

## Logistic Regression concepts

*In class exercise: Week 4*

1. Suppose we have collected data for a group of students in the Analytics Edge lecture from the previous year with variables  $X_1$  = hours studied,  $X_2$  = CGPA until the previous term, and  $Y$  = binary variable which indicates if they received an A or not. We fit a logistic regression and produce estimated coefficients:

$$\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1.$$

- (a) Estimate the probability that a student who studies for 40 hours and has a CGPA of 3.5 gets an A in the class.
  - (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?
2. Consider the following two confusion matrices:

**Confusion matrix 1:**

	Actual = 0	Actual = 1
Predicted = 0	15	5
Predicted = 1	10	20

**Confusion matrix 2:**

	Actual = 0	Actual = 1
Predicted = 0	20	10
Predicted = 1	5	15

- (a) What is the sensitivity of Confusion Matrix 1?
- (b) What is the specificity of Confusion Matrix 1?
- (c) To go from Confusion Matrix 1 to Confusion Matrix 2, did we increase or decrease the threshold value?

3. The question involves the following ROC curve. In this healthcare set up, we categorise patients according to the kind of care they get. The baseline is supposed to be good care which gives a value 0 and you want to detect cases where poor care of patients happen which is categorized as 1.

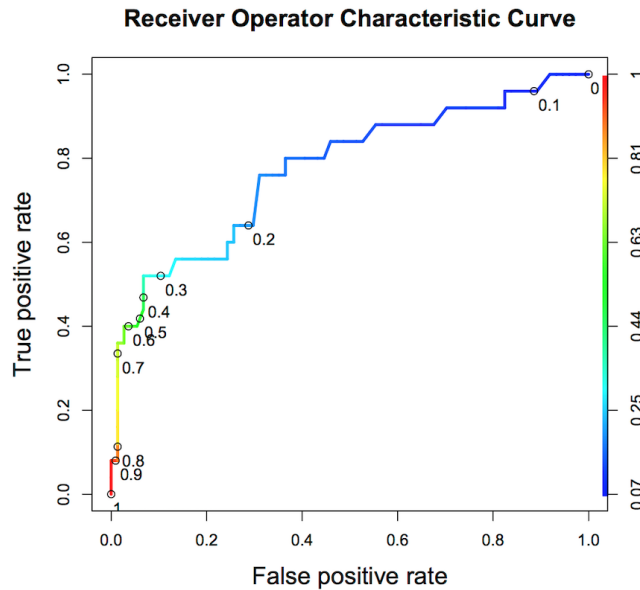


Figure 1: ROC curve

- (a) Given this ROC curve, which threshold would you pick if you wanted to correctly identify a small group of patients who are receiving the worst care with high confidence?

Select the best option.

- (i)  $t = 0.2$ .
- (ii)  $t = 0.3$ .
- (iii)  $t = 0.7$ .
- (iv)  $t = 0.8$ .

- (b) Which threshold would you pick if you wanted to correctly identify half of the patients receiving poor care, while making as few errors as possible?

Select the best option.

- (i)  $t = 0.2$ .
- (ii)  $t = 0.3$ .
- (iii)  $t = 0.7$ .
- (iv)  $t = 0.8$ .