Evaluation

Prof. Eric A. Suess

February 17, 2021

Introduction

We have primarily talked about **Classification methods**, such as, kNN, Naive Bayes, C5.0, RIPPER, CART, Logistic Regression, etc.

In the Classification setting we have used **Accuracy/Success Rate** to Evaluate the "usefulness" of an algorithm.

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

So we have looked at the **Confussion Matrix**.

acc <- mean(pred == testy)

Introduction

We have started to look at **Prediction methods**, such as, Linear Regression, Multiple Linear Regression, etc.

So we looked at "Accuracy" as the **correlation** between the test values of the **response** and the **predicted/fitted** values from the model.

When using Prediction methods a quantitative response is predicted.

Question:

But Logistic Regression is used for Classification, right?

Answer:

Yes, but it uses the predicted probabilties.

In R we can classify using the **ifelse()** function to convert the probabilities into 0 and 1.

> *ifelse(prob $<$ 0.5, 0, 1)*

Beyond Accuracy

There are a number of values that can be calculated to evaluate accuracy using Classification algorithms.

Beyond Accuracy

- **Kappa** - adjusts accuracy by accounting for the possibility of a correct prediction by chance alone. So should be a bit smaller than what we have discussed as Accuracy.

Beyond Accuracy

- **Sensitivity**

$$Sensitivity = \frac{TP}{TP+FN} \approx P(+|D)$$

- **Specificity**

$$Specificity = \frac{TN}{TN+FP} \approx P(-|D^c)$$

Beyond Accuracy

- **Precision**

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

- **Recall**

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

Beyond Accuracy

- **F-measure** or F1 or F-score

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

The F-measure assumes equal weight for the Precision and Recall. This may not always be the case.

Visualizing Performance Tradeoffs - ROC

Visualizations can be very helpful for understanding how the performance of learning algorithms differ.

Useful for comparing two or more learners side-by-side.

The Receiver Operating characteristic (ROC) is commonly used.

To use the ROC we need:

1. the class values/labels
2. the predicted probabilities of the **positive class**

ROC - Sensitivity/Specificity plot

See page 312/332 for an example.

The ROC plots the **Sensitivity** versus **1 - Specificity**.

For the MS Statistics students this is:

- **True Positive Rate** versus **False Positive Rate**

or

- **Power** versus α.

ROC - Sensitivity/Specificity plot

No predictive value, 45 degree line

Perfect predictive value, up and across. 100% true positives with no errors.

ROC - AUC

The **Area Under the Curve** (AUC) is commonly used to compare Classifiers.

Holdout Method

- Training
- Validation
- Testing

Repeated Holdout

Cross-Validation

k-fold cross validation

10-fold cross validation

Train on 9 of the folds and test on the last. Average the accuracy measure.

Bootstrap sampling

Random sample with replacement. Train on the sample and test on the remaining examples.

$$error = 0.632 \times error_{test} + 0.368 \times error_{train}$$