Classification Metrics

-Applied Multivariate Analysis & Statistical Learning-

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

The total number of each combination is recorded in the table

The overall misclassification rate is

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

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

Kappa score

The Kappa score is the degree to which the classifications match the truth relative to what would be expected if they were independent

$$\kappa = \frac{O - E}{1 - E}$$

- $O = (TP + TN)/n_{\text{total}}$

  (This is 1 - misclassification rate)

- $E = \left(\frac{(TP+FP)(TP+FN)}{n_{\text{total}}^2}\right) + \left(\frac{(FN+TN)(FP+TN)}{n_{\text{total}}^2}\right)$

  (This strange formula is estimating the probability that the classifier and the truth would take the same level if they are independent)

(Also, it only makes sense if the confusion matrix is computed on test data)

Receiver operating characteristic

The probability threshold

Example: We could train a classifier to classify whether a plane needs to be serviced

		Truth	
		Serviced	Not serviced
Our	Serviced	Plane gets fixed	plane could have been flying
Preds.	Not serviced	Potential crash	plane flies

Here, the economic costs of not having a plane available when it could have been don't compare to flying an unfit plane

One way to incorporate these ideas is via adjusting the threshold

Sensitivity and specificity

In this example, the 'interesting' case is the plane needs to be serviced

Hence,

- sensitivity is $\mathbb{P}(\hat{Y} = \text{serviced} | Y = \text{serviced})$

  (and the $\widehat{\mathbb{P}}(\hat{Y} = \text{serviced} | Y = \text{serviced}) = TP/P$)
- specificity is $\mathbb{P}(\hat{Y} = \text{not serviced} | Y = \text{not serviced})$

  (and the $\widehat{\mathbb{P}}(\hat{Y} = \text{not serviced} | Y = \text{not serviced}) = TN/N$)

The probability of the critical error in this case is 1 - sensitivity

Using the 0.5 threshold might result in a specificity of .87 and sensitivity of .34

Question: If we have the event as 'serviced' and adjust the threshold to $\tau = 0.8$, then will the sensitivity go up or down?

Sensitivity and specificity

In this example, the 'interesting' case is the plane needs to be serviced

Hence,

- sensitivity is $\mathbb{P}(\hat{Y} = \text{serviced}|Y = \text{serviced})$
  (and the $\widehat{\mathbb{P}}(\hat{Y} = \text{serviced}|Y = \text{serviced}) = TP/P$)
- specificity is $\mathbb{P}(\hat{Y} = \text{not serviced}|Y = \text{not serviced})$
  (and the $\widehat{\mathbb{P}}(\hat{Y} = \text{not serviced}|Y = \text{not serviced}) = TN/N$)

The probability of the critical error in this case is 1 - sensitivity

Using the 0.5 threshold might result in a specificity of .87 and sensitivity of .34

Question: If we have the event as 'serviced' and adjust the threshold to $\tau = 0.8$, then will the sensitivity go up or down? Down!

Receiver operating characteristic

We can therefore adjust the sensitivity and specificity by adjusting the threshold

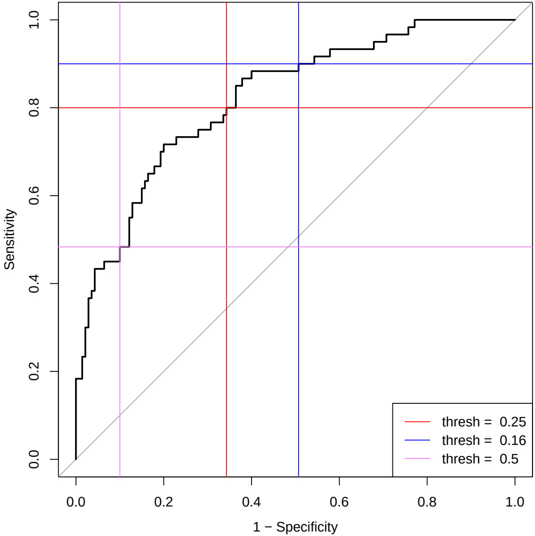
In fact, this threshold is another example of a tuning parameter

We will get a (potentially) new classifier for any value of the threshold from [0,1]

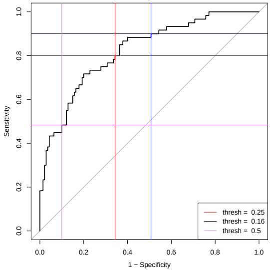
The receiver operating characteristic (ROC) plots three things:

- The sensitivity/recall
- 1-specificity (false positive)
- the threshold

Receiver operating characteristic



Receiver operating characteristic



The ROC plot can be used in a few different ways

- Perhaps we state that we will not accept a classifier unless it has at least a sensitivity of .8 on test data
  $\rightarrow$ set the threshold less than 0.25
- Alternatively, we can get a quantitative measurement about a classifier averaged across all possible values of the threshold
  $\rightarrow$ get the area under the ROC curve

Area under the ROC curve

The area under the ROC curve give a summary of 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

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

The test misclassification rate is

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

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

The test misclassification rate is

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

What is the sensitivity/specificity?

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

Multi-class classification

Most of these notes pertain to the 2-class problem: $|\mathcal{G}| = 2$

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

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

Then the following are essentially undefined:

- kappa
- ROC/AUC
- sensitivity/specificity
- precision/recall

(There do exist some extensions, but they are very awkward)

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