Lecture 10: Classification examples

Reading: Chapter 4

STATS 202: Data mining and analysis

Jonathan Taylor, 10/15
Slide credits: Sergio Bacallado

Example. Predicting `default`

Used LDA to predict credit card default in a dataset of 10K people.

Predicted "yes" if $P(\texttt{default} = \text{yes}|X) > 0.5$.

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

Example. Predicting `default`

Used LDA to predict credit card default in a dataset of 10K people.

Predicted "yes" if $P(\texttt{default} = \text{yes}|X) > 0.5$.

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

- The error rate among people who do **not** default (false positive rate) is very low.
- However, the rate of false negatives is 76%.
- It is possible that false negatives are a bigger source of concern!
- One possible solution: Change the threshold.

Example. Predicting `default`

Changing the threshold to 0.2 makes it easier to classify to "yes".

Predicted "yes" if $P(\texttt{default} = \textsf{yes}|X) > 0.2$.

		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
default status	Yes	235	195	430
	Total	9,667	333	10,000

Example. Predicting `default`

Changing the threshold to 0.2 makes it easier to classify to "yes".

Predicted "yes" if $P(\texttt{default} = \mathsf{yes}|X) > 0.2$.

		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
default status	Yes	235	195	430
	Total	9,667	333	10,000

Note that the rate of false positives became higher! That is the price to pay for fewer false negatives.
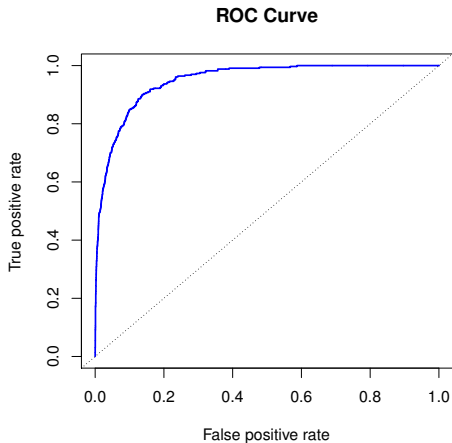
Example. Predicting `default`

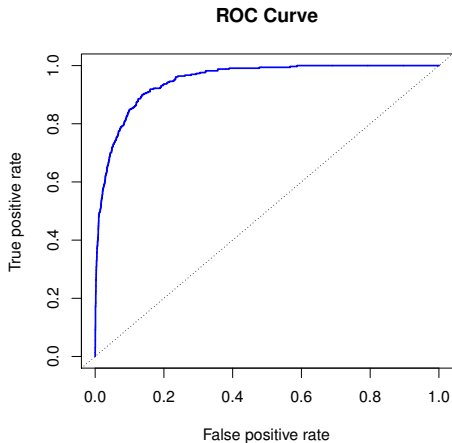Let's visualize the dependence of the error on the threshold:
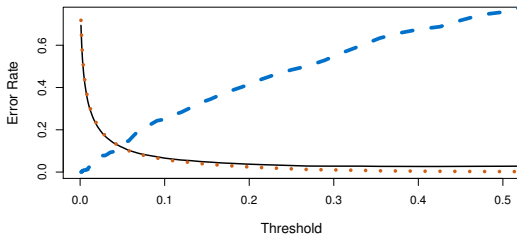
- ▸ – – – False negative rate (error for defaulting customers)
- ▸ · · · · False positive rate (error for non-defaulting customers)
- ▸ ——— 0-1 loss or total error rate.

Example. The ROC curve

ROC Curve

- Displays the performance of the method for any choice of threshold.

Example. The ROC curve

ROC Curve

- Displays the performance of the method for any choice of threshold.

- The area under the curve (AUC) measures the quality of the classifier:

    - 0.5 is the AUC for a random classifier

    - The closer AUC is to 1, the better.