Classification Assessment

Data Science with Python CS677

Farshid Alizadeh-Shabdiz, PhD, MBA
Alizadeh@bu.edu
Fall 2021

1

Classification Assessment

2

How to Quantify the Fitness in Logistic Regression?

Let us assume the output of logistic regression is 0.37 and 0.84.

- If all the samples are decided correctly, it is perfect!
- What if the result is not perfect!

3

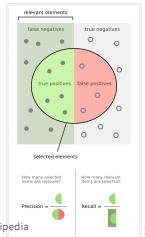
True/False Positive and True/False Negative

- **True positive.** Rate of **Correct** identification of **Positive** category.
  Example: Sick people correctly identified as sick
- **False positive.** Rate of **Incorrect** identification of **Positive** category.
  Example: Healthy people incorrectly identified as sick
- **True negative.** Rate of **Correct** identification of **Negative** category.
  Example: Healthy people correctly identified as healthy
- **False negative.** Rate of **Incorrect** identification of **Negative** category.
  Example: Sick people incorrectly identified as healthy

In general, Positive = identified and negative = rejected.

Therefore:

- **True positive = correctly identified**
- **False positive = incorrectly identified**
- **True negative = correctly rejected**
- **False negative = incorrectly rejected**



Reference: Wikipedia

4

Example

- We have identified 30 sick people **correctly identified** as sick.
- We have identified 4 healthy people **incorrectly identified** as sick.
- We have identified 25 healthy people **correctly identified** as healthy.
- We have identified 3 sick people **incorrectly identified** as healthy.

5

Example

- We have identified 30 sick people **correctly identified** as sick.
- We have identified 4 healthy people **incorrectly identified** as sick.
- We have identified 25 healthy people **correctly identified** as healthy.
- We have identified 3 sick people **incorrectly identified** as healthy.

- **True Positive:** 30
- **False Positive:** 4
- **True Negative:** 25
- **False Negative:** 3

6

Example - Confusion Matrix

- We have identified 30 sick people **correctly identified** as sick.
- We have identified 4 healthy people **incorrectly identified** as sick.
- We have identified 25 healthy people **correctly identified** as healthy.
- We have identified 3 sick people **incorrectly identified** as healthy.

- **True Positive (TP):** 30
- **False Positive (FP):** 4
- **True Negative (TN):** 25
- **False Negative (FN):** 3

		True Conditions	
		Condition Positive	Condition Negative
Predicted Class	Positive	TP=30	FP=4
	Negative	FN=3	TN=25

7

Sensitivity & Specificity

Sensitivity (also called Recall, Hit Rate, or true positive rate (TPR)) **measures the proportion of actual positives** that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

$$\text{Sensitivity or Recall} = \frac{TP}{TP+FN}$$

Specificity (also called Selectivity or true negative rate (TNR)) **measures the proportion of actual negatives** that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

$$\text{Specificity} = \frac{TN}{TN+FP}$$

8

Precision

Precision (also called positive predictive value) is the **fraction of relevant instances among the retrieved instances**.

$$\text{Precision} = \frac{TP}{TP+FP}$$

9

F-Measure, F1 Measure

F-Measure provides a way to **combine Precision and Recall** as a measure of the overall effectiveness of a Classification algorithm.

F-Measure is calculated as a ratio of the weighted importance of Precision and Recall. The general formula for **positive real β number** is:

$$F_\beta = (1+\beta^2) * \frac{Precision \times Recall}{(\beta \times Precision) + Recall}$$

"F Measure measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision" (Rijsbergen et al.)

Using Precision and Recall Formula:

$$F_\beta = \frac{(1+\beta^2) \times TP}{(1+\beta^2) \times TP + \beta^2 \times FN + FP}$$

10

F1 Measure, F2, F0.5

F1 Score: A measure that combines precision and recall is the **harmonic mean** of precision and recall.

$$\text{F1 Score} = 2 * \frac{Precision \times Recall}{Precision + Recall}$$

- F_2 weighs recall higher than precision (by placing **more emphasis on false negatives**)
- $F_{0.5}$ weighs recall lower than precision (by attenuating the **influence of false negatives**)

11

Example

- **True Positive:** 30
- **False Positive:** 4
- **True Negative:** 25
- **False Negative:** 3

- What is the Sensitivity or Recall for this example?
- What is the Specificity for this example?
- What is the F1-Score for this example?

12

Example

- **True Positive (TP):** 30
- **False Positive (FP):** 4
- **True Negative (TN):** 25
- **False Negative (FN):** 3

$$\text{Sensitivity or Recall} = \frac{TP}{TP+FN} = \frac{30}{30+3} = 0.9091 \text{ or } 90.9\%$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{25}{25+4} = 0.8621 \text{or } 86.21\%$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{30}{30+4} = 0.8823 \text{ or } 90.9\%$$

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.88 \times 0.91}{0.88 + 0.91} = 0.89$$

13

Receiver Operating Characteristic Curve (ROC Curve)

14

ROC (Receiver Operating Characteristic)

Sensitivity and specificity are important measure in assessing the fit of a logistic regression model.

Models with one or more continuous explanatory or independent variables have more possible values for the predicted probabilities and therefore **there are often many cutoffs that produce distinct values of sensitivity and specificity**.

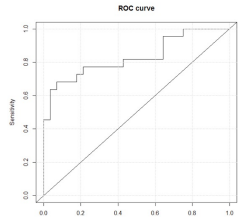
The area under the **ROC (receiver operating characteristic) curve (also known as the c-statistic)** is a measure of the sensitivity and specificity across the range of all possible cutoffs.

It is used to measure the goodness of fit of a logistic regression model.

15

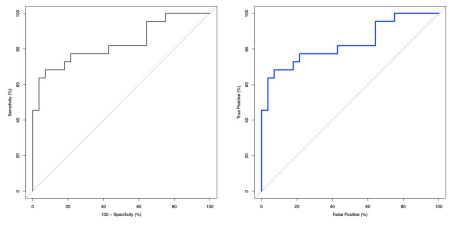
ROC (Receiver Operating Characteristic)

- The ROC curve is a plot of corresponding pairs of **sensitivity (y-axis)** and **1 minus the specificity (x-axis) for each possible cutoff point**.
- **It ranges between 0.5 and 1.0 with larger values indicating better fit.**
- **When the area under the curve is equal to 0.50**, it is said that the model **does no better at classifying events than at random or by chance**.



16

ROC - False Positive Rate, True Positive Rate Rate



17

Cross Entropy as an Assessment Tool

- Cross entropy is an appropriate tool to compare classifiers
- But
  - Cross entropy doesn't relate to a meaningful fact
  - In the absolute sense doesn't provide any insight into classifier
  - It is not bounded. It goes from zero to infinity

18