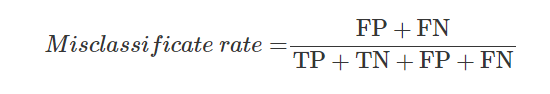
Topic 2 Assign Part1 response

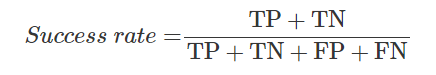
Gayathri Rajkumar

Grand Canyon University: DSC 540

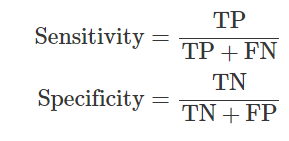
09/16/2020

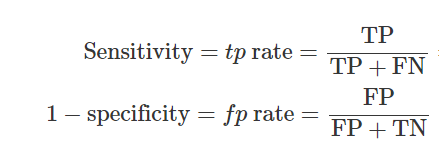
Decisions made on the basis of classifications that are based on misclassification error rate result in poor performance when data is unbalanced or highly skewed (Gopal, 2019). The important role of the confusion matrix is to evaluate the performance of the outcome of a classification model. These metrics help to measure the accuracy of the classification problems. A classification model will accept as input the attributes and then it will find the pattern within the data, and then it will classify the data. In this process, the model is getting trained using training data and then predict using test data. For example, in the binary sentiment classification analysis model, the prediction can be either positive or negative sentiment. The confusion matrix has two rows and 2 two columns in the matrix. This matrix has both actual and predicted values, such as predicated positive, predicted negative, actual positive, and actual negative. Once the model makes the predication using test data, the results of the prediction get loaded into the confusion matrix to understand the performance of the model. For example, the cell in the top left (first row and first column) represents Positive sentiments and true positives which means correctly predicted as positive. The bottom right of the cell in the confusion matrix represents the true negative which means predicated as negative correctly. The top right box is false negative which means the actual value is positive but classified as negative. The bottom left is false positive which means the sentiments are negative, but the model is classified as positive. This is the simplest way to identify the misclassification error in pattern recognition. Using the confusion matrix, both accuracy and misclassification rate can be calculated using the below formula.





The confusion matrix also helps to find other metrics such as sensitivity and specificity. A sensitivity (true positive rate) of 1.0 implies that the model predicts all positive observations in a correct manner; simply put, the approach fails to make any false-negative errors (Gopal, 2019).  A specificity of 1.0 or false positive rate of 0 indicates that the model predicts all negative observations accurately; that is, the model does not make any false positive predictions (Gopal, 2019). When the model’s true positive rate is high and the false positive rate is low, that model is considered as good, otherwise, the model is not accurate and not acceptable.

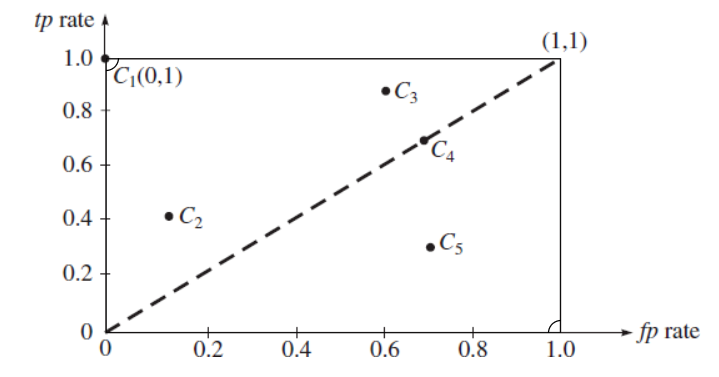




Although the classifier has high accuracy, it has low sensitivity, and therefore its ability to accurately recognize positive tuples is poor. Because of low fp rate, it can accurately predict negative class (Gopal, 2019).

The Receiver Operator Characteristic (ROC) is a two-dimensional graph that evaluates the rate of detecting true positive against detecting false positive. False-negative has a high impact than false positive. For example, false negative can wrongly predict a patient does not have disease. So the accuracy of the classification model is important. The ROC graph and confusion matrix play important role in identifying the misclassification error in pattern recognition. In the ROC graph, the true positive rate is on the y-axis which is also known as sensitivity, and the false-positive lays on the x-axis which is known as specificity.

The below chart is retrieved from Applied Machine learning.



In the above chart, (0,1) is true-positive and (1,1) is false-positive. The lower left point (0,0) is not classified as positive or negative. The (1,1) is positive, but not classified as true or false positive. The C1, C2, C3, C4, and C5 are classifier instances. The performance of C1 is perfect as it is classified as true positive. The classifier C3 has high true positive and C2 has low true positive. Both classifiers are in above the triangle. The area above the triangle is liberal. Even though C4 is below the line, there is a slight chance to have true positive as it is close to the border. The classifier C5 is in the lower right triangle, which is close to false-positive and it has poor performance. When a classifier algorithm is applied to a test set, it yields a single confusion matrix, which in turn corresponds to one ROC point (Gopal, 2019). Roc Curve are created using the values using confusion matrix to identify the classification issue.

References

Gopal, M. (2019). *Applied machine learning*. New York: McGraw-Hill Education.