Machine Learning

Supervised

Regression

dependent variable is discrete, (eg. 15000)

classification

dependent variable is categorical (eg. true, false)

binary

output column has only two categories
↳ 1 or 0, true or false

multinomial

output column has multiple categories
↳

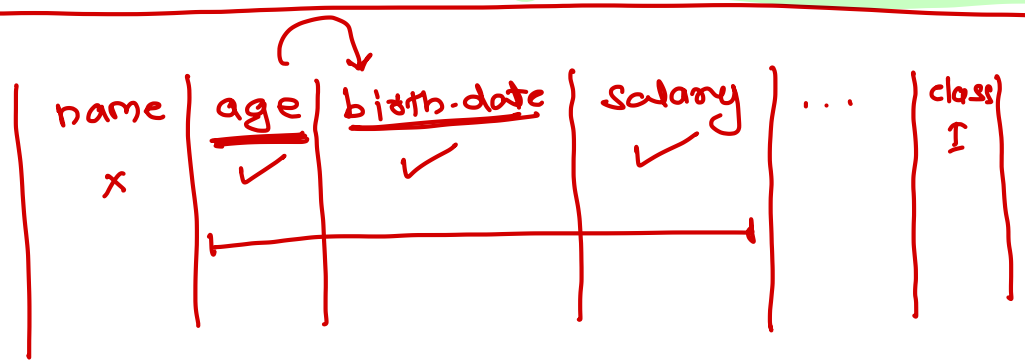
Logistic Regression

Overview

- Logistic Regression was used in the biological sciences in early twentieth century
- It was then used in many social science applications
- Logistic Regression is used when the dependent variable(target) is categorical
- The dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.)
- Unlike linear regression, logistic regression can directly predict probabilities (values that are restricted to the (0,1) interval)
- Furthermore, those probabilities are well-calibrated when compared to the probabilities predicted by some other classifiers
- E.g.
  - To predict whether an email is spam (1) or (0)
  - Whether the tumor is malignant (1) or not (0)

Assumptions

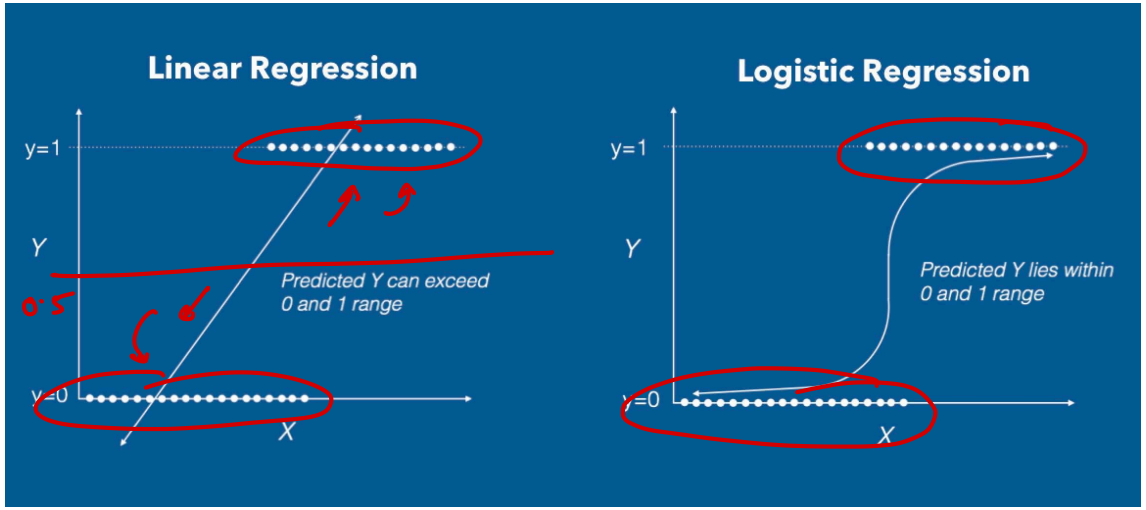
- Binary logistic regression requires the dependent variable to be binary
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome
- Only the meaningful variables should be included [feature engineering]
- The independent variables should be independent of each other. That is, the model should have little or no multi-collinearity
- The independent variables are linearly related to the log odds
- Logistic regression requires quite large sample sizes



Linear vs Logistic

- Target variable is an interval variable
- Predicted values are the mean of the target variable at the given values of the input variable

- Target variable is a discrete (binary or ordinal) variable
- Predicted values are the probability of a particular level(s) of the target variable at the given values of the input variables



How does it work ?

- Logistics regression analysis starts with calculating the “Odds” of the dependent variable
- It is the ratio of the probability that an individual is a member of a particular group or category, p(y) divided by the probability that an individual is not a member of the group or category [1– p(y)]
- It is represented as:

$$Odds = \frac{p(y)}{[1 - p(y)]}$$

- In order to establish a linear relationship between the odds and the independent variables in the logistic regression model, the odds need to be transformed to logit (log-odds) by taking the natural logarithm (ln) of odds
- The logarithmic transformation creates a continuous dependent variable of the categorical dependent variable

Types

- **Binary Logistic Regression**
  - Most useful when you want to model the event probability for a categorical response variable with two outcomes
  - For example, its often used in credit analysis in determining the risk whether the next customer is likely to default — or not default — on a loan

- **Multinomial Logistic Regression**
  - Used to classify subjects into groups based on a categorical range of variables to predict behaviour
  - For example, a survey can be conducted to aid advertising strategy where participants are asked to select one of several competing products as their favorite

Advantages

- Widely used technique due to its simplicity, efficiency, easy interpretation, and usage of limited computational resources.
- Allows easy regularization of outputs to prevent overfitting, yielding probabilities as prediction results.
- Logistic Regression allows easy model updating using stochastic gradient descent.
- Logistic Regression models does not get effected to predict output probabilities on removal of variables uncorrelated to the output or multi-collinear variables

Disadvantages

- Logistic regression's greatest disadvantage is fails to solve non-linear problems and it underperforms when there are multiple or non-linear decision boundaries. It fails to capture more complex relationships.
- Another important requirement for Logistic Regression to function properly is Feature Engineering as it helps to identify independent variables. Without proper identification of independent variables Logistic Regression fails to perform correctly.
- Logistic Regression can only predict a categorical outcome with discrete probability outcome

Applications of Logistic Regression

- Logistic regression has been widely used in medical research, in the field of predictive food microbiology, to describe bacterial growth/no growth interface, in the 0–1 format.
- Logistic regression may be used to predict the risk of developing a given disease based on observed characteristics of the patient.
- Logistic regression is widely used in credit risk modelling in identifying potential loan defaulters.
- Logistic regression is used in different predictive models which finds its usage in response modelling problems like Normal, Poisson, binomial responses and other distributions, even including hypothesis testing.

X_test

age	physical_score
83	40
50	37
52	24
58	31

y_test

test_result
1
1
0
0

y_prediction

test_result	
1	✓
0	✗
0	✓
1	✗

observed predicted

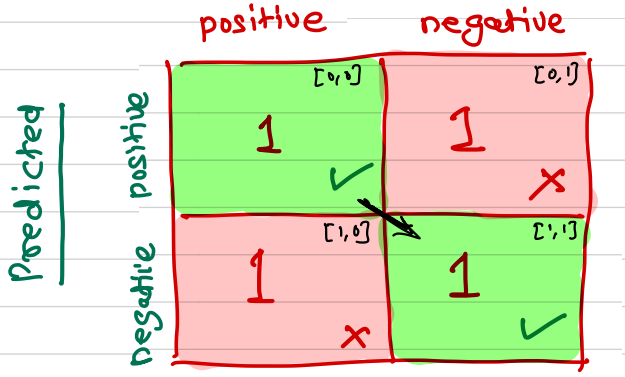
total = 4

correct = 2

wrong = 2

accuracy = $\frac{\text{correct}}{\text{total}} = \frac{2}{4} = 50\%$.

confusion matrix



correct = cm[0,0] + cm[1,1] = 2

wrong = cm[0,1] + cm[1,0] = 2

total = cm[0,0] + cm[0,1] + cm[1,0] + cm[1,1] = 4

accuracy = 2/4 = 50%.