Practical Session 2

Logistic Regression

Alex Ferrer Ferre
Antonio Darder Bennassar

July 2022

CIMNE
International Centre
for Numerical Methods in Engineering

EXCELENCIA
SEVERO
OCHOA

1 Introduction

This session will be focused on the classification task. Similarly to the previous session a logisitc regression will be coded and tested with a real dataset.

The logistic regression uses the same linear transformation as the linear regression

$$h = X \cdot \theta \tag{1}$$

But now we will add another function afterwards. The logistic function which will transform the results into a (0,1) interval that represents the probability of the data belonging to a group.

$$\sigma = \hat{y} = \frac{1}{1 + e^{-h}} \tag{2}$$

In this case θ is a matrix which has as many rows as x has columns and has the same number of columns as groups in the dataset. The derivative of the sigmoid can be writen in terms of itself

$$\frac{\partial \sigma}{\partial h} = \sigma(1 - \sigma) \tag{3}$$

The function that models the system is obtained concatenation of equations (1) and (2). Regarding the metrics, in classification tasks the most common equation is the negative log-likelihood or sometimes called cross entropy. This function takes the likelihood probability and it applies a logarithm

$$J = \frac{1}{m} \sum_{i}^{m} \sum_{j}^{n} y_i^j \cdot ln(\hat{y}_i^j) + (1 - y_i^j) \cdot ln(1 - \hat{y}_i^j) \tag{4}$$

$$\nabla J_{\hat{y}} = \frac{1}{m} \sum_{i}^{m} \frac{\hat{y}_i - y_i}{\hat{y}_i(1 - \hat{y}_i)} \tag{5}$$

Combining equations (3), (5) and the derivative of h respect the parameters which is equal to X we obtain the gradient of the whole model:

$$\nabla J_{\theta} = \frac{1}{m} \sum_{i}^{m} X_i^T \times (\hat{y}_i - y_i) \equiv \frac{1}{m} X^T \times (\hat{y} - y) \tag{6}$$

Unfortunately the addition of the sigmoid function transforms the problem in such a way that it is no longer analytically solvable. The solutions will be computed numerically using for example the minimize function from the scipy library or using gradient descent methods.

1.1 The datasets

In this session we will use a real-world dataset. This contains around 120 data points of two different tests performed to a set of microchips. After completing the tests the microchips are labeled as "accepted" or "rejected".

The boundary of the groups kind of follows an elliptical shape, however there is some noise which will make a polynomial with degree 2 not fit perfectly. This fact will help to analyze the overfitting.

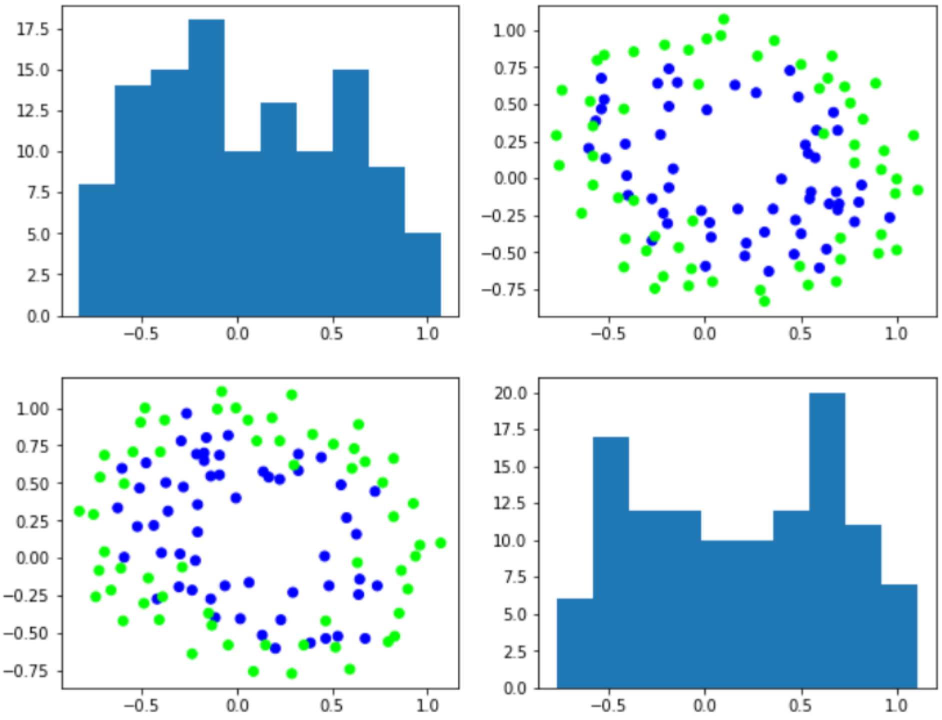

Figure 1 Microchip dataset