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

The aim of this project is to implement and use logistic regression and a multilayer perceptron (MLP) to classify data **PLEASE SPECIFY WHAT KIND OF DATA** and to further use the MLP to perform regression analysis on Franke's function.

2 Theory

2.1 Stochastic Gradient Descent (SGD)

2.2 Logistic Regression (LR)

2.3 Artificial Neural Networks (ANN)

In an artificial neural network, something something nodes. The output of the nodes in each layer is given by the value of a chosen activation function $f(z)$.

2.3.1 Multilayer perceptron

The multilayer perceptron is a feedforward neural network.

The activation of the jth neuron of layer l is defined as

$$z_j^l = \sum_{i=1}^{M_{l-1}} w_{ij}^l a_j^{l-1} + b_j^l, \tag{1}$$

where b_j^l and w_{ij}^l are the biases and weights at layer l, and $a_j^l = f(z_j^l)$.

To calculate the optimal biases 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

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

$$\frac{\partial \mathcal{C}}{\partial b_j^L} = \delta_j^L, \tag{3}$$

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

before propagating backwards through the hidden layers using the general equations

$$\frac{\partial \mathcal{C}}{\partial w^l_{jk}} = \delta^l_j a^{l-1}_k, \tag{5}$$

$$\frac{\partial \mathcal{C}}{\partial b^l_j} = \delta^l_j, \tag{6}$$

$$\delta^l_j = \sum_k \delta^{l+1}_k w^{l+1}_{kj} f'(z^l_j). \tag{7}$$

Looking at these equations, it is clear that the chosen cost function $\mathcal{C}$ should be differentiable.

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 original 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 [1]:

- 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 Model evaluation

4.1 Regression

To evaluate the performance of our regression model, we consider the R^2 score, given by

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}, \tag{8}$$

where $\boldsymbol{y}$ is the given data, $\hat{\boldsymbol{y}}$ is the model and $\bar{y}$ is the mean value of $\boldsymbol{y}$.

4.2 Classification

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

$$\text{accuracy} = \frac{\sum_{i=1}^{n} I(t_i = y_i)}{n}, \tag{9}$$

where t_i is the target, y_i is the model output, n is the number of samples and I is the indicator function,

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

5 Method

6 Results

7 Discussion

fill

8 Conclusion

fill

References

[1] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.