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?

$$\hat{Y}_i = \begin{cases} 1 & \hat{p}_i \geq 0.5 \\ 0 & \hat{p}_i < 0.5 \end{cases} \quad \text{(threshold)}$$

(pick value (0 or 1) with the higher probability)

Confusion matrix

		Actual	
	(TN)	$Y = 0$	$Y = 1$
Predicted	$\widehat{Y} = 0$	344 ✓	70
	$\widehat{Y} = 1$	80	220 ✓

FN

(TP)

FP

Question: Did we do a good job predicting survival?

Accuracy : $\dfrac{TP + TN}{\text{total \# observations}} = \dfrac{220 + 344}{714} \approx 0.79$

Accuracy = probability of a correct prediction for a randomly selected observation

Classification error = 1 − accuracy

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?

explanatory variables

binary outcome

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

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

e.g., this is what logistic regression models

Our binary predictions are

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

threshold

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?

$$\underbrace{\mathbb{E}[Y]}_{\text{expectation wrt } Y} = \underbrace{\mathbb{E}[\underbrace{\mathbb{E}[Y|X]}_{\text{function of } X}]}_{\text{expectation in terms of } X}$$

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

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

Pf: Let $C(X)$ denote classification function for a binary classifier. $C(X) = \hat{Y} \in \{0,1\}$

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

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

$$= \mathbb{E}[\mathbb{1}\{Y=0\} | X] \quad \text{if } C(X) = 1$$

$$= P(Y=0|X)$$

$$= \mathbb{E}[\mathbb{1}\{Y=1\} | X] \quad \text{if } C(X) = 0$$

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

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

Why a threshold of 0.5?

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

$$\underbrace{P(\hat{Y} \neq Y)}_{\text{want to } \underline{\text{minimize}}} = \mathbb{E}\left[(1-p(X))\,C(X) + p(X)(1-C(X))\right]$$

$$= \int_x \underbrace{\left[(1-p(x))\,C(x) + p(x)(1-C(x))\right]}_{\underline{\text{minimize}} \text{ integrand for each } x \text{ to minimize integral}} f(x)\,dx$$

Integrand: either $1-p(x)$ or $p(x)$

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

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

$\boxed{\text{Bayes classifier}}$ $\Rightarrow$ $C(x) = \begin{cases} 1 & p(x) \geq 0.5 \\ 0 & p(x) < 0.5 \end{cases}$ to minimize $P(\hat{Y} \neq Y)$ //

Connection to N-P:

$$P(Y=1 \mid X) = \frac{f(X \mid Y=1)\, P(Y=1)}{f(X \mid Y=1) P(Y=1) + f(X \mid Y=0) P(Y=0)}$$

$$= \frac{f(X \mid Y=1)}{f(X \mid Y=1) + f(X \mid Y=0) \dfrac{P(Y=0)}{P(Y=1)}}$$

$$\Rightarrow \quad P(Y=1 \mid X) > 0.5$$

$$\Leftrightarrow \quad f(X \mid Y=1) > f(X \mid Y=0) \frac{P(Y=0)}{P(Y=1)}$$

$$\Leftrightarrow \quad \frac{f(X \mid Y=1)}{f(X \mid Y=0)} > \frac{P(Y=0)}{P(Y=1)}$$

Summary:

N-P test:

- reject H_0 if $\frac{f(X|\theta_1)}{f(X|\theta_0)} > k$
- choose k st $\beta(\theta_0) = \alpha$
- No probabilities on hypotheses (i.e, no $P(\theta=\theta_0)$)

Bayes classifier:

- $\hat{Y} = 1$ if $\frac{f(X|Y=1)}{f(X|Y=0)} > \frac{P(Y=0)}{P(Y=1)}$
- This threshold minimizes $P(\hat{Y} \neq Y)$ (maximizes accuracy)
- If $P(Y=1) = P(Y=0) = \frac{1}{2}$, just pick Y with higher $f(X|Y)$

Another confusion matrix

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

FN

TP

FP

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

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

Exactly the same accuracy as if we set $\hat{Y} = 0$ for everyone in the data

Problem : imbalanced classes ($P(Y=0) = 70\%$)

Sensitivity : $\hat{P}(\hat{Y} = 1 \mid Y = 1) = \frac{TP}{TP + FN)} = \frac{66}{66 + 1631} = 0.039$

Specificity : $\hat{P}(\hat{Y} = 0 \mid Y = 0) = \frac{TN}{TN + FP} = 0.984$

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

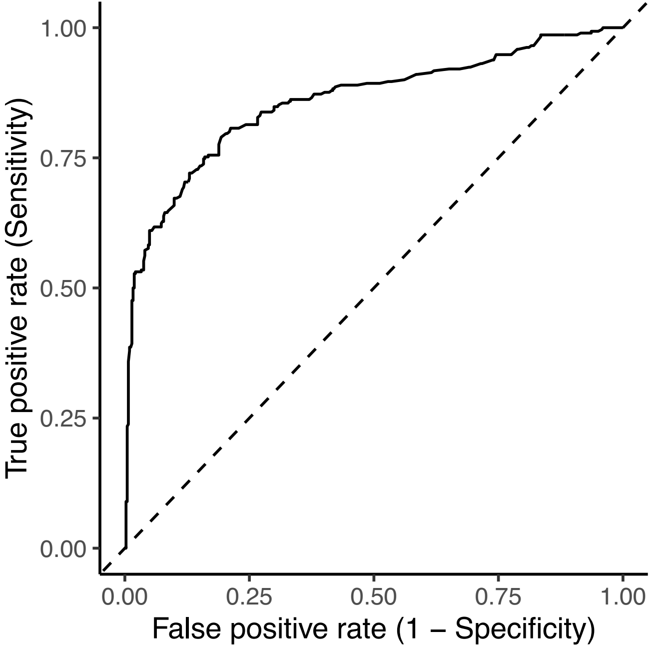
Threshold of 0.7:

		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.