

FYS-STK3155/4155 Applied Data Analysis and Machine Learning - Project 2: Classification and Regression

Lotsberg, Bernhard Nornes
Nguyen, Anh-Nguyet Lise

<https://github.com/liseanh/FYS-STK4155-project2/>

October - November 2019

Abstract

blip blop opnion wrong

1 Introduction

fill

2 Theory

To evaluate the performance of our classification model, we consider the accuracy score, given by

$$\frac{\sum_{i=1}^n I(t_i = y_i)}{n}, \quad (1)$$

where t_i is the target, y_i is our model output and I is the indicator function,

$$I = \begin{cases} 1 & t_i = y_i \\ 0 & t_i \neq y_i \end{cases}.$$

2.1 Stochastic Gradient Descent (SGD)

2.2 Logistic Regression (LR)

2.3 Artificial Neural Networks (ANN)

2.3.1 Multilayer perceptron

The multilayer perceptron is a feedforward neural network. To calculate the optimal bi-

ases and weights for the problem, we initialize the gradients of the cost function \mathcal{C} with respect to the weights W and biases b at the output layer $l = L$ and the output error δ_L as

$$\frac{\partial \mathcal{C}}{\partial w_{jk}^L} = \delta_j^L a_k^{L-1}, \quad (2)$$

$$\frac{\partial \mathcal{C}}{\partial b_j^L} = \delta_j^L, \quad (3)$$

$$\delta_j^L = f'(z_j^L) \frac{\partial \mathcal{C}}{\partial a_j^L}, \quad (4)$$

before propagating backwards through the hidden layers using the general equations

$$\frac{\partial \mathcal{C}}{\partial w_{jk}^l} = \delta_j^l a_k^{l-1}, \quad (5)$$

$$\frac{\partial \mathcal{C}}{\partial b_j^l} = \delta_j^l, \quad (6)$$

$$\delta_j^l = \sum_k \delta_k^{l+1} w_{kj}^{l+1} f'(z_j^l). \quad (7)$$

3 Data

In this paper we are using credit card payment data from a Taiwanese bank downloaded from the UCI Machine Learning

Repository. The response variable is a binary variable of default payment with Yes = 1, No = 0. The data set consists of 30 000 observations, with X amount of observations with default payments. There are 23 explanatory variables, cited from the original paper they are described as:

- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . . ; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . . ; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . . ; X17 = amount of bill statement in April, 2005.
- X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . . ; X23 = amount paid in April, 2005.

4 Method

fill

5 Results

6 Discussion

fill

7 Conclusion

fill

References