621 Assignment2 - Classification Metrics

Raghu

Oct 3, 2018

Contents

```
## Loading required package: knitr

## Loading required package: ggplot2

## Loading required package: caret

## Loading required package: lattice

## Loading required package: pROC

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Question 1

Download the classification output data set

Answer: Summary of the data set.

pregnant	glucose	diastolic	skinfold	insulin	bmi	pedigree	age	class	scored.class	scored.probability
7	124	70	33	215	25.5	0.161	37	0	0	0.3284523
2	122	76	27	200	35.9	0.483	26	0	0	0.2731904
3	107	62	13	48	22.9	0.678	23	1	0	0.1096604
1	91	64	24	0	29.2	0.192	21	0	0	0.0559984
4	83	86	19	0	29.3	0.317	34	0	0	0.1004907
1	100	74	12	46	19.5	0.149	28	0	0	0.0551546

Question 2

The data set has three key columns we will use:

. class: the actual class for the observation

. scored.class: the predicted class for the observation (based on a threshold of 0.5)

. scored.probability: the predicted probability of success for the observation

Use the table() function to get the raw confusion matrix for this scored dataset. Make sure you understand the output. In particular, do the rows represent the actual or predicted class? The columns?

```
##
##       0   1
##   0 119  30
##   1   5  27
```

Answer: Rows represents the predicted class.
Columns represents the actual class.

Question 3 to 8

Write a function that takes the data set as a dataframe, with actual and predicted classifications identified, and returns the following of the predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$classificationErrorRate = \frac{FP + FN}{TP + FP + TN + FN}$$

Verify that you get an accuracy and an error rate that sums to one.

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1Score = \frac{2 * Precision * Sensitity}{Precision + Sensitity}$$

Sensitivity (also called the true positive rate) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

The positive and negative predictive values (PPV and NPV respectively) are the proportions of positive and negative results in statistics and diagnostic tests that are true positive and true negative results, respectively.

The PPV and NPV describe the performance of a diagnostic test or other statistical measure. A high result can be interpreted as indicating the accuracy of such a statistic. The PPV and NPV are not intrinsic to the test; they depend also on the prevalence.The PPV can be derived using Bayes' theorem.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.806 which means our model is approx. 80% accurate.

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.High precision relates to the low false positive rate. We have got 0.798 precision which is pretty good.

F1 score - F1 Score is the weighted average of Precision and Sensitivity. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Sensitivity. In our case, F1 score is 0.871.

accuracy	error_rate	precision	sensitivity	specificity	f1
0.8066298	0.1933702	0.7986577	0.9596774	0.4736842	0.8717949

Answer:

Sum of accuracy and error rate

```
## [1] 1
```

Question 9

Before we move on, let's consider a question that was asked: What are the bounds on the F1 score? Show that the F1 score will always be between 0 and 1. (Hint: if $0 < a < 1$ and $0 < b < 1$ then $ab < a$)

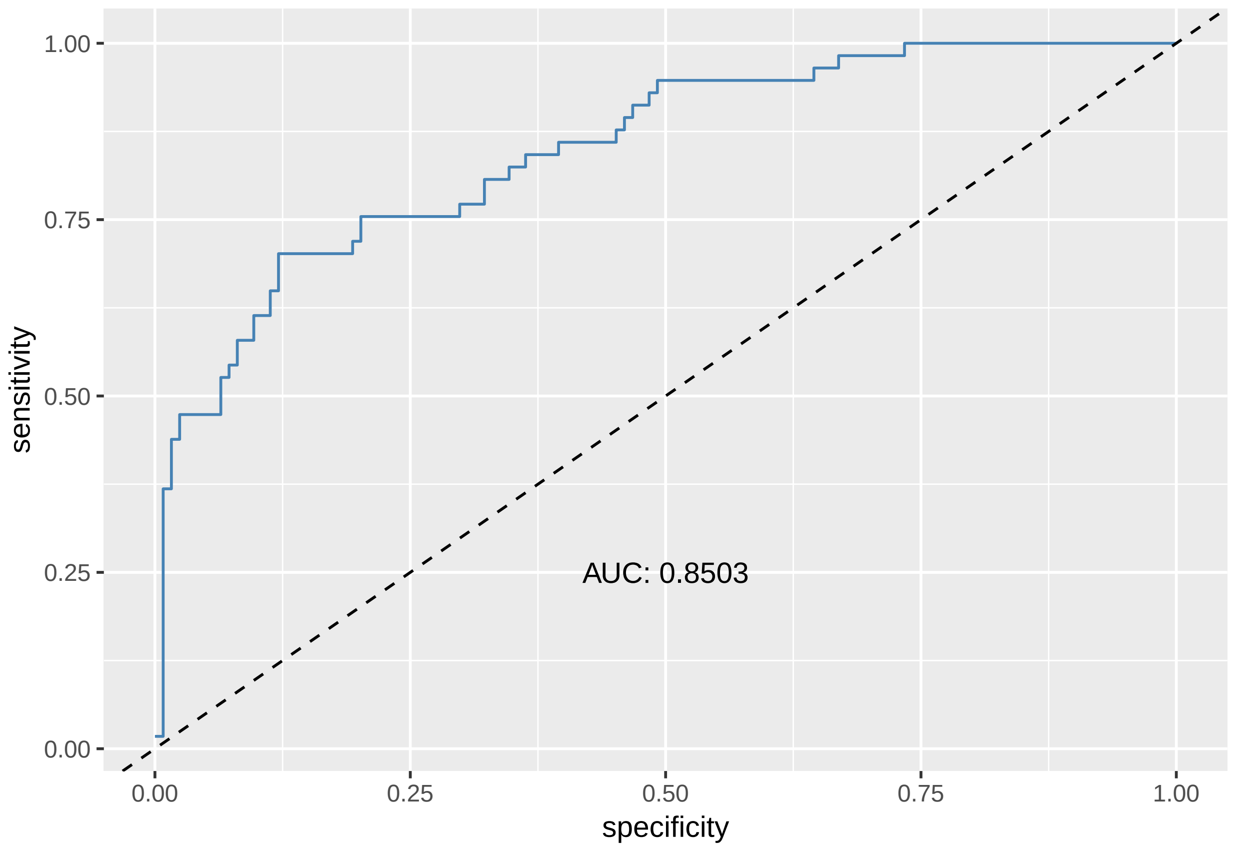
Answer:

F1 is calculated using the precision and sensitivity scores. Since each of those are bounded by 0 and 1, we can be confident that the values when substitued in the formula will be between 0 and 1.

Question 10 and 11.

Write a function that generates an ROC curve from a data set with a true classification column (class in our example) and a probability column (scored.probability in our example). Your function should return a list that includes the plot of the ROC curve and a vector that contains the calculated area under the curve (AUC). Note that I recommend using a sequence of thresholds ranging from 0 to 1 at 0.01 intervals.

Use your created R functions and the provided classification output data set to produce all of the classification metrics discussed above.



ROC curve is used to show in a graphical way the connection/trade-off between sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

As the area under an ROC curve is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests.

The closer an ROC curve is to the upper left corner, the more efficient is the test.

Question 12.

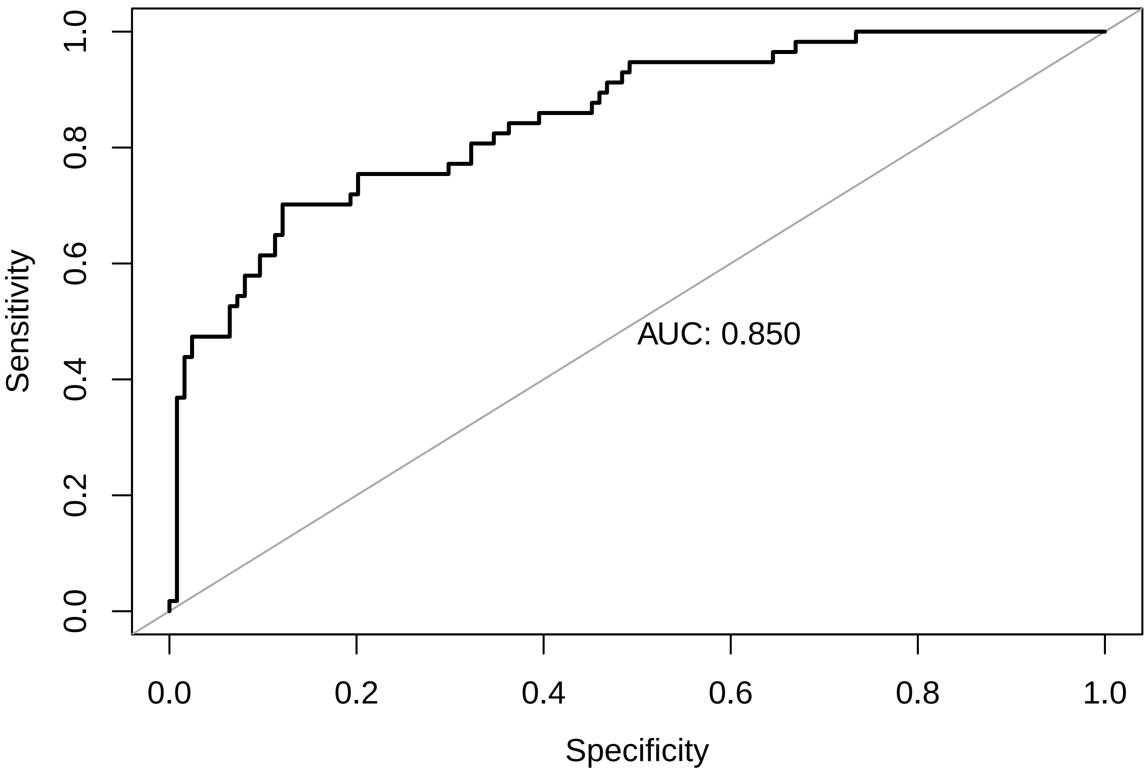
Investigate the caret package. In particular, consider the functions confusionMatrix, sensitivity, and specificity. Apply the functions to the data set. How do the results compare with your own functions?

	accuracy	error_rate	precision	sensitivity	specificity	f1
Manual	0.8066298	0.1933702	0.7986577	0.9596774	0.4736842	0.8717949
CaretPackage	0.8066298	0.1933702	0.7986577	0.9596774	0.4736842	0.8717949

In comparison of manual calculation and Caret package, the results are same.

Question 13.

Investigate the pROC package. Use it to generate an ROC curve for the data set. How do the results compare with your own functions?



ROC curve of the pROC package appears to be the same with that of the calculation created manually.