Lecture 22: Binary classification

Ciaran Evans

Types of research questions

For a logistic regression model, we have learned how to answer the following types of questions:

- What is the predicted probability for each observation in the data?
- What is the relationship between the explanatory variable(s) and the response?
- Do we have strong evidence for a relationship between these variables?

Another research question:

- How well do we predict the response?

Making predictions with the Titanic data

- For each passenger, we calculate $\widehat{p}_i$ (estimated probability of survival)
- But, we want to predict *which* passengers actually survive

Question: How do we turn $\widehat{p}_i$ into a binary prediction of survival / no survival?

Confusion matrix

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	344	70
	$\widehat{Y} = 1$	80	220

Question: Did we do a good job predicting survival?

Why a threshold of 0.5?

Question: Why might a threshold of 0.5 be a common choice when making binary predictions?

Why a threshold of 0.5?

Consider data (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. Fit a model to estimate

$$p(x) = P(Y = 1|X = x)$$

Our binary predictions are

$$\widehat{Y} = \begin{cases} 1 & p(x) \geq h \\ 0 & p(x) < h \end{cases}$$

The **classification error** is given by $P(\widehat{Y} \neq Y)$.

Claim: For any binary classifier, $h = 0.5$ minimizes classification error.

Why a threshold of 0.5?

Claim: For any binary classifier, $h = 0.5$ minimizes classification error.

Another confusion matrix

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	3957	1631
	$\widehat{Y} = 1$	66	66

Question: Did we do a good job predicting the response?

Classification metrics

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	3957	1631
	$\widehat{Y} = 1$	66	66

Accuracy: $\widehat{P}(\widehat{Y} = Y) = \dfrac{TP + TN}{\text{total}}$

Sensitivity: $\widehat{P}(\widehat{Y} = 1 | Y = 1) = \dfrac{TP}{TP + FN}$

Specificity: $\widehat{P}(\widehat{Y} = 0 | Y = 0) = \dfrac{TN}{TN + FP}$

Changing the threshold

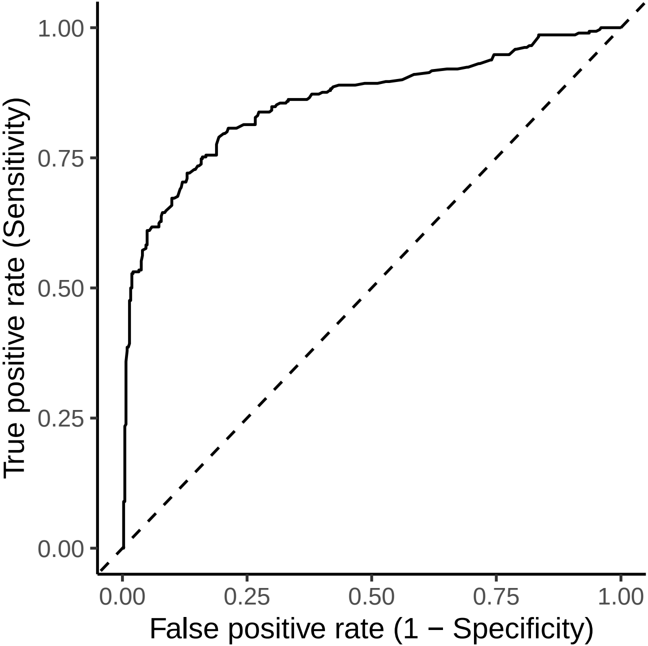
Threshold of 0.7:

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	412	136
	$\widehat{Y} = 1$	12	154

Threshold of 0.3:

		Actual	
		$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	309	49
	$\widehat{Y} = 1$	115	241

ROC curve: consider all thresholds



Binary classification vs. hypothesis testing

- Both binary classification and hypothesis testing involve deciding between two options
- Error metrics for both involve looking at correct decisions, false positives (type I errors), false negatives (type II errors)

Question: How do binary classification and hypothesis testing *differ*?

Binary classification vs. hypothesis testing

Binary classification:

- Can use training data to estimate performance and so choose a threshold
- Thresholds are chosen to maximize some combination of sensitivity and specificity

Hypothesis testing:

- Conceptually a two-step approach: control type I error, then hope to have good power (i.e., don't consider tests which have high type I error)
- Only see one test result; don't get to estimate type I error or power from a single test
- Want theoretical guarantees that (if assumptions are met) type I error can be controlled at desired level

Binary classification vs. hypothesis testing

- Usual approach to binary classification: maximize some combination of sensitivity and specificity
- Neyman-Pearson classification[1]: control probability of false positives (1 - specificity) at desired level, then try to maximize sensitivity

Question: Why might you choose one of these approaches over the other?

[1] Scott, C., & Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11), 3806-3819.