Classification Metrics

-Introduction to Data Science-

ISL: Chapters 2.2.3, 4.4.3

Lecturer: Darren Homrighausen, PhD

Evaluating Classifications

Misclassification rate

The loss function for classification is the 0-1 loss:

$$\ell(g(X), Y) = \mathbf{1}(Y \neq g(X)) \;\Rightarrow\; R(g) = \mathbb{P}(g(X) \neq Y)$$

Suppose we have training data $\mathcal{D}_{\text{train}}$ with $|\mathcal{D}_{\text{train}}| = n$,

We can define the training error (with respect to 0-1 loss) as

$$\hat{R}_{\text{train}}(g) = \frac{1}{n} \sum_{(X,Y) \in \mathcal{D}_{\text{train}}} \mathbf{1}(Y \neq g(X)) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(Y_i \neq g(X_i))$$

Likewise, with test data $\mathcal{D}_{\text{test}}$ with $|\mathcal{D}_{\text{test}}| = n_{\text{test}}$, we can define the test error (with respect to 0-1 loss) as

$$\hat{R}_{\text{test}}(g) = \frac{1}{n_{\text{test}}} \sum_{(X,Y) \in \mathcal{D}_{\text{test}}} \mathbf{1}(Y \neq g(X)) \to R(g) \quad \text{as } n_{\text{test}} \to \infty$$

An example

Suppose we are interested in predicting whether or not the economy will be in a recession

We have quarterly measurements of

- State level economic growth

  (Larger number is better)
- Federal level variables such as GDP, interest rates, employment, S&P 500, ...

Here, we will code the supervisor as

$$Y = \begin{cases} 1 & \text{if recession} \\ 0 & \text{if growth} \end{cases}$$

Confusion matrix

We can report our results in a matrix:

		Truth		
		Recession	No Recession	Totals
Our Preds.	Recession	TP	FP	$P^* = TP + FP$
	No Recession	FN	TN	$N^* = FN + TN$
	Totals	$P = TP + FN$	$N = FP + TN$	n_{total}

The total number of each combination is recorded in the table

The overall misclassification rate is

$$1 - \frac{\text{TP} + \text{TN}}{n_{\text{total}}} = \frac{\text{FP} + \text{FN}}{n_{\text{total}}}$$

Sensitivity and specificity

Sensitivity: The fraction of true positives (TP) out of the total number of actual positives (P)

(Notationally: TP/P)

Specificity: The fraction of true negatives (TN) out of the total number of actual negatives (N)

(Notationally: TN/N)

We can think of this in terms of hypothesis testing

$$H_0 : \text{no recession}$$
$$H_A : \text{recession}$$

Sensitivity: $\mathbb{P}(\text{reject } H_0 | H_0 \text{ is false})$ or 1 - $\mathbb{P}$(Type II error)

(This is the same as power)

Specificity: $\mathbb{P}(\text{accept } H_0 | H_0 \text{ is true})$ or 1 - $\mathbb{P}$(Type I error)

Precision and recall

Other commonly used criteria are precision and recall

Precision: This is the fraction of true positives (TP) out of total number of predicted positives (P^*)

(Notationally: TP/P^*)

Recall: This is the fraction of true positives (TP) out of the total number of actual positives (P)

(Notationally: TP/P. This is the same as sensitivity and power)

There is a combination of these two known as F1 score:

$$\mathrm{F1} = (2) \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

(This is the harmonic mean of precision and recall and $0 \leq \mathrm{F1} \leq 1$)

A larger F1 score indicates a better procedure

Receiver operating characteristic

Many classifiers have a tuning parameter for taking a probability estimate and forming classifications

Example: Logistic regression where $\mathcal{G} = \{0, 1\}$

$$\widehat{\mathbb{P}(Y = 1|X)} = \hat{\pi}(X) = \frac{e^{X^\top \hat{\beta}}}{1 + e^{X^\top \hat{\beta}}}$$

This gets converted to a classifier as:

$$\hat{g}(X) = \begin{cases} 0 & \text{if } \hat{\pi}(X) < \tau \\ 1 & \text{if } \hat{\pi}(X) > \tau \end{cases}$$

The Bayes' rule is $g_*(X) = \arg\max_g \mathbb{P}(Y = g|X)$

In the two class problem, $\mathcal{G} = \{0, 1\}$, this is the same as

$$g_*(X) = \begin{cases} 0 & \text{if } \mathbb{P}(Y = 1|X) < 0.5 = \tau_* \\ 1 & \text{if } \mathbb{P}(Y = 1|X) > 0.5 = \tau_* \end{cases}$$

Receiver operating characteristic

As the Bayes' rule uses $\tau = \tau_* = 0.5$, often that is the default

However, τ is a tuning parameter

The receiver operating characteristic (ROC) curve compares the true positive and false positive rates for $0 \leq \tau \leq 1$

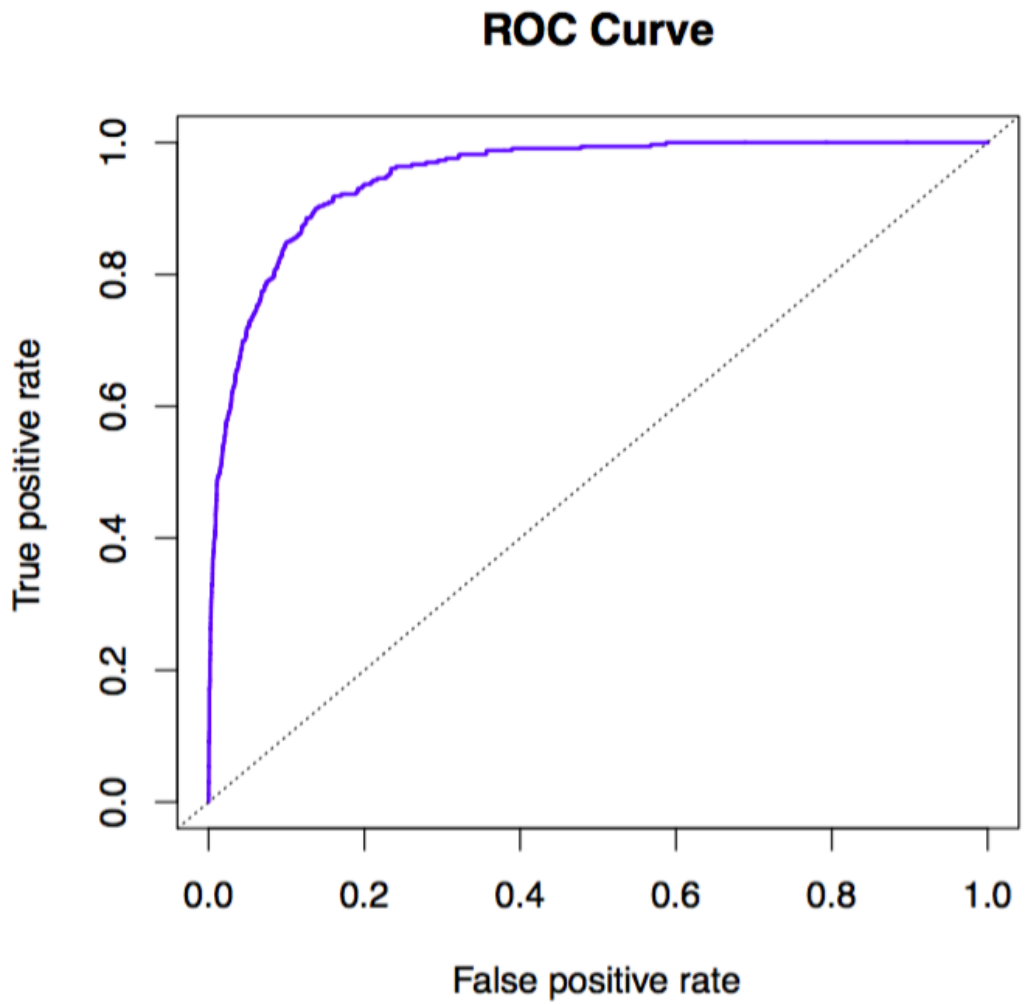
These are sensitivity/recall and 1-specificity, respectively

The area under the ROC curve summarizes the plot and is called AUC

$(0 \leq \mathrm{AUC} \leq 1)$

The interpretation: a procedure with larger AUC is better and AUC ≈ 1 is best

ROC curve


ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Confusion matrix

We can compute any of these metrics with training or test data

Like with estimating the risk, use the test based version

Suppose we train some procedure and get the following test confusion matrix

		Truth	
		Recession	No Recession
Our Predictions	Recession	(A)	(B)
	No Recession	(C)	(D)

The test misclassification rate is

$$\frac{(\mathrm{B}) + (\mathrm{C})}{(\mathrm{A}) + (\mathrm{B}) + (\mathrm{C}) + (\mathrm{D})} = \frac{(\mathrm{B}) + (\mathrm{C})}{n_{\text{test}}}$$

What is the sensitivity/specificity?

Confusion matrix

We can compute any of these metrics with training or test data

Like with estimating the risk, use the test based version

Suppose we train some procedure and get the following test confusion matrix

		Truth	
		Recession	No Recession
Our Predictions	Recession	(A)	(B)
	No Recession	(C)	(D)

The test misclassification rate is

$$\frac{(\mathrm{B}) + (\mathrm{C})}{(\mathrm{A}) + (\mathrm{B}) + (\mathrm{C}) + (\mathrm{D})} = \frac{(\mathrm{B}) + (\mathrm{C})}{n_{\text{test}}}$$

What is the sensitivity/specificity?

(Sensitivity is (A)/[(A) + (C)], Specificity is (D)/[(B) + (D)])

Multi-class classification

Suppose $|\mathcal{G}| > 2$

(That is, suppose there are more than two possible classes for the supervisor)

The confusion matrix and misclassification rates can be generalized to any number of classes

However, sensitivity/specificity, precision/recall, or ROC are not well defined