Logistic Regression Validation

Jeffrey Woo

MSDS, University of Virginia

1. Motivation
2. Confusion Matrix
3. ROC Curve

Titanic Example

Going back to our previous example, we fit a logistic regression model to estimate the probability of survival on the Titanic, based on a number of predictors. How would you evaluate the predictive ability of our logistic regression model with our data?

Evaluating Classification Models

- With no access to more observations, we will have to split our data into two: training data and test data.
- Use the training data to fit a logistic regression model.
- Then use the model to estimate the probability of the observations in the test data of being in each class.
- Use a **decision rule / threshold** to classify the observations in the test data (e.g. if estimated probability > 0.5, classify as "Yes").
- Compute the number of: true positives, false positives, true negatives, and false negatives.

Titanic Example

Evaluate predictive ability of logistic regression of survival on fare paid and gender.

- I randomly split the data into the training and test data with equal number of observations.
- Fit logistic regression model with Fare and Gender as predictors using the training data.
- Use logistic regression model to estimate the probabilities for each observation in test data.
- Classify observation in test data as "Yes" if estimated probability > 0.5. Otherwise, classify as "No".

Titanic Example: Test Data

Survived	Sex	Age	Fare	preds	preds > 0.5
0	male	1.0	39.6875	0.2071538	FALSE
1	female	1.0	11.1333	0.6571846	TRUE
1	female	1.0	15.7417	0.6707167	TRUE
1	male	1.0	37.0042	0.2014132	FALSE
1	female	2.0	26.0000	0.6998214	TRUE
1	female	3.0	41.5792	0.7410628	TRUE
1	male	3.0	31.3875	0.1897844	FALSE
1	male	3.0	15.9000	0.1603996	FALSE

Titanic Example

```
> table(test$Survived, preds>0.5)

    FALSE TRUE
  0   180   30
  1    44  100
```

A **confusion matrix** can be used to summarize the number of true/false positives and true/false negatives.

Choice of Threshold

Choice of threshold should be based on what kind of error we are more concerned with, and chosen before the fitting the model. Discuss the following:

- If we are more concerned with a false positive (classifying someone who did not survive as someone survived), how should the threshold be changed?
- If we are more concerned with a false negative (classifying someone survived as someone who did not survive), how should the threshold be changed?

Confusion Matrix

Table: Classification Table of Model

	Model classifies as "No"	Model classifies as "yes"
True "no"	True negative (TN)	False positive (FP)
True "yes"	False negative (FN)	True positive (TP)

- Overall error rate:
- False Positive Rate / Type I error rate:
- False Negative Rate / Type II error rate:
- Sensitivity:
- Specificity:

Titanic Example

With threshold of 0.5:

	FALSE	TRUE
0	180	30
1	44	100

- Overall error rate:
- False Positive Rate / Type I error rate:
- False Negative Rate / Type II error rate:
- Sensitivity:
- Specificity:

Titanic Example

With threshold of 0.7:

	FALSE	TRUE
0	201	9
1	98	46

- Overall error rate:
- False Positive Rate / Type I error rate:
- False Negative Rate / Type II error rate:
- Sensitivity:
- Specificity:

The ROC Curve

Another way to present how well a model does in classifying observations is using a Receiver Operating Characteristic (ROC) curve. The name is derived from its initial use during World War II when analyzing radar signals. Users of radar wanted to distinguish signals due to enemy aircraft from signals due to noise such as a flock of birds.

- An ROC curve is a two-dimensional plot, with the sensitivity on the y-axis and $1 -$ specificity (false positive rate, or $\frac{FP}{TN+FP}$) on the x-axis.
- The ROC curve plots the sensitivity and $1-$ specificity for **every possible value** of the decision rule cutoff (i.e., between 0 and 1).

The ROC Curve

- A model that classifies at random (i.e. without using information from the predictors) will have an ROC curve that **lies on the diagonal**.
- A model that classifies all observations correctly will have a sensitivity and specificity of 1, so it will belong at the (0,1) position on the plot. The further the curve is from the diagonal and closer to (0,1), the better the model is in classifying observations correctly.
- A model that classifies all observations incorrectly will have a sensitivity and specificity of 0, so it will belong at the (1,0) position on the plot. The further the curve is from the diagonal and closer to (1,0), the worse the model does in classifying observations.

Titanic Example

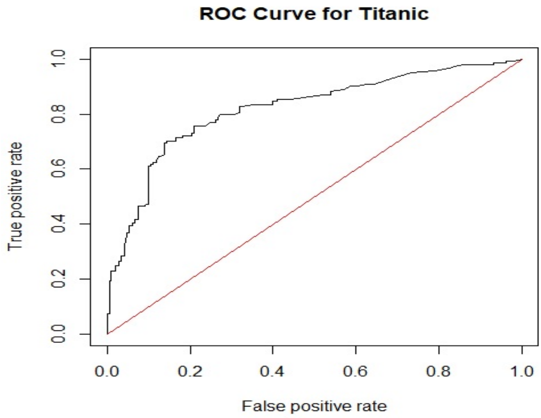

Figure: ROC Curve for Titanic Data

The AUC

Another way to measure the classification ability of a model is using the Area Under the ROC Curve (AUC).

- For a model that randomly guesses, the AUC will be 0.5.
- An AUC closer to 1 indicates the model does better than random guessing in classifying observations. An AUC of 1 indicates a model that classifies all observations correctly.
- An AUC closer to 0 indicates a model that does worse than random guessing.

The AUC in our example is around 0.8184.

Cautions

- Accuracy can be a misleading measure especially when you have **unbalanced sample sizes** of the two classes. You can have high accuracy, yet either the false positive or the false negative to be really high. In this situation, it is more informative to look at the false positive and false negative rates.
- The ROC curve shows the true positive and false positive rates as the threshhold is varied. It does not immediately inform you of the true positive and false positive rates for your specific threshold.
- The AUC, just like the ROC, is a summary of the predictive performance for all possible values of the threshold. It does not inform you of the accuracy for your specific threshold.