

# Data Analysis and Machine Learning: Logistic Regression

Morten Hjorth-Jensen<sup>1,2</sup>

<sup>1</sup>Department of Physics, University of Oslo

<sup>2</sup>Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

Oct 17, 2019

## Logistic Regression

In linear regression our main interest was centered on learning the coefficients of a functional fit (say a polynomial) in order to be able to predict the response of a continuous variable on some unseen data. The fit to the continuous variable  $y_i$  is based on some independent variables  $\hat{x}_i$ . Linear regression resulted in analytical expressions for standard ordinary Least Squares or Ridge regression (in terms of matrices to invert) for several quantities, ranging from the variance and thereby the confidence intervals of the parameters  $\hat{\beta}$  to the mean squared error. If we can invert the product of the design matrices, linear regression gives then a simple recipe for fitting our data.

Classification problems, however, are concerned with outcomes taking the form of discrete variables (i.e. categories). We may for example, on the basis of DNA sequencing for a number of patients, like to find out which mutations are important for a certain disease; or based on scans of various patients' brains, figure out if there is a tumor or not; or given a specific physical system, we'd like to identify its state, say whether it is an ordered or disordered system (typical situation in solid state physics); or classify the status of a patient, whether she/he has a stroke or not and many other similar situations.

The most common situation we encounter when we apply logistic regression is that of two possible outcomes, normally denoted as a binary outcome, true or false, positive or negative, success or failure etc.

## Optimization and Deep learning

Logistic regression will also serve as our stepping stone towards neural network algorithms and supervised deep learning. For logistic learning, the minimization of the cost function leads to a non-linear equation in the parameters  $\hat{\beta}$ . The optimization of the problem calls therefore for minimization algorithms. This

forms the bottle neck of all machine learning algorithms, namely how to find reliable minima of a multi-variable function. This leads us to the family of gradient descent methods. The latter are the working horses of basically all modern machine learning algorithms.

We note also that many of the topics discussed here on logistic regression are also commonly used in modern supervised Deep Learning models, as we will see later.

## Basics

We consider the case where the dependent variables, also called the responses or the outcomes,  $y_i$  are discrete and only take values from  $k = 0, \dots, K - 1$  (i.e.  $K$  classes).

The goal is to predict the output classes from the design matrix  $\hat{X} \in \mathbb{R}^{n \times p}$  made of  $n$  samples, each of which carries  $p$  features or predictors. The primary goal is to identify the classes to which new unseen samples belong.

Let us specialize to the case of two classes only, with outputs  $y_i = 0$  and  $y_i = 1$ . Our outcomes could represent the status of a credit card user that could default or not on her/his credit card debt. That is

$$y_i = \begin{bmatrix} 0 & \text{no} \\ 1 & \text{yes} \end{bmatrix}.$$

## Linear classifier

Before moving to the logistic model, let us try to use our linear regression model to classify these two outcomes. We could for example fit a linear model to the default case if  $y_i > 0.5$  and the no default case  $y_i \leq 0.5$ .

We would then have our weighted linear combination, namely

$$\hat{y} = \hat{X}^T \hat{\beta} + \hat{\epsilon}, \quad (1)$$

where  $\hat{y}$  is a vector representing the possible outcomes,  $\hat{X}$  is our  $n \times p$  design matrix and  $\hat{\beta}$  represents our estimators/predictors.

## Some selected properties

The main problem with our function is that it takes values on the entire real axis. In the case of logistic regression, however, the labels  $y_i$  are discrete variables. A typical example is the credit card data discussed below here, where we can set the state of defaulting the debt to  $y_i = 1$  and not to  $y_i = 0$  for one the persons in the data set (see the full example below).

One simple way to get a discrete output is to have sign functions that map the output of a linear regressor to values  $\{0, 1\}$ ,  $f(s_i) = \text{sign}(s_i) = 1$  if  $s_i \geq 0$  and 0 if otherwise. We will encounter this model in our first demonstration of neural networks. Historically it is called the “perceptron” model in the machine learning literature. This model is extremely simple. However, in many cases it

is more favorable to use a “soft” classifier that outputs the probability of a given category. This leads us to the logistic function.

## The logistic function

The perceptron is an example of a “hard classification” model. We will encounter this model when we discuss neural networks as well. Each datapoint is deterministically assigned to a category (i.e  $y_i = 0$  or  $y_i = 1$ ). In many cases, it is favorable to have a “soft” classifier that outputs the probability of a given category rather than a single value. For example, given  $x_i$ , the classifier outputs the probability of being in a category  $k$ . Logistic regression is the most common example of a so-called soft classifier. In logistic regression, the probability that a data point  $x_i$  belongs to a category  $y_i = \{0, 1\}$  is given by the so-called logit function (or Sigmoid) which is meant to represent the likelihood for a given event,

$$p(t) = \frac{1}{1 + \exp -t} = \frac{\exp t}{1 + \exp t}.$$

Note that  $1 - p(t) = p(-t)$ .

## Examples of likelihood functions used in logistic regression and neural networks

The following code plots the logistic function, the step function and other functions we will encounter from here and on.

## Two parameters

We assume now that we have two classes with  $y_i$  either 0 or 1. Furthermore we assume also that we have only two parameters  $\beta$  in our fitting of the Sigmoid function, that is we define probabilities

$$\begin{aligned} p(y_i = 1|x_i, \hat{\beta}) &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \\ p(y_i = 0|x_i, \hat{\beta}) &= 1 - p(y_i = 1|x_i, \hat{\beta}), \end{aligned}$$

where  $\hat{\beta}$  are the weights we wish to extract from data, in our case  $\beta_0$  and  $\beta_1$ .

Note that we used

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

## Maximum likelihood

In order to define the total likelihood for all possible outcomes from a dataset  $\mathcal{D} = \{(y_i, x_i)\}$ , with the binary labels  $y_i \in \{0, 1\}$  and where the data points are drawn independently, we use the so-called [Maximum Likelihood Estimation](#) (MLE) principle. We aim thus at maximizing the probability of seeing the

observed data. We can then approximate the likelihood in terms of the product of the individual probabilities of a specific outcome  $y_i$ , that is

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

from which we obtain the log-likelihood and our **cost/loss** function

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

### The cost function rewritten

Reordering the logarithms, we can rewrite the **cost/loss** function as

$$\mathcal{C}(\hat{\beta}) = \sum_{i=1}^n (y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp(\beta_0 + \beta_1 x_i))).$$

The maximum likelihood estimator is defined as the set of parameters that maximize the log-likelihood where we maximize with respect to  $\beta$ . Since the cost (error) function is just the negative log-likelihood, for logistic regression we have that

$$\mathcal{C}(\hat{\beta}) = - \sum_{i=1}^n (y_i(\beta_0 + \beta_1 x_i) - \log(1 + \exp(\beta_0 + \beta_1 x_i))).$$

This equation is known in statistics as the **cross entropy**. Finally, we note that just as in linear regression, in practice we often supplement the cross-entropy with additional regularization terms, usually  $L_1$  and  $L_2$  regularization as we did for Ridge and Lasso regression.

### Minimizing the cross entropy

The cross entropy is a convex function of the weights  $\hat{\beta}$  and, therefore, any local minimizer is a global minimizer.

Minimizing this cost function with respect to the two parameters  $\beta_0$  and  $\beta_1$  we obtain

$$\frac{\partial \mathcal{C}(\hat{\beta})}{\partial \beta_0} = - \sum_{i=1}^n \left( y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right),$$

and

$$\frac{\partial \mathcal{C}(\hat{\beta})}{\partial \beta_1} = - \sum_{i=1}^n \left( y_i x_i - x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right).$$

## A more compact expression

Let us now define a vector  $\hat{y}$  with  $n$  elements  $y_i$ , an  $n \times p$  matrix  $\hat{X}$  which contains the  $x_i$  values and a vector  $\hat{p}$  of fitted probabilities  $p(y_i|x_i, \hat{\beta})$ . We can rewrite in a more compact form the first derivative of cost function as

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

If we in addition define a diagonal matrix  $\hat{W}$  with elements  $p(y_i|x_i, \hat{\beta})(1 - p(y_i|x_i, \hat{\beta}))$ , we can obtain a compact expression of the second derivative as

$$\frac{\partial^2 \mathcal{C}(\hat{\beta})}{\partial \hat{\beta} \partial \hat{\beta}^T} = \hat{X}^T \hat{W} \hat{X}.$$

## Extending to more predictors

Within a binary classification problem, we can easily expand our model to include multiple predictors. Our ratio between likelihoods is then with  $p$  predictors

$$\log \frac{p(\hat{\beta}\hat{x})}{1 - p(\hat{\beta}\hat{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Here we defined  $\hat{x} = [1, x_1, x_2, \dots, x_p]$  and  $\hat{\beta} = [\beta_0, \beta_1, \dots, \beta_p]$  leading to

$$p(\hat{\beta}\hat{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}.$$

## Including more classes

Till now we have mainly focused on two classes, the so-called binary system. Suppose we wish to extend to  $K$  classes. Let us for the sake of simplicity assume we have only two predictors. We have then following model

$$\log \frac{p(C=1|x)}{p(K|x)} = \beta_{10} + \beta_{11} x_1,$$

$$\log \frac{p(C=2|x)}{p(K|x)} = \beta_{20} + \beta_{21} x_1,$$

and so on till the class  $C = K - 1$  class

$$\log \frac{p(C=K-1|x)}{p(K|x)} = \beta_{(K-1)0} + \beta_{(K-1)1} x_1,$$

and the model is specified in term of  $K - 1$  so-called log-odds or **logit** transformations.

## More classes

In our discussion of neural networks we will encounter the above again in terms of a slightly modified function, the so-called **Softmax** function.

The softmax function is used in various multiclass classification methods, such as multinomial logistic regression (also known as softmax regression), multiclass linear discriminant analysis, naive Bayes classifiers, and artificial neural networks. Specifically, in multinomial logistic regression and linear discriminant analysis, the input to the function is the result of  $K$  distinct linear functions, and the predicted probability for the  $k$ -th class given a sample vector  $\hat{x}$  and a weighting vector  $\hat{\beta}$  is (with two predictors):

$$p(C = k|\mathbf{x}) = \frac{\exp(\beta_{k0} + \beta_{k1}x_1)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1}x_1)}.$$

It is easy to extend to more predictors. The final class is

$$p(C = K|\mathbf{x}) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_{l1}x_1)},$$

and they sum to one. Our earlier discussions were all specialized to the case with two classes only. It is easy to see from the above that what we derived earlier is compatible with these equations.

To find the optimal parameters we would typically use a gradient descent method. Newton's method and gradient descent methods are discussed in the material on [optimization methods](#).

## A simple classification problem

### The Credit Card example

Here we use the [credit card data](#). The data are from an extensive database from Taiwan and include more than ten predictors.

For categorical data -Scikit-Learn- provides a so-called **one-hot encoder**. This is called one-hot encoding, because only one attribute will be equal to 1 (hot), while the others will be 0 (cold). **Scikit-Learn** provides a `OneHotEncoder` encoder to convert integer categorical values into one-hot

### How to read the Credit Card data