

Business Analytics & Machine Learning

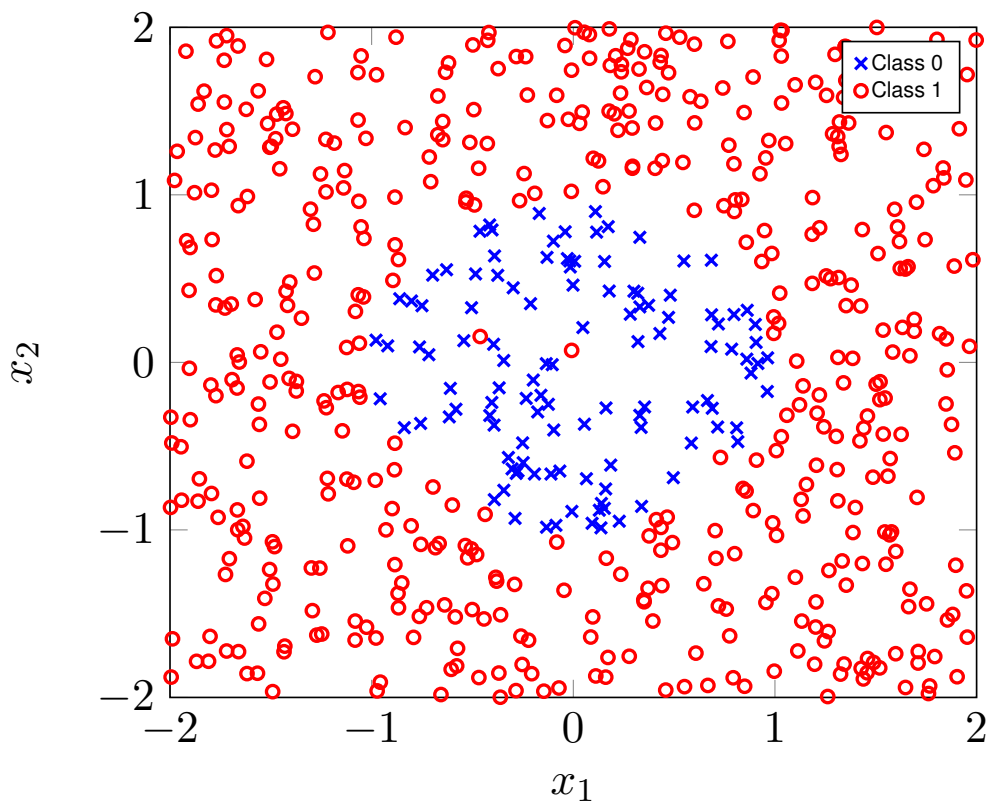
Tutorial sheet 3: Logistic Regression – Solution

Prof. Dr. Martin Bichler

Julius Durmann, Markus Ewert, Yutong Chao, Dr. Mete Ahunbay

Exercise T3.1 *Logistic regression for a 2D classification problem*

You are given the data set in *2d-classification-data.csv* which is also visualized below. The data consists of two-dimensional points specified by coordinates x_1 and x_2 , belonging to one of two classes: 0 or 1. For convenience, the data set also includes square coordinates, $z_1 = x_1^2$ and $z_2 = x_2^2$.



In this exercise, we are going to find a predicting model which can classify new data points. Consider the logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\iff$$

$$p(y=1|x_1, x_2) = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

- a) For which values $x = (x_1, x_2)$ does the logistic regression model output $p(y=1|x) = p(y=0|x)$?
Derive a functional description $x_2 = f(x_1)$ which describes the corresponding decision boundary.

- b) The model in a) is obviously not appropriate for the data at hand. Looking at the plot, convince yourself that aside from some very random noise, class 0 data points are those that lie within some elliptic disc¹. Come up with an appropriate logistic regression model, using the fact that the set of points of a standard ellipse satisfies $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} \leq r^2$.
Hint: The simplest such model will have three coefficients.
- c) Using Python and scikit-learn (or statsmodels), derive the optimal parameters for the logistic regression model. If you like, you may use the provided template notebook for this purpose.
- d) You are now told that the actual data is class 1 with certainty outside the circular disc of radius 1, and class 1 with probability 2% if it is inside it. Draw the decision boundary of the optimal model and observe the classification of points about it. Can you explain why the model might perform worse (with respect to the number of misclassified samples) than the actual model you could have even come up with by yourself in b)?

Solution

- a) By inserting the definition of $p(y = 1|x)$ and the fact that $p(y = 1|x) + p(y = 0|x) = 1$, we find:

$$\begin{aligned}
 & p(y = 1|x) = p(y = 0|x) \\
 \Leftrightarrow & p(y = 1|x) = 1 - p(y = 1|x) \\
 \Leftrightarrow & p(y = 1|x) = \frac{1}{2} \\
 \Leftrightarrow & \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = \frac{1}{2} \\
 \Leftrightarrow & \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}} = \frac{1}{2} \\
 \Leftrightarrow & e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} = \frac{1}{2} + \frac{1}{2} e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \\
 \Leftrightarrow & e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} = 1 \\
 \Leftrightarrow & \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \ln(1) = 0
 \end{aligned}$$

From the last equation, we can derive:

$$x_2 = f(x_1) = -\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1$$

We observe that this is a linear function, represented by a straight.

This straight describes the decision boundary:

$$\begin{aligned}
 x_2 < f(x_1) & \implies p(y = 1|x) > p(y = 0|x) \\
 x_2 > f(x_1) & \implies p(y = 1|x) < p(y = 0|x)
 \end{aligned}$$

(*Note:* The direction of the inequalities depend on the sign of β_2 .)

- b) We see that, for some coefficients a, b, r , we want

$$\log\left(\frac{p}{1-p}\right) \leq 1/2 \Leftrightarrow \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} \leq r^2$$

¹You might even deduce that it's a circle!

i.e. the decision boundary should be given by an ellipse. The inequality on the right is equivalent to

$$\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} - r^2 \leq 0.$$

Write $\beta_0 = -r^2$, $\beta_1 = 1/a^2$ and $\beta_2 = 1/b^2$, and we have our logistic model,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2.$$

c) Please have a look at the provided Jupyter Notebook for the solution.

The optimal parameter values are:

β_0	-10.286
β_1	10.639
β_2	10.476

Note: If you use scikit-learn's `LogisticRegression` and don't specify that there is no penalty, you might end up with the following parameters:

β_0	-10.3198
β_1	10.6802
β_2	10.5038

They are the optimal (minimizing) values for the *regularized* loss

$$\hat{\mathcal{L}}(\beta) = -LL(\beta) + \frac{1}{2}\|\beta\|_2^2$$

where $LL(\beta)$ is the log-likelihood.

d) The fitted logistic regression model several misclassified samples. The reason for this is that the model does *not* minimize the number of misclassified samples. Instead, it maximizes the (log-)likelihood ($LL(\beta)$) of the data

$$LL(\beta) = \sum_{i=1}^n y_i \ln \sigma(\beta^T x_i) + (1 - y_i) \ln(1 - \sigma(\beta^T x_i))$$

Values β^* which maximize $LL(\beta)$ not necessarily minimize the number of misclassified data points! Still, this formulation makes sense since it helps us to explain the data from a probabilistic perspective and lets us use gradient-based optimization methods.

Note: If you used scikit-learn's `LogisticRegression` with a penalty term, the loss is modified further. This may also have a (minor) effect on the number of misclassified samples.

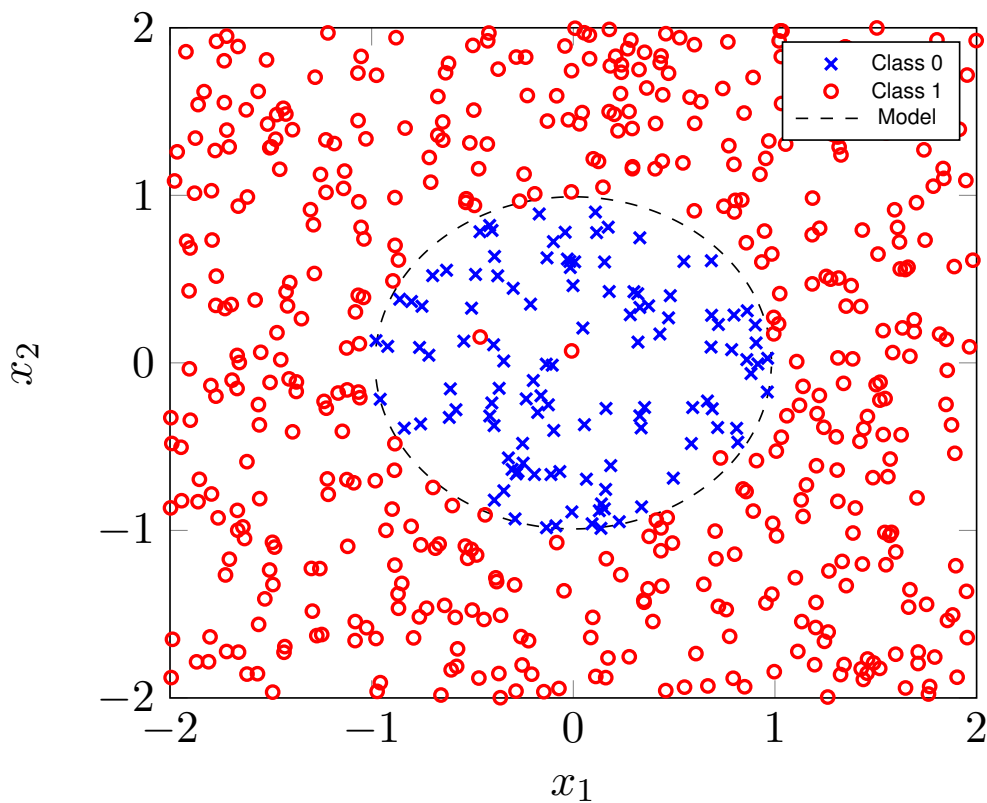


Figure 1 Decision boundary by logistic regression model (Model).

Exercise T3.2 *Maximum likelihood estimation*

You are given the following dataset with the dependent binary variable y and the independent variable x .

x	y
1	0
2	0
2.5	1
4	1

Based on these data points we want to create a logistic regression model with the logistic function σ (or more broadly a sigmoid function):

$$\Pr[Y|X] = p(x) = \sigma(\beta_0 + \beta_1 x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

To estimate the logistic regression coefficients, we will use *maximum likelihood estimation*.

To simplify notation, let $p_i = p(x_i) = \sigma(z_i)$ and $z_i = \beta_0 + \beta_1 x_i$.

a) Determine the likelihood function $L(\beta)$.

Hint: To keep everything simple, it is sufficient to formulate L in terms of p_i (which includes the dependency on β).

b) Find the gradient for the log of the likelihood function $LL(\beta)$. The gradient is defined as:

$$\nabla LL(\beta) = \begin{pmatrix} \frac{\partial LL(\beta)}{\partial \beta_0} \\ \frac{\partial LL(\beta)}{\partial \beta_1} \end{pmatrix}.$$

Hint: Use the chain rule: $\frac{\partial LL}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial LL}{\partial p_i} \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial \beta_j}$ with $z_i = \beta_0 + \beta_1 x_i$. You may use the following derivative of the logistic function σ without proof: $\sigma'(z_i) = \sigma(z_i)(1 - \sigma(z_i))$.

- c) Given the initial values $\beta^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and a learning rate $\alpha = 0.2$, calculate the coefficients after the first iteration of gradient ascent.
- d) If a linear regression model was fitted to a logistic regression dataset, what could be the problems w.r.t. the Gauss Markov properties?

Solution

- a) The likelihood function is the probability that the dependent variables are observed, assuming the data set is produced by the logistic model. For a single data point, its likelihood is p_i if $y_i = 1$ and $1 - p_i$ if $y_i = 0$. Therefore,

$$L(y | x, \beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} = p_3 p_4 (1 - p_1) (1 - p_2). \quad (1)$$

- b) Simply taking the logarithm of the likelihood function gives us

$$\begin{aligned} LL &= \sum_{i=1}^n y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \\ &= \ln(p_3) + \ln(p_4) + \ln(1 - p_1) + \ln(1 - p_2). \end{aligned} \quad (2)$$

Now,

$$\nabla_{\beta} LL = \begin{pmatrix} \frac{\partial LL}{\partial \beta_0} \\ \frac{\partial LL}{\partial \beta_1} \end{pmatrix}, \quad \text{where } \frac{\partial LL}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial}{\partial p_i} LL(y_i, p_i) \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial \beta_j}.$$

Further,

$$\frac{\partial z_i}{\partial \beta_0} = 1, \quad \frac{\partial z_i}{\partial \beta_1} = x_i, \quad \frac{\partial p_i}{\partial z_i} = p_i(1 - p_i), \quad \frac{\partial LL}{\partial p_i} = \frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i}.$$

Calculate partial derivative for β_0 :

$$\begin{aligned} \frac{\partial}{\partial \beta_0} LL(\beta) &= \sum_{i=1}^n \frac{\partial LL}{\partial p_i} \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial \beta_0} = \sum_{i=1}^n \left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right) \cdot (p_i(1 - p_i)) \cdot 1 \\ &= \sum_{i=1}^n y_i(1 - p_i) - p_i(1 - y_i) \\ &= -p_1 - p_2 + (1 - p_3) + (1 - p_4). \end{aligned} \quad (3)$$

Calculate partial derivative for β_1 :

$$\begin{aligned}\frac{\partial}{\partial \beta_1} LL(\beta) &= \sum_{i=1}^n \frac{\partial LL}{\partial p_i} \frac{\partial p_i}{\partial z_i} \frac{\partial z_i}{\partial \beta_1} = \sum_{i=1}^n \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \cdot (p_i(1-p_i)) \cdot x_i \\ &= \sum_{i=1}^n (y_i(1-p_i) - (1-y_i)p_i) x_i \\ &= -p_1 - 2p_2 + 2.5(1-p_3) + 4(1-p_4).\end{aligned}\tag{4}$$

c) Here, for $\beta^{(0)} = (0, 0)^T$, one calculates $z_i = \beta_0 + \beta_1 x_i = 0$ and

$$p_i(z_i) = \frac{\exp(0)}{1 + \exp(0)} = 0.5 \quad \text{for all } i = 1, \dots, 4.$$

Plugging these values into the calculated partial derivatives, we obtain

$$\begin{aligned}\frac{\partial}{\partial \beta_0} LL(\beta^{(0)}) &\stackrel{\text{Equation 3}}{=} -0.5 - 0.5 + 0.5 + 0.5 = 0, \\ \frac{\partial}{\partial \beta_1} LL(\beta^{(0)}) &\stackrel{\text{Equation 4}}{=} -0.5 - 2 \cdot 0.5 + 2.5 \cdot 0.5 + 4 \cdot 0.5 = \frac{7}{4}.\end{aligned}$$

Finally, the update step is given by

$$\begin{aligned}\beta_0^{(1)} &= \beta_0^{(0)} + \alpha \frac{\partial LL}{\partial \beta_0} = 0 + 0.2 \cdot 0 = 0, \\ \beta_1^{(1)} &= \beta_1^{(0)} + \alpha \frac{\partial LL}{\partial \beta_1} = 0 + 0.2 \cdot \frac{7}{4} = 0.35,\end{aligned}$$

which lets us conclude that $\beta^{(1)} = \begin{pmatrix} 0 \\ 0.35 \end{pmatrix}$.

d) First, if a linear regression model were fitted to this data, the predicted values of \hat{y} could be outside $[0, 1]$, which in a binary logistic setting would not be a good prediction since the interpretation as a probability would not make sense.

Second, the properties of *autocorrelation* and *homoscedasticity* would be violated.

- **Autocorrelation:** It can be noted that the residuals of a linear model fitted to a setting with a binary dependent variable would result in positive residuals on one side, negative on the other and in the range of $\hat{y} = [0, 1]$ they would be in the interval $[-1, 1]$. This would form a pattern that indicates autocorrelation.
- **Homoscedasticity:** Error terms do not have constant variance, because true y values take on only two values from the set $\{0, 1\}$, therefore they are heteroscedastic.

Exercise T3.3 *Poisson regression*

You are provided the following numbers from the result of a *Poisson regression model*.

Variable	Estimate	Std. Error
Intercept	1.5499	0.0503
Age	-0.0047	0.0009

- a) According to the model above, what *qualitative* effect does a change in the independent variable age (+1) have on the dependent variable dv .
- b) According to the model above, what *quantitative* effect (on the incidence rate and log-incidence rate) does a change in the independent variable age (+1) have on the dependent variable dv .

Solution

- a) Given the coefficients in the result, we can write the equation

$$\ln(dv) = 1.5499 + (-0.0047)\text{age}.$$

As can be seen, an increase in age will decrease the dependent variable dv .

- b) The log-incidence rate decreases by 0.0047 with an increase of 1 in the variable age:

$$\Delta(\log \text{ incidence rate}) = \beta_1 = -0.0047.$$

The incidence rate decreases by a factor of 0.9953 when the variable age increases by 1 unit:

$$\Delta(\text{incidence rate}) = e^{\beta_1} = 0.9953.$$

This means that there is an approximate 0.5% reduction for each increase in the age by 1 unit.