30 March 2018
05:50

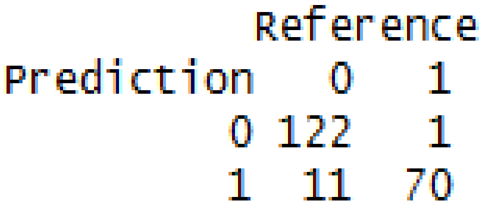
Evaluation Metrics for Classification Models

- Computing just the accuracy score for a classification model gives an incomplete picture of your model's performance. The following evaluation metrics should help gain perspective of the practical usability of your classifier model.

- Earlier you saw how to build a logistic regression model to classify malignant tissues from benign, based on the original Breast cancer dataset.

- The computed the accuracy from the above model turned out to be 94%, which sounds pretty good. But, it doesn't reveal much information about how well the model actually did in predicting the 1's and 0's independently.

The Confusion Matrix:



```
          Reference
Prediction   0   1
         0 122   1
         1  11  70
```

- Look at the 1 in top-right of the table. This means the model predicted 1 instance as benign which was actually 'malignant' (positive).

- This is a classic case of '**False Negative' or Type II error**. You want to avoid this at all costs, because, it says the patient is healthy when he is actually carrying malignant cells

- Also, the model predicted 11 instances as 'Malignant' when the patient was actually 'Benign'. This is called '**False Positive' or Type I error**. This condition should also be avoided but in this case is not as dangerous as Type II error.

Sensitivity:

Sensitivity is the percentage of actual 1's that were correctly predicted. It shows what percentage of 1's were covered by the model.
TPR = TP/(TP + FN)
The total number of 1's is 71 out of which 70 was correctly predicted. So, sensitivity is 70/71 = 98.59%

Similarly, you can call it **Recall** as well.

Specificity:

Specificity is the proportion of actual 0's that were correctly predicted.
TNR = TN/(TN + FP)
So in this case, it is 122 / (122+11) = 91.73%.

Detection rate:

It is the proportion of the whole sample where the events were detected correctly.
So, it is 70 / 204 = 34.31%

Precision:

The approach here is to find what percentage of the model's positive (1's) predictions are accurate.
So, it is 70/81 = 0.8642

F1 Score:

A **good model** should have a **good precision** as well as a **high recall**. So ideally, I want to have a measure that combines both these aspects in one single metric – **the F1 Score**.

F1 Score = (2 * Precision * Recall) / (Precision + Recall)

= (2*0.8642*0.9859)/(0.8642+0.9859)
= 0.9210

Cohen's Kappa:

Kappa is similar to Accuracy score, but it takes into account the accuracy that would have happened anyway through random predictions.

Kappa = (Observed Accuracy - Expected Accuracy) / (1 - Expected Accuracy)

ROC Curve:

Choosing the best model is sort of a balance between predicting the one's accurately or the zeroes accurately. In other words sensitivity and specificity. But it would be great to have something that captures both these aspects in one single metric.

This is nicely captured by the '**Receiver Operating Characteristics**' curve, also called as the **ROC curve**. In fact, the area under the ROC curve can be used as an evaluation metric to compare the efficacy of the models.

So, if we trace the curve from bottom left, the value of probability cutoff decreases from 1 towards 0. If you have a good model, more of the real events should be predicted as events, resulting in high sensitivity and low FPR. In that case, the curve will rise steeply covering a large area before reaching the top-right.

Therefore, the larger the area under the ROC curve, the better is your model.