_2}{ln(C_1) - ln(C_2)}(t_2 - t_1) \tag{21}$$

This method assumes that C1 > C2. The fraction represents the logarithmic average of the two concentrations. Just as with the linear method, the average concentration is multiplied by the time interval.

D. Dataset Used

The data used is technical spec of cars. The dataset called Cars data or Auto-Mpg data is available at the UCI Machine Learning Repository. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.

The details of this dataset are as follows:

1. Number of Instances: 398
2. Number of Attributes: 9 including the class attribute
3. Attribute Information:
   a. mpg: continuous
   b. cylinders: multi-valued discrete
   c. displacement: continuous
   d. horsepower: continuous
   e. weight: continuous
   f. acceleration: continuous
   g. model year: multi-valued discrete
   h. origin: multi-valued discrete
   i. car name: string (unique for each instance)

The dataset was split into training and testing dataset with training dataset containing 196 instances. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model.

E. Training using Regression

Linear Regression is a machine learning algorithm based on supervised learning that performs a regression task. While training the model, we are given x (input training data) and y (labels to data). When training the model, it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best intercept and coefficient of x, say q1 and q2. Once we find these values, we get the best fit line. By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is the minimum. So, it is very important to update these q1 and q2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value(y). Cost function (J) of Linear Regression is the Mean Squared Error (MSE) between predicted y value (pred) and true y value (y).

Linear regression requires the relation between the dependent variable and the independent variable to be linear. But since the distribution of data is more complex in most of the cases, polynomial regression comes in use. While training the model we are given: x (input training data) and y (labels to data). When training the model, it fits the best quadratic equation (we used both quadratic as well as cubic during our lab sessions) to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best intercept and coefficient of x, say q1 and q2. The model gets the best regression fit line by finding the best coefficients say a, b and c. Once we found these values, we got the best quadratic equation. By achieving the best -fit regression equation, the model aims to predict y value such that the error difference between predicted value and the true value is the minimum. So, it is very important to update these coefficient values, to reach the best value that minimizes the error between predicted y value and true y value.

F. Calculation of MSE

Mean Square Error is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points. In other words, the MSE of an estimator measures the average of the

squares of the errors i.e. the average squared difference between the estimated values and the actual value. In mathematical notation,

$$C = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (21)$$

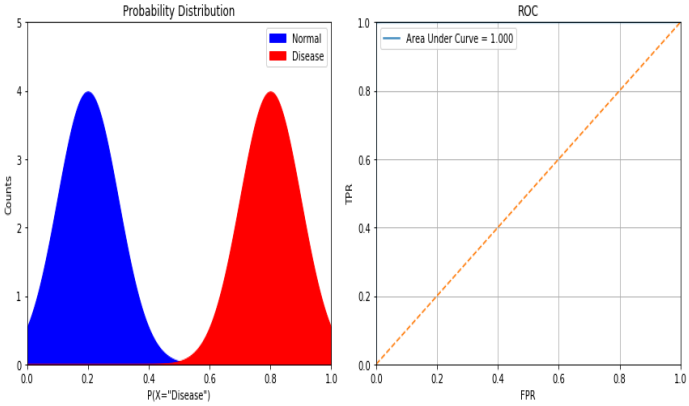
Where, y_i is the target value for input-output pair $(\vec{x_i}, yi)$ and $\hat{y}_i$ is the computed output of the network on the input $\vec{x_i}$.

VI. Result

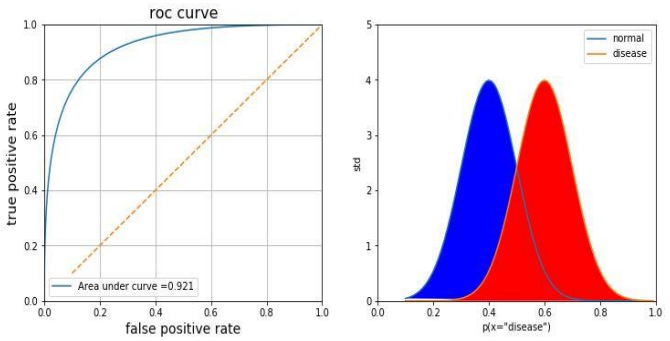
A. ROC Curve

Two normal distribution functions with same standard deviation and means or expected value summing up to 1 indicating probabilities were used for this purpose. The expected values were changed for normal and diseased Gaussian Curves in order, 0.2/0.8, 0.4/0.6, 0.5/0.5 and 0.6/0.4, and the Gaussian and ROC curves so formed were studied.

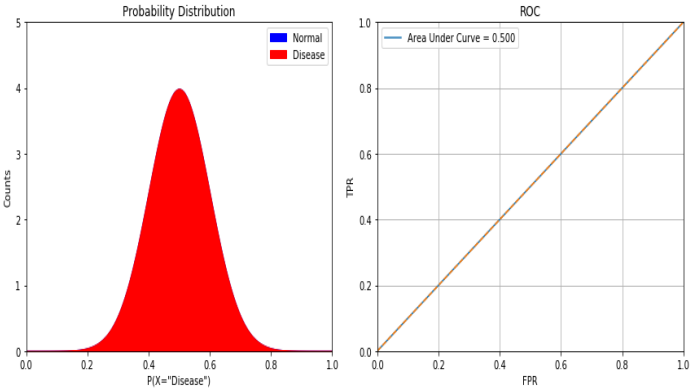
As we know, ROC is a curve of probability. So plotting the distributions of those probabilities:



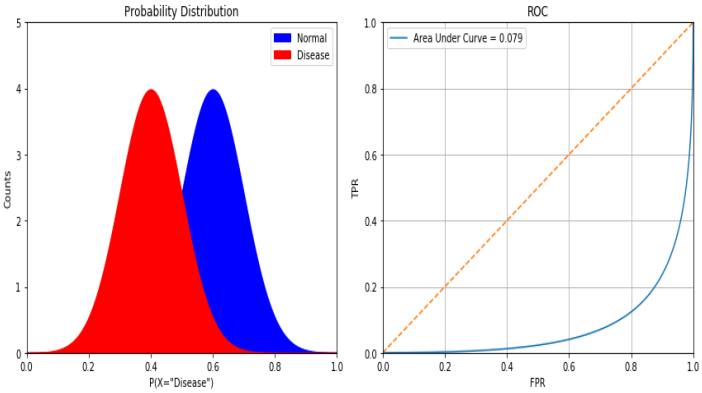
This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.



When two distributions overlap, we introduce type 1 and type 2 error. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.921, as in this case, it means there is 92.1% chance that model will be able to distinguish between positive class and negative class.



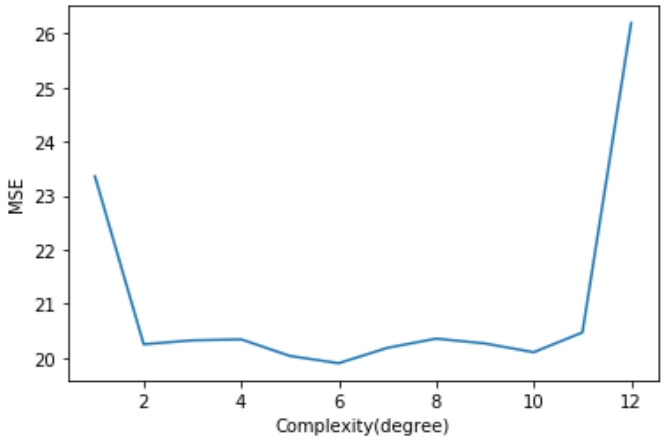
This is the worst situation. When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class.



When AUC is approximately 0, model is actually reciprocating the classes. It means, model is predicting negative class as a positive class and vice versa.

B. MSE vs. Complexity

The MSE versus Complexity in degree was observed for degrees in the range of [1, 13] is plotted as follows:



The error encountered is high for both very low complexity and high complexity, so it is imperative to find the global minimum which is neither too high nor too low. It can be seen from the graph that the minimum value of error is at degree 6 and was calculated to be 19.9.

VII. Discussion

The two Gaussian distribution curves are plotted by using the gaussian function and filled with the two different curves to visualize clearly. Here the true positive rate gets almost constant while increasing the false positive rate after the 0.5 and the area under the curve was 0.92 which was very useful to choose cut off for test.

VIII. Conclusion

In this way, the lab of Data Mining on the on the study of Gaussian distribution, Receiver Operating Characteristics (ROC) Curve, Area Under the Curve (AUC) and Linear as well as Polynomial Regression was successfully done.

Appendix

https://github.com/krishnabojha/Data_Mining/tree/master/5.%20Gaussian

References

[1] Ethem Alpaydin, *Introduction to Machine Learning*.: MIT Press, 2010.

[2] Gonen Mithat, Analyzing Receiver Operating Characteristic Curves Using SAS, SAS Press, 2007