Binary predictions

- Challenge 5 (logistic regression in Python) released

Types of research questions

So far, we have learned how to answer the following questions:

+ what is the probability for observation i in the data?
+ What is the relationship between the explanatory variable(s) and the response? (fitting & interpreting model)
+ What is a "reasonable range" for a parameter in this relationship? (confidence intervals)
+ Do we have strong evidence for a relationship between these variables? (hypothesis testing)

> What other kinds of research questions might we ask?

- How well will I predict on new observations? / How well do I predict the response?
- what model should I use to predict the response? / what variables are important?

Making predictions with the Titanic data

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

> How do we turn $\hat{p}_i$ into a binary prediction of survival / no survival?

$$\hat{y}_i = \begin{cases} 1 & \hat{p}_i \geq 0.5 \leftarrow \text{threshold} \\ 0 & \hat{p}_i < 0.5 \end{cases}$$

Confusion matrix

correct (True negatives)

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

Incorrect (False Negatives)

correct (True Positives)

Incorrect (False positives)

> Did we do a good job predicting survival?

$$\text{Accuracy} = \frac{\#\text{ correct predictions}}{\#\text{ observations}} = \frac{TP + TN}{n} = \frac{220 + 344}{714}$$

$$= 0.79$$

If I randomly select an observation, what is the probability my prediction is correct?

1 - Accuracy = classification error

Another confusion matrix (Dengue):

precision: $P(Y=1 \mid \hat{Y}=1)$
(PPV)

	$Y=0$	$Y=1$
$\hat{Y}=0$	3957	1631
$\hat{Y}=1$	66	66

Accuracy: $\frac{3957+66}{5720} = 0.703$

$\approx$ 70% of the patients don't have Dengue

Problem: Accuracy is misleading with imbalanced data

How well did we do within group?

$$\underbrace{P(\hat{Y}=1 \mid Y=1)}_{\text{sensitivity}} = \frac{66}{1631+66} = \frac{TP}{TP+FN} = 0.039$$

sensitivity (aka recall, or TPR)

$$\underbrace{P(\hat{Y}=0 \mid Y=0)}_{\text{specificity}} = \frac{3957}{3957+66} = \frac{TN}{TN+FP} = 0.984$$

specificity (aka TNR)

Why a threshold of 0.5?

Consider data (X, Y) $X \in \mathbb{R}^d$ and $Y \in \{0,1\}$

Fit a model to estimate $p(x) = P(Y=1 \mid X=x)$

we want binary predictions: $\hat{Y} = \begin{cases} 1 & p(X) \geq h \\ 0 & p(X) < h \end{cases}$

<u>Classification error</u>: $P(Y \neq \hat{Y})$

(1 - accuracy)

<u>Claim</u>: $h = 0.5$ minimizes classification error among <u>all</u> binary classifiers

<u>Proof</u>: Let $C(X)$ be an arbitrary classification function. $C(X) \in \{0,1\}$, $\hat{Y} = C(X)$

$$P(Y \neq \hat{Y}) = P(Y \neq C(X)) = \mathbb{E}[\mathbb{1}\{C(X) \neq Y\}]$$

$$\mathbb{1}_A = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

$$P(X \in A) = \int_A f(x)\,dx = \int_{-\infty}^{\infty} f(x) \cdot \mathbb{1}_A(x)\,dx = \mathbb{E}[\mathbb{1}_A(X)]$$

(iterated expectations)

$$\mathbb{E}[\mathbb{1}\{C(X) \neq Y\}] = \mathbb{E}\Big[\underbrace{\mathbb{E}[\mathbb{1}\{C(X) \neq Y\} \mid X]}\Big]$$

$$p(X) = P(Y = 1 \mid X)$$

$$= \begin{cases} \mathbb{E}[\mathbb{1}\{Y=0\}] & C(X) = 1 \\ \mathbb{E}[\mathbb{1}\{Y=1\}] & C(X) = 0 \end{cases}$$

$$= \mathbb{E}[\mathbb{1}\{Y=0\}]\,C(X) + \mathbb{E}[\mathbb{1}\{Y=1\}]\,(1 - C(X))$$

$$= P(Y=0)\,C(X) + P(Y=1)\,(1 - C(X))$$

$$= (1 - p(X))\,C(X) + p(X)\,(1 - C(X))$$

$$\Rightarrow P(Y \neq \hat{Y}) = \mathbb{E}\big[(1 - p(X))\,C(X) + p(X)\,(1 - C(X))\big]$$

$$\Rightarrow P(Y \neq \hat{Y}) = \mathbb{E}\left[(1-p(X))\, c(X) + p(X)\,(1-c(X)) \right]$$

$$= \int_{x} \left[(1-p(x))\, c(x) + p(x)\,(1-c(x)) \right] f(x)\, dx$$

<u>Minimize</u> $P(Y \neq \hat{Y})$: make $(1-p(x))c(x) + p(x)(1-c(x))$ small!

If $1-p(x) < p(x)$: $c(x) = 1$

$1-p(x) > p(x)$: $c(x) = 0$

$$\Rightarrow \boxed{\text{Bayes classifier}}\quad c(x) = \begin{cases} 1 & 1-p(x) \leq p(x) \\ 0 & 1-p(x) > p(x) \end{cases} = \begin{cases} 1 & p(x) \geq 0.5 \\ 0 & p(x) < 0.5 \end{cases}$$

$\Rightarrow$ to minimize $\underbrace{\text{classification error}}_{\text{Bayes risk}}$, choose $h = 0.5$

//

Tradeoff: increase threshold ⇒ sensitivity ↓
specificity ↑

Changing the threshold

Accuracy = sensitivity $P(Y=1)$ + spec. $P(Y=0)$

Using a threshold of 0.7:

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

sensitivity: $\frac{154}{290}$

specificity = $\frac{412}{424}$

Using a threshold of 0.3:

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

sensitivity: $\frac{241}{290}$

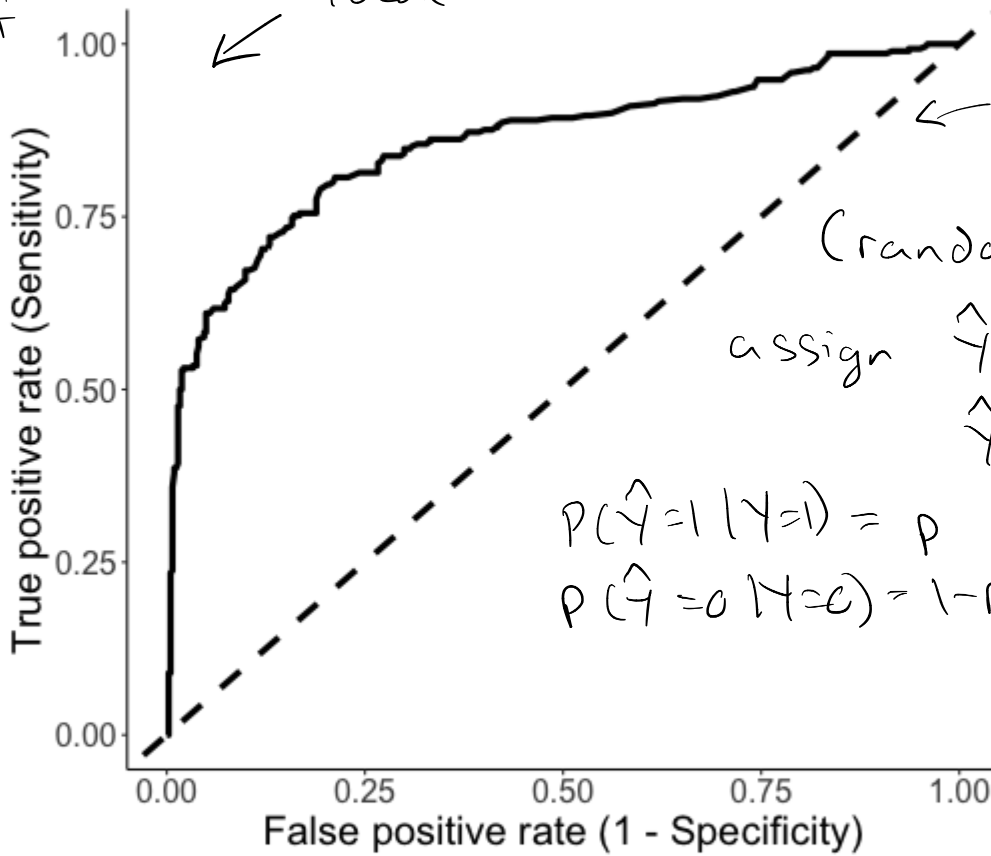
specificity: $\frac{309}{424}$

receiver operating characteristic

ROC curve: consider all thresholds

ideal if ROC curve is close to top left

plots trade-off between sensitivity & specificity



random classifier (random guessing)

assign $\hat{Y} = 1$ w/ prob p

$\hat{Y} = 0$ w/ prob $1-p$

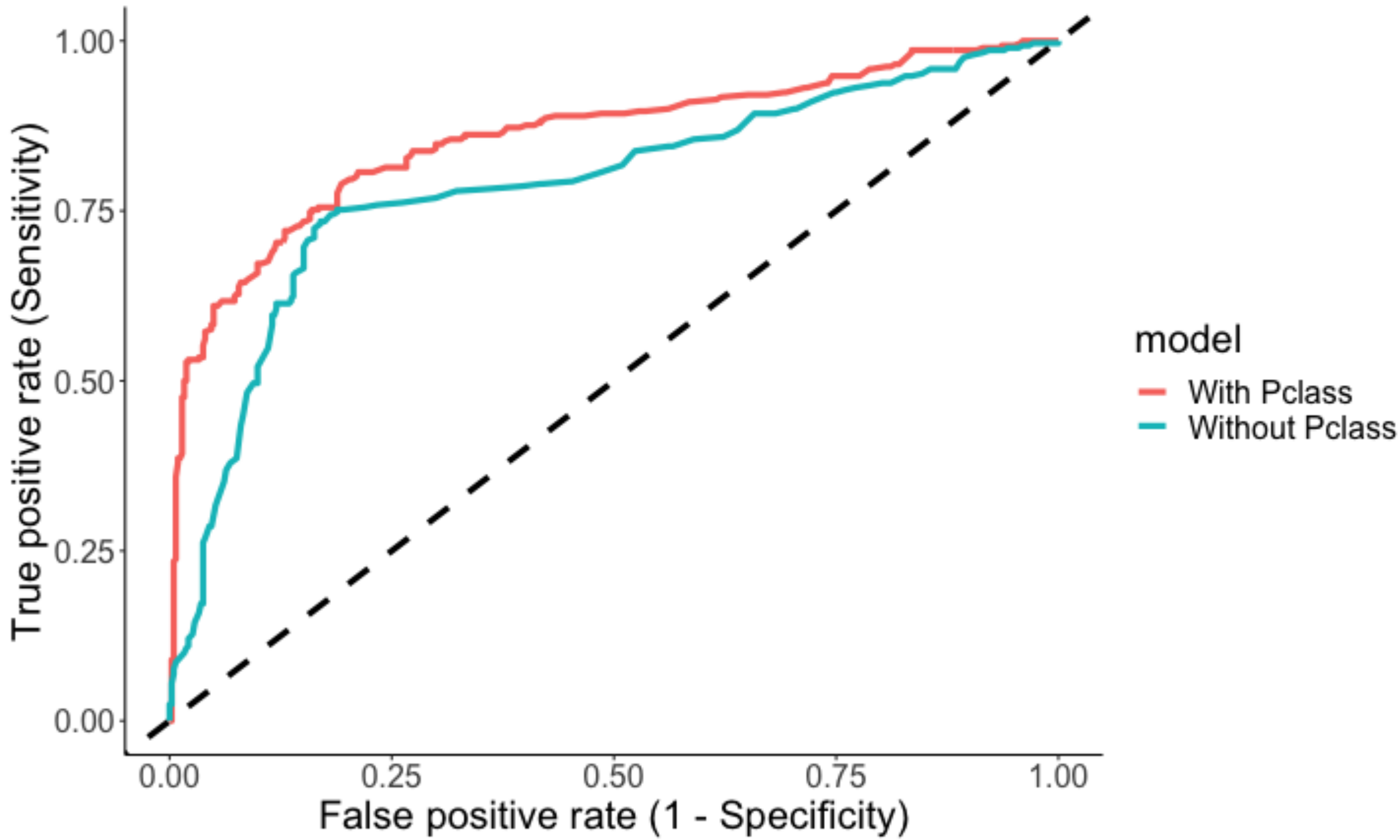
$P(\hat{Y}=1 \mid Y=1) = p \Rightarrow$ Sens. $= p$

$P(\hat{Y}=0 \mid Y=0) = 1-p \Rightarrow$ 1-spec. $= p$

summarizing: Area under curve (AUC)

$0.5 \leq AUC \leq 1$, closer to 1 is better

Comparing models with ROC curves



Problem: reusing data...

It is generally a bad idea to assess performance of a model on the same data we used to train it. This can lead to overfitting.

What can we do instead?