BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Classification

- Regression assumes that the response variable Y is numerical.
- In many situations, the response variable is a categorical variable, e.g. the solvency of a customer.
- The process of prediction of a categorical variable is called classification.
- Predicting a categorical variable implies assigning a new observation to one of the categories.

2

2

Classification

- There are several classification techniques.
- The most widely used are:

1. Logistic regression
2. *k*-nearest neighbors

3

3

Classification problem with 2 categories

- Example of a classification problem:

  customers of a bank are classified on the basis of their solvency (solvent / insolvent)

- For each customer we observe the categorical variable Solvency (Solvent / insolvent) and some predictors:

1. Income
2. Number of members of the household
3. …

- We want to identify a classification rule to link the categorical response variable Solvency to the numerical predictors.

4

4

Categorical response variable

- The categorical response variable is transformed into a binary numerical variable (dummy variable):

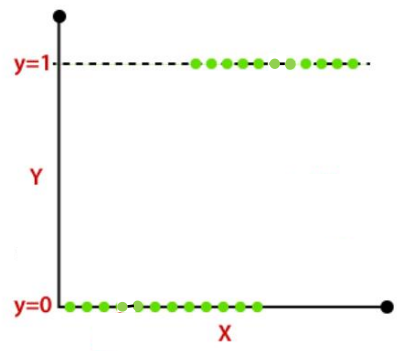
$$Y = \begin{cases} 0 & solvent \\ 1 & insolvent \end{cases}$$

5

5

Scatterplot

- Suppose to have 1 predictor.
- Then we can draw a scatterplot with the predictor X on the x-axis and Y on the y-axis.



6

Simple regression (not a good idea)

- If we applied the simple regression we would get the predicted value of Y,

$$\hat{Y} = b_0 + b_1 X$$

that is the probability for a customer to be insolvent,

$$\hat{Y} = P(Y=1 \mid X)$$

In this example

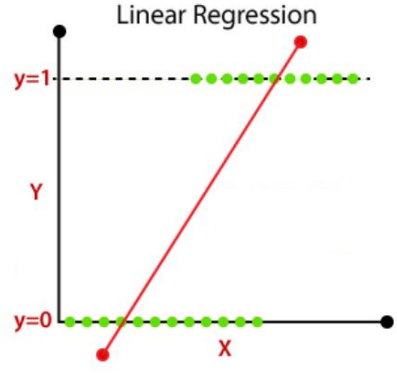
$$\hat{Y} = P(Y=1 \mid X) = P(Insolvent \mid X) = b_0 + b_1 X$$

7

7

P(Insolvent | X) = $b_0 + b_1 X$?

- The use of linear regression does not ensure that the estimated probability is between 0 and 1!
- But probability is always a number between 0 and 1.



8

Logistic regression

- To avoid this problem, we must use a (non-linear) function that provides an output in the interval [0,1].
- Many functions satisfy this requirement.
- Logistic regression uses the logistic function.

9

9

Logistic regression

- In logistic regression

$$\hat{Y} = P(insolvent \mid X) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

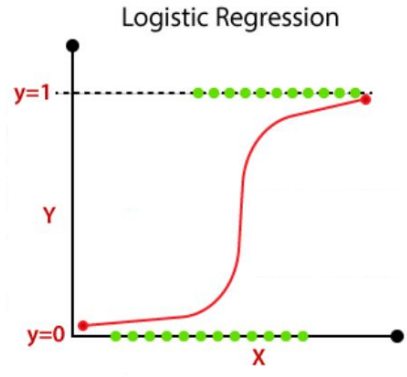
$$0 < \mathrm{P}(insolvent \mid X) < 1$$

10

10

Logistic regression

- Logistic curve provides an S-shaped curve. The estimated probability is between 0 and 1!



11

11

Logistic regression

- If $b_1 > 0$, $X \uparrow \implies \hat{Y} = \mathrm{P}(Y = 1) \uparrow$
- However, in logistic regression, b_1 is <u>not</u> the expected change of the probability when X increases by one unit.
- This expeced change does depend on the value of X!

- If $b_1 < 0$, $X \uparrow \implies \hat{Y} = \mathrm{P}(Y = 1) \downarrow$

12

12

Multiple logistic regression

- If we have more predictors, X_1, X_2, …, X_k

$$\Pr(insolvent \mid X) = \frac{e^{b_0+b_1X+\cdots+b_kX_k}}{1+e^{b_0+b_1X+\cdots+b_kX_k}}$$

13

13

Logistic regression with more than 2 categories

- The categorical response variable can have more than 2 categories.

14

14