

Using Neural Networks and Logistic Regression for Classification and Regression Problems

Eina B. Jørgensen, Anna Lina P. Sjur and Jan-Adrian H. Kallmyr

November 11, 2019

Abstract

1 Introduction

Artificial Neural Networks (ANNs) offer a flexible way of performing data analysis, and can be used for tasks such as classification and regression.

In this article, we will study the 2005 Taiwan credit card data set with the purpose of classifying risky/non-risky credit card holders. For this we will use a ANN as well as the LR method. Furthermore, we will try to model the Franke function with randomised noise using a ANN.

2 Theory and methods

2.1 Logistic Regression

Classification problems aim to predict the behaviour of a given object, and look for patterns based on discrete variables (i.e categories). Logistic regression can be used to solve such problems, commonly by the use of variables with binary outcomes such as true/false, positive/negative, success/failure etc., or in the specific credit card case: *risky/non-risky* ?

As opposed to linear regression, the equation one gets as a result of minimization of the cost function by $\hat{\beta}$ using logistic regression, is non-linear, and is solved using minimization

algorithms called *gradient descent methods*.

When predicting the the output classes in which an object belongs, the prediction is based on the design matrix $\hat{\mathbf{X}} \in \mathbb{R}^{n \times p}$ that contain n samples that each carry p features.

A distinction is made between *hard classification* - deterministically determine the variable to a category, and *soft classification* - determines the probability that a given variable belongs in a certain category. The latter is favorable in many cases, and logistic regression is the most used example of this type of classifier.

When using logistic regression, the probability that a given data point x_i belongs in a category y_i is given by the Sigmoid-function (or logistic function):

$$p(t) = \frac{1}{1 + e^{-t}} = \frac{e^t}{1 + e^t} \quad (1)$$
$$1 - p(t) = p(-t)$$

Assuming a binary classification problem, i.e. y_i can be either 0 or 1, and a set of predictors $\hat{\beta}$ the Sigmoid function (1) gives the probabilities with relation:

$$p(y_i = 0 | x_i, \hat{\beta}) = 1 - p(y_i = 1 | x_i, \hat{\beta})$$

The total likelihood for all possible outcomes $\mathcal{D} = \{(y_i, x_i)\}$ is used in the Maximum Like-

likelihood Estimation (MLE), aiming at maximizing the log/likelihood function (2). The likelihood function can be expressed with \mathcal{D} :

$$P(\mathcal{D}|\hat{\beta}) = \prod_{i=1}^n \left[p(y_i = 1|x_i, \hat{\beta}) \right]^{y_i} \left[1 - p(y_i = 0|x_i, \hat{\beta}) \right]^{1-y_i}$$

And the log/likelihood function is then:

$$P_{\log}(\hat{\beta}) = \sum_{i=1}^n \left(y_i \log \left[p(y_i = 1|x_i, \hat{\beta}) \right] + (1 - y_i) \log \left[1 - p(y_i = 0|x_i, \hat{\beta}) \right] \right) \quad (2)$$

The cost/error-function \mathcal{C} (also called cross-entropy in statistics) is the negative of the log/likelihood. Maximizing P_{\log} is thus the same as minimizing the cost function. The cost function is:

$$\mathcal{C}(\hat{\beta}) = -P_{\log}(\hat{\beta}) = -\sum_{i=1}^n \left(y_i \log \left[p(y_i = 1|x_i, \hat{\beta}) \right] + (1 - y_i) \log \left[1 - p(y_i = 0|x_i, \hat{\beta}) \right] \right) \quad (3)$$

Finding the parameters $\hat{\beta}$ that minimize the cost function is then done through derivation. Defining the vector \hat{y} containing n elements y_i , the $n \times p$ matrix \hat{X} containing the x_i elements, and the vector \hat{p} that is the fitted probabilities $p(y_i|x_i, \hat{\beta})$, the first derivative of \mathcal{C} is

$$\nabla_{\beta} \mathcal{C} = \frac{\partial \mathcal{C}(\hat{\beta})}{\partial \hat{\beta}} = -\hat{X}^T (\hat{y} - \hat{p}) \quad (4)$$

This gives rise to set of linear equations, where the aim is to solve the system for $\hat{\beta}$.

With $\hat{x} = [1, x_1, x_2, \dots, x_p]$ and p predictors $\hat{\beta} = [\beta_0, \beta_1, \beta_2, \dots, \beta_p]$ the relation between likelihoods of outcome is:

$$\log \frac{p(\hat{\beta}\hat{x})}{1 - p(\hat{\beta}\hat{x})} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (5)$$

and $p(\hat{\beta}\hat{x})$ defined by:

$$p(\hat{\beta}\hat{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \quad (6)$$

2.2 Gradient Descent Methods

The General Idea

With the gradient of \mathcal{C} defined as in (4), we use this to find the minimum of the cost function. The basic idea is that by moving in the direction of the negative gradient of a function, we can move towards the value (in this case the β) that minimizes the function (in this case $\mathcal{C}(\beta)$). ?

This is done by repeating the algorithm

$$\beta_{j+1} = \beta_j - \gamma \nabla_{\beta} \mathcal{C}(\beta) \quad j = 0, 1, 2, \dots \quad (7)$$

When a minimum is approached, $\nabla_{\beta} \mathcal{C}(\beta) \rightarrow 0$, and thus we can set a limit when $\beta_{k+1} \approx \beta_k$ given a certain tolerance, and the β which minimizes the cost function is found. γ is in this case called the *learning rate*, and is a parameter that must be tuned to each specific case in order to optimize the regression.

Stochastic Gradient Descent

In this project we use a stochastic version of gradient descent, which is an improvement upon the regular gradient descent. This is done by expressing the cost function (and thus also its gradient) as a sum

$$\nabla_{\beta} \mathcal{C}(\beta) = \sum_i^n \nabla_{\beta} c_i(\mathbf{x}_i, \beta), \quad (8)$$

and by only taking calculating the gradient of a subset of the data at the time. These subsets, called *minibatches* are of size \mathbf{M} , and the total amount is $\frac{n}{\mathbf{M}}$ where n is the amount of data points. The minibatches are denoted \mathbf{B}_k , with $k = 1, 2, \dots, \frac{n}{\mathbf{M}}$.

Instead of a sum over all the data points $i \in [1, n]$ we now in each step, sum over all

the data points in the given minibatch $i \in \mathbf{B}_k$ where k is picked randomly with uniform probability from $[1, \frac{n}{M}]$.

The stochastic and final version of (7) is therefore given by the algorithm

$$\beta_{j+1} = \beta_j - \gamma_j \sum_{i \in \mathbf{B}_k} \nabla_{\beta} c_i(\mathbf{x}_i, \beta) \quad (9)$$

An iteration over the total number of minibatches is commonly referred to as an *epoch*.

By using the stochastic gradient descent method (9) to minimize the cost function (3) we can find the β values that give the most accurate classification, by doing *logistic regression*.

2.3 Neural networks

In this section, the equations used are based off the book by ?, unless otherwise specified.

The structure of a network

Neural networks, as the name suggests, are inspired by our understanding of how networks of neurons function in the brain. As can be seen in the example network in Figure 1, neurons are structured in layers. We always have an input and an output layer, in addition to a varying number of hidden layers. The input layer has as many neurons as there are input variables, while the output layer has one neuron for each output. How many neurons you have in the output layer depends on the specific problem. The number of neurons in each hidden layer, on the other hand, is not directly related to inputs or outputs, and must be decided in some other way.

As the diagram in Figure 1 suggests, the neurons in each layer are not connected with each other, but take in inputs from the previous layer and pass on an output to the neurons in the next layer, as illustrated with arrows.

This way, the inputs are fed through the network and processed, resulting in an output.

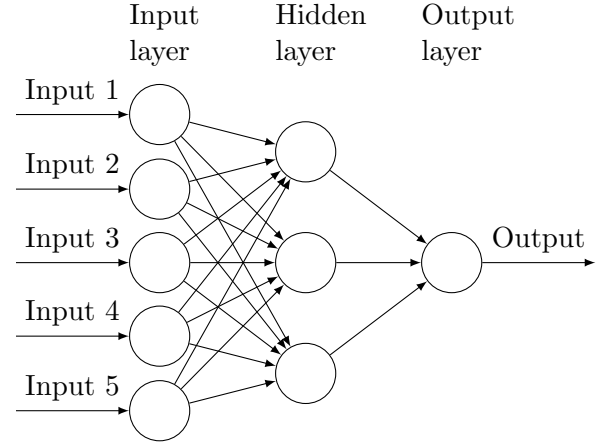


Figure 1: Schematic diagram of a neural network with five input neurons in the input layer, one hidden layer with three neurons and a single output neuron in the output layer.

Forward feeding

Each neuron has one or multiple inputs, as illustrated with arrows in Figure 1. Each of these inputs has a weight associated with it. To clarify the notation used, let's take a look at the j th neuron in the l th layer. The weight associated with the input coming from the k th neuron in the previous layer is denoted as w_{jk}^l . In addition, each neuron has a bias associated with it, for the neuron in question denoted as b_j^l . Summing the weighted inputs and the bias, and feeding this to a function f , gives the activation a_j^l :

$$a_j^l = f \left(\left(\sum_k w_{jk}^l a_k^{l-1} \right) + b_j^l \right)$$

This activation is then fed forward as input to all the neurons in the next layer.

In matrix notation, the activation for the whole layer l can be written as

$$\mathbf{a}^l = f(\mathbf{w}^l \mathbf{a}^{l-1} + \mathbf{b}^l) \quad (10)$$

Here, \mathbf{a}^l and \mathbf{b}^l are vertical vectors containing the activations and biases of the l th layer, while \mathbf{w}^l is a matrix with elements w_{jk}^l , i.e. the j th column contains the weights of the inputs reaching the j th neuron.

Let's look at the activation function in Eq. (??) denoted with a f . In the case of classification, the sigmoid function stated in Eq. (1) is often used in introductory texts. As we will see in the backpropagation algorithm, the sigmoid is a good choice for activation function, since a small change in the output can be propagated backwards, resulting in small changes in the weights and biases through the network.

Another activation function is the so-called rectified linear unit (ReLU) function, given as $f(z) = \max(0, z)$. This gives an activation function which only fires when z is positive. In the case of regression, a variant of ReLU, called leaky ReLU, was used in this project. This function is given as

$$f(z) = \begin{cases} z & z \geq 0 \\ 0.01z & z < 0 \end{cases} \quad (11)$$

This form of ReLU has a small positive gradient for negative values, which allows for gradient based learning even when the z -value is negative (?).

With a basis in Eq. (??), the algorithm for forward feeding is given in Algorithm ???. Here L is the total number of layers.

Note that the output will have values between 0 and 1, when the sigmoid function is used to compute the activations of all the layers. In a classification problem, this corresponds to the likelihood of an outcome. For example, in a classification problem with five classes, the network would have five output

```
Set  $\mathbf{a}^1 = \text{input}$ ;
foreach  $l=2:L$  do
    Compute  $\mathbf{z} = \mathbf{w}^l \mathbf{a}^{l-1} + \mathbf{b}^l$ ;
    Compute  $\mathbf{a}^l = f(\mathbf{z})$ ;
Set output to  $\mathbf{a}^L$ ;
```

Algorithm 1: The forward feeding algorithm.

neurons, each representing a class. The final classification of an input would then be the class with the highest probability. In the regression case, we see that when using the leaky ReLU activation function from Eq. (??), the output can be any positive value (strictly also negative values, due to the leakage, but this would call for large negative z -values).

Backpropagation

When training the network, the goal is to find the weights and biases that minimize the cost function C . In this project, the cross-entropy cost function was used for classification. For a single neuron, this is given as

$$C = - \sum_{i=1}^n [y_i \ln a_i^L + (1 - y_i) \ln(1 - a_i^L)] \quad (12)$$

Here, we are summing over n points of training data, where a_i^L is the output with the corresponding correct value y_i .

In the regression case, the quadratic cost function, given as

$$C = \frac{1}{2} \sum_{i=1}^n \|y_i - a_i^L\|^2 \quad (13)$$

is used.

To find the weights and biases that minimize Eq.(??) and (??), one can use Stochastic Gradient Decent, as described previously. But in order to use SGD, the derivatives of C with

respect to all the weights and biases must be computed, and it is here that backpropagation comes in. It can be shown that the derivatives are given as in Eq. (??).

$$\begin{aligned}\delta^L &= \nabla_a C \odot f'(z^L) \\ \delta^l &= ((\mathbf{w}^{l+1})^T \delta^{l+1}) \odot f'(\mathbf{z}^l) \\ \frac{\partial C}{\partial b_j^l} &= \delta_j^l \\ \frac{\partial C}{\partial w_{jk}^l} &= a_k^{l-1} \delta_j^l\end{aligned}\tag{14}$$

Taking a look at the expression for δ^L , one can show that this reduces to $\delta^L = \mathbf{a} - \mathbf{y}$, when the cross-entropy cost function is combined with the sigmoid activation function, and similarly for the quadratic cost and ReLU when we have positive output values.

Equation (??) are the basis for the backpropagating algorithm, described in Algorithm ??.

```

Compute  $\{\mathbf{a}^l\}_{l=1}^L$  with feed forward;
Compute  $\delta^L$ ;
Set  $\frac{\partial C}{\partial \mathbf{b}^L} = \delta^L$ ;
Compute  $\frac{\partial C}{\partial \mathbf{w}^L} = \delta^L (\mathbf{a}^{L-1})^T$ ;
foreach  $l=L-1:2$  do
    Compute  $\delta^l$ ;
    Set  $\frac{\partial C}{\partial \mathbf{b}^l} = \delta^l$ ;
    Compute  $\frac{\partial C}{\partial \mathbf{w}^l} = \delta^l (\mathbf{a}^{l-1})^T$ ;

```

Algorithm 2: The backpropagation algorithm.

2.4 Data sets

In this report, the default of credit card clients data set (?) was used to study the performance and compare the logistic regression method and neural networks. A description of the attributes of the data set can be found in ?. Upon inspection, it is notable that the data set

contains a considerable amount of values different from their valid values as described by ?. Mostly, this applies for the categorical variables, where the given value does not correspond to any category. By removing data points with invalid values, as well as entries where the client does not have any history of past payments or bill statements, the data set is reduced from 30000 data points to 3792 points. As this is a considerable reduction of data points, a second reduced data set was constructed, where (OGSÅ HVA VI GJORDE — HER!)

In addition the default of credit card clients data set, data produced with Franke’s function was applied to neural networks. For details on Franke’s function, see ?.

Reduction of Dimension

(HER!)

2.5 Quality of Measurements

(HER!)

3 Results

3.0.1 Tuning Learning Rate and Mini-batch Size

Logistic Regression

Optimal Parameters and Associated Results

3.1 Verification by Comparison to Scikit-Learn

3.2 Neural Networks for Regression on Franke’s Function

4 Discussion

5 Conclusion

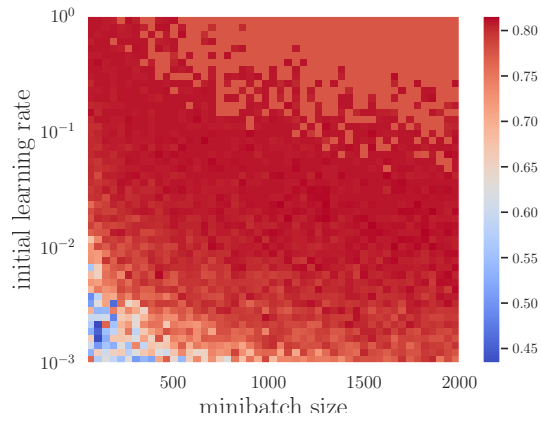


Figure 2: Heatmap showing the accuracy of the Logistic Regression for different values of the minibatch sizes and initial learning rates.

Table 1: Fraction of true and false negatives and positives for Neural network (NN), Logistic regression (LR) and SciKitLearn’s neural network (SKL)

	LR	NN	SKL
Positive	True		
	False		
Negative	True		
	False		

Figure 3