The Analytics Edge SUMMER 2022

Logistic Regression concepts

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.

   Answer. $\dfrac{e^{-6+0.05\,\texttt{Hour}+\texttt{CGPA}}}{1+e^{-6+0.05\,\texttt{Hour}+\texttt{CGPA}}} = 0.3775.$

   (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?

   Answer. We equate $\dfrac{e^{-6+0.05\,\texttt{Hour}+3.5}}{1+e^{-6+0.05\,\texttt{Hour}+3.5}} = \dfrac{1}{2}.$

   Now solve $e^{-2.5+0.05\,\texttt{Hour}} = 1$ to get `Hours = 50`.

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?

   Answer. Sensitivity $= \dfrac{\texttt{TP}}{\texttt{TP+FN}} = \dfrac{20}{20+5} = 0.8.$

   (b) What is the specificity of Confusion Matrix 1?

   Answer. Specificity $= \dfrac{\texttt{TN}}{\texttt{TN+FP}} = \dfrac{15}{15+10} = 0.6.$

   (c) To go from Confusion Matrix 1 to Confusion Matrix 2, did we increase or decrease the threshold value?

   Answer. In Confusion Matrix 2, we predict the outcome 1 fewer number of times. Hence the threshold must have increased.

3. The question involves the following ROCR 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.

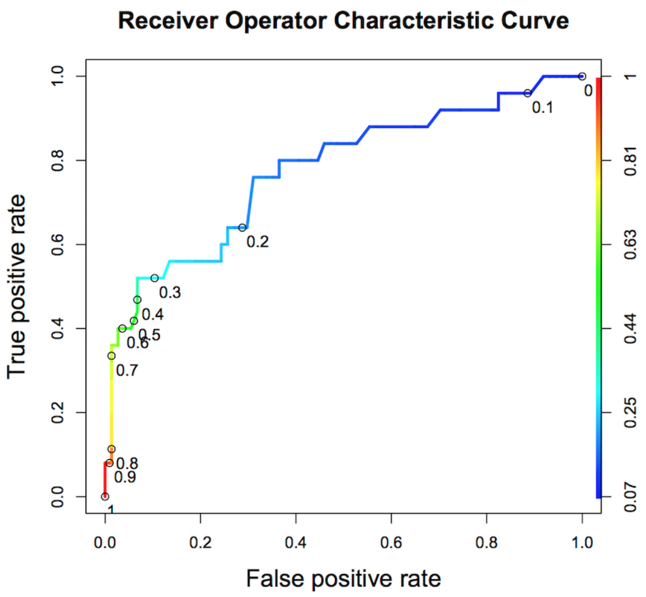

Figure 1: ROC curve

(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) $t = 0.8$.

The threshold 0.7 is best to identify a small group of patients who are receiving the worst care with high confidence, since at this threshold we make very few false positive mistakes, and identify about 35% of the true positives. The threshold t = 0.8 is not a good choice, since it makes about the same number of false positives, but only identifies 10% of the true positives. The thresholds 0.2 and 0.3 both identify more of the true positives, but they make more false positive mistakes, so our confidence decreases.

(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) $t = 0.8$

The threshold 0.3 is the best choice in this scenario. The threshold 0.2 also identifies over half of the patients receiving poor care, but it makes many more false positive mistakes. The thresholds 0.7 and 0.8 don't identify at least half of the patients receiving poor care.