Logistic Regression with Maximum Likelihood Estimation

Daniel Yates Follow
Feb 24, 2019 · 7 min read

Logistic regression is a statistical model that is used in classification problems.
It allows us to take some features and predict the correct class.

This post will cover:

- When to use logistic regression
- Where logistic regression comes from
- How to fit our model to our data
- How to interpret the parameters

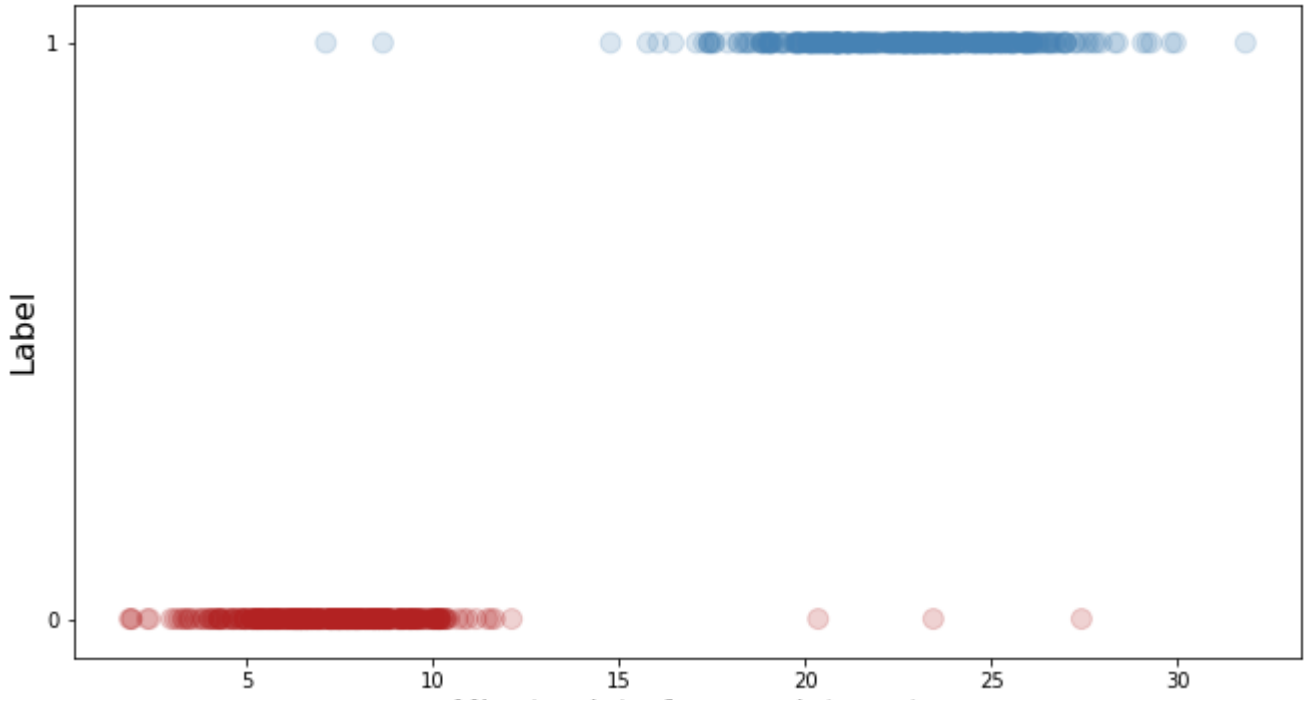
. . .

Setting the scene

As an example, lets say that based on some features we want to predict if a customer is likely to complain.

Where $\boldsymbol{X}$ is our matrix of features, $\boldsymbol{y}$ is our label (1 = will complain / 0 = will not complain)

If we plot our first feature (number of minutes late to appointment) and whether there was a complaint or not we can see that in general customers are more likely to complain the later that someone is for an appointment.

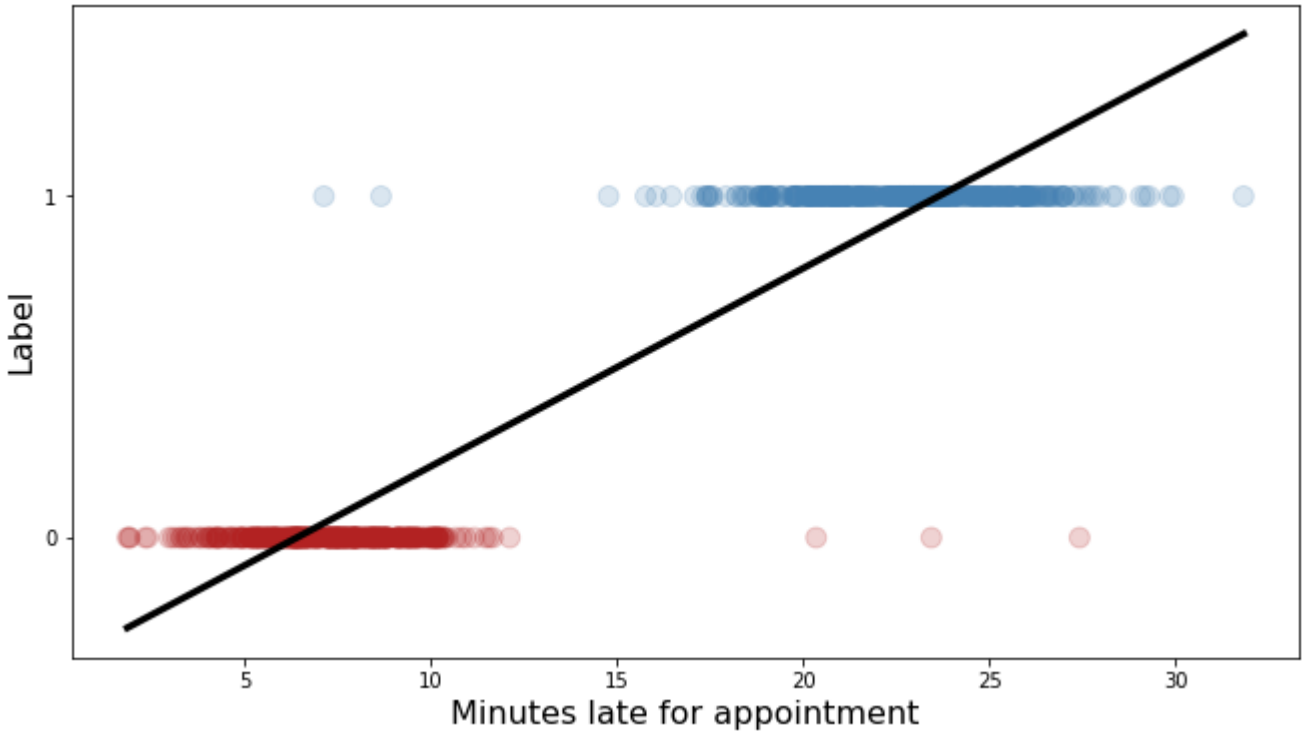


Minutes late for appointment

When to use Logistic regression

Why not linear regression?

If we try to tackle this problem with good ol' linear regression we run in to a few problems.



As you can see using a linear regression model doesn't fit the data and the domain of our linear function is $[-\infty, \infty]$.
So we need a function that captures what the above plot is showing us and also has domain $[0, 1]$.
This is where logistic regression comes in.

Assumptions

Before we go ahead and start throwing this at all our classification problems we need to make sure we're aware of the assumptions. The assumptions are listed at https://www.statisticssolutions.com/assumptions-of-logistic-regression/.
I've summarized them below:

- Dependent variable must be binary/ordinal for binary/ordinal logistic regression: Binary output [0, 1], ordinal output [1, 2, …, k]
- The observations are independent of each other: The observations should not come from repeated measurements or matched data.
- Little/no multicollinearity between features: The features will not be highly correlated.
- Linearity of features and log odds: the features must be linearly correlated. (we'll go through log odds below).

- Large sample size. A general guideline is provided in the original source (link above)

Where logistic regression comes from

A few definitions...

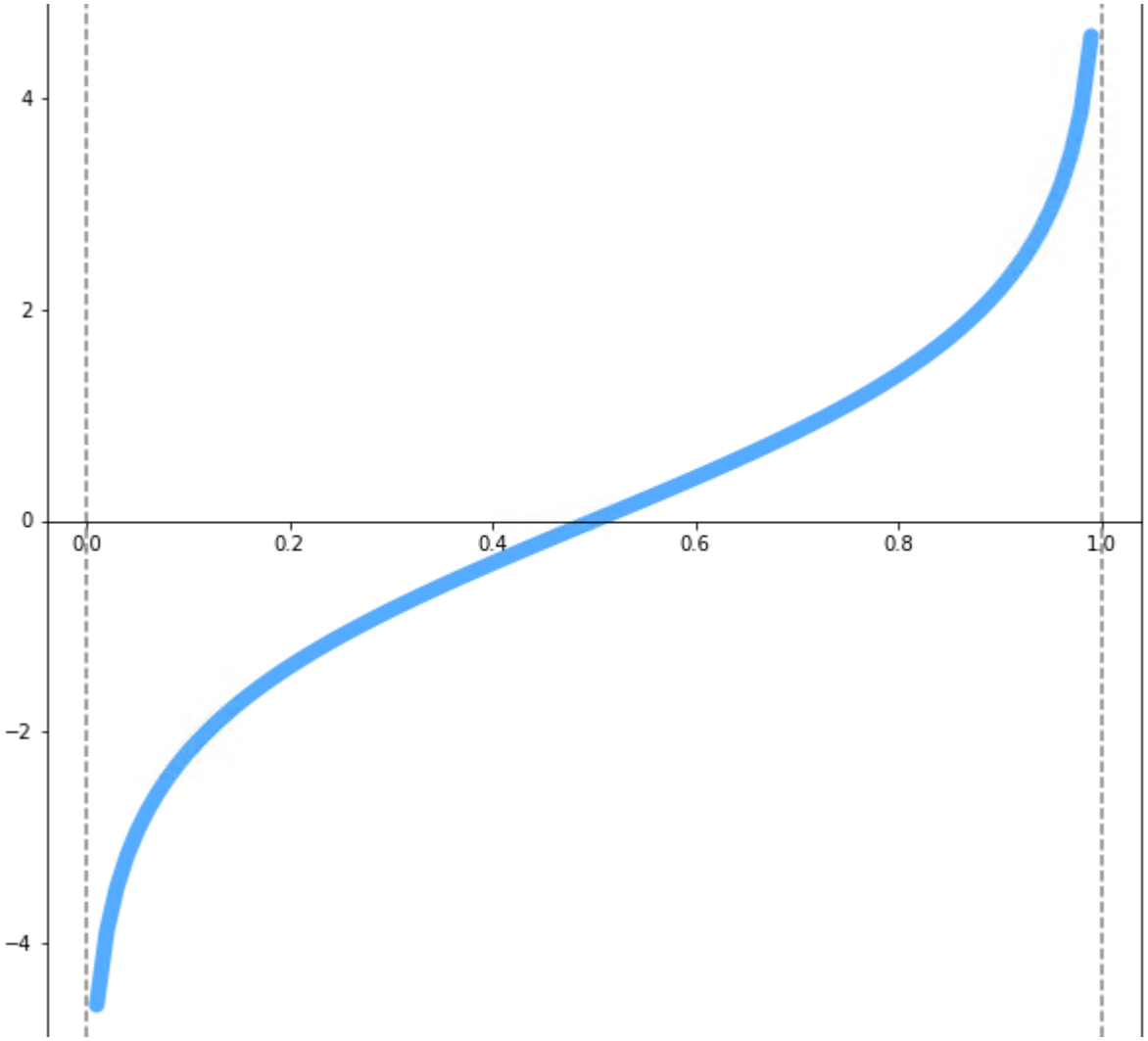
Bernoulli distribution: The Bernoulli distribution is the distribution that takes value 1 with probability p and 0 with probability 1-p. This is a handy distribution for our problem above where we have classes 0 and 1.

Link function: Links the linear model $X\beta$ to the mean of a particular distribution. So the link function will essentially take our linear model and make it output within the desired distribution.

Logit and inverse logit

As the problem we are trying to model outputs within the Bernoulli distribution, we will use the link function for this distribution for our linear model.

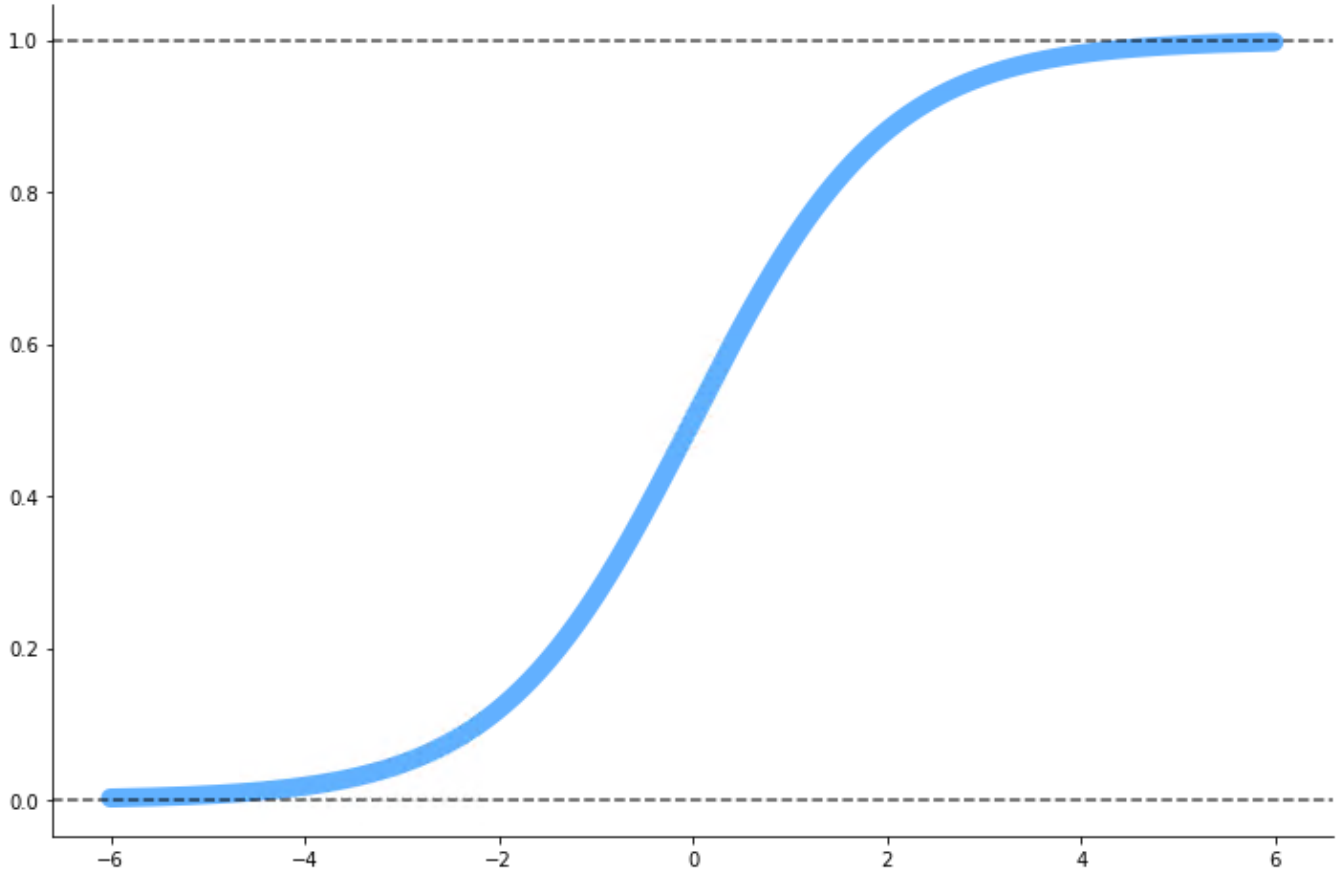
The link function for the Bernoulli Distribution is the Logit.



$$Logit(x) = \frac{x}{1-x}$$

This line is certainly different to our linear function and looks a lot closer to the shape of our data. It just requires some rotation.

If we take the inverse of the logit we get…



$$Logit^{-1}(x) = \frac{e^x}{1 + e^x}$$

This is the function that we will use for logistic regression. It will fit the data nicely and as you can see it outputs in the domain [0, 1].

$$Logit^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{1 + e^{-(X\beta)}}$$

Now we have the logistic regression formula!

*Note: **X**β assumes there is a constant column (column of ones) as 1× β_0 will be our intercept.*

How to fit our model to our data

Now back to our customer complaint problem, we're going to use logistic regression to take our features ***X*** and tell us the probability that the customer will complain.

$$Logit^{-1}(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} = P(X;\beta)$$

Without fitting our model it's unlikely that our model will produce anything meaningful. So we need a method of training our model to make it produce the best possible results.

What is fitting?

Fitting a model is updating a models parameters {β0, β1, …, βk} such that it gives us the most accurate outputs.
There are different methods of doing this but we're going to focus on maximum likelihood estimation (great article explaining MLE https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1).

Before we dive into maximum likelihood estimation we need to understand likelihood and log likelihood.

likelihood and log likelihood

We need a measure that tells us how well our model fits our observed data. For this we're going to use the log likelihood.

The likelihood function is

$$n$$

$$L(\beta) = \prod_{i=1} (P(X_i; \beta))^{(y_i)} (1 - P(X_i; \beta))^{(1-y_i)}$$

It is the product of the probabilities that y is equal to 1. We want this to be as close to 1 as possible. However this is a tricky function to maximize so we take the log of it and that makes things easier to work with.

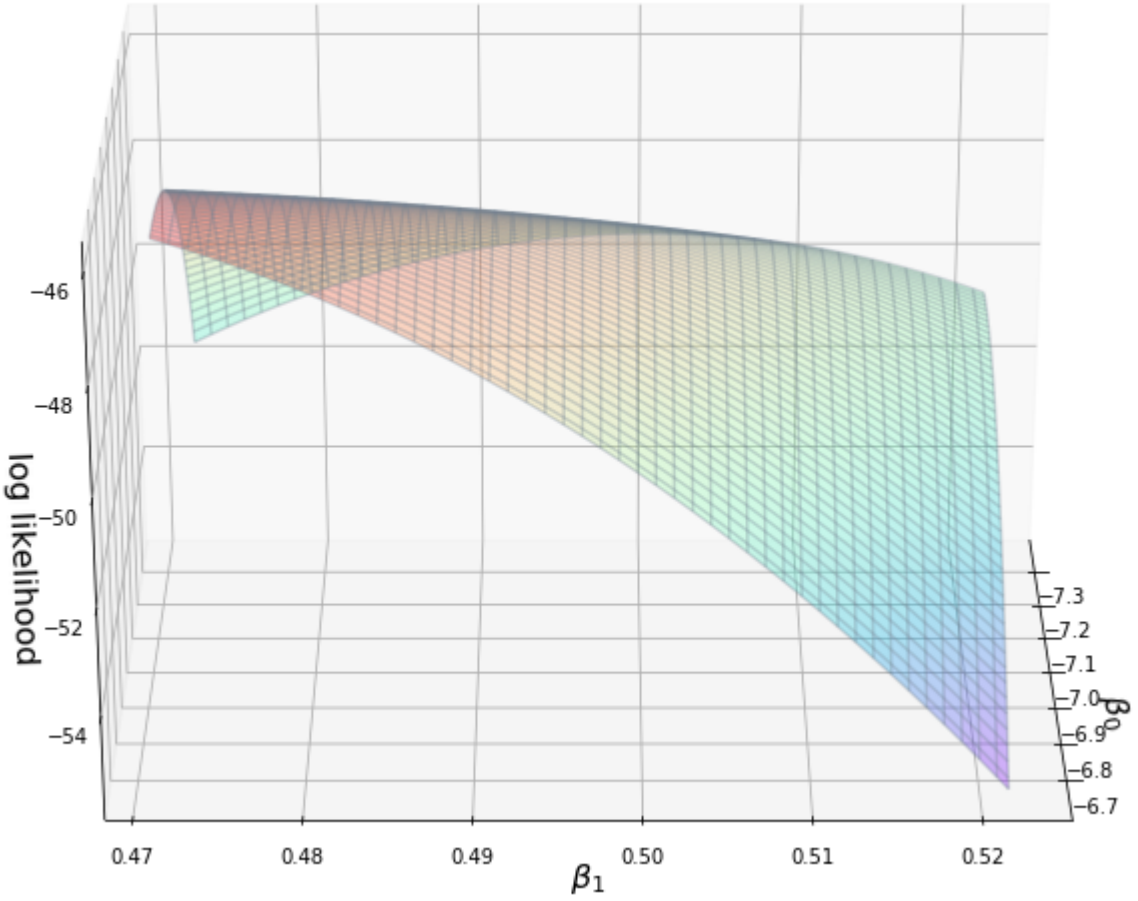
Thanks to some snazzy logarithm laws the natural log of the likelihood function becomes

$$\begin{aligned} \ell(\beta) &= ln \left(\sum_{i=1}^{n} (P(X_i; \beta))^{(y_i)} (1 - P(X_i; \beta))^{(1-y_i)} \right) \\ &= \sum_{i=1}^{n} (y_i)\ ln(P(X_i; \beta)) \ + \ (1 - y_i)\ ln(1 - P(X_i; \beta)) \end{aligned}$$

Since we're now taking the log of the probabilities, and our probabilities are in the range 0 to 1, all of the domain of this function is [0, -∞]. Therefore we want to get the output as close to 0 as possible.

Maximum likelihood estimation

Using maximum likelihood estimation we will find the β parameters that reach the global maximum of the log likelihood function.
If we plot the log likelihood function and some β parameters we can see the plane of our function.



By tweaking the β parameters we can see the effect that this has on our log likelihood function.

So how do we update our β parameters to maximize this function?

Newton-Raphson

There are a variety of different methods to converge to the maximum of this function however we are just going to focus on the Newton-Raphson method.
The equation for a single update is

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial \ell}{\partial \beta}$$

We take the derivative of the log likelihood function with respect to our β parameters. This tells us how much a change in β affects our log likelihood function.
We then multiply this by the inverse of the second derivative of our log likelihood function with respect to β .

The second derivative tells us how much the first derivative is changing with respect to the log likelihood.

Below are the derivatives.

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial z} \cdot \frac{\partial z}{\partial \beta}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \frac{y_i - \sigma}{\sigma(1-\sigma)} \cdot \sigma(1-\sigma) \cdot x_i$$
$$= \sum_{i=1}^{N} x_i(y_i - \sigma)$$

=

$$\ell = \sum_{i=1}^{N} (y_i)(ln(\sigma)) + (1-y_i)(ln(1-\sigma))$$
$$\frac{\partial \ell}{\partial \sigma} = \sum_{i=1}^{N} \frac{y_i}{\sigma} - \frac{1-y_i}{1-\sigma}$$
$$= \sum_{i=1}^{N} \frac{y_i(1-\sigma)}{\sigma(1-\sigma)} - \frac{\sigma(1-y_i)}{\sigma(1-\sigma)}$$
$$= \sum_{i=1}^{N} \frac{y_i(1-\sigma) - \sigma(1-y_i)}{\sigma(1-\sigma)}$$
$$= \sum_{i=1}^{N} \frac{y_i - y_i\sigma - \sigma + y_i\sigma}{\sigma(1-\sigma)}$$
$$= \sum_{i=1}^{N} \frac{y_i - \sigma}{\sigma(1-\sigma)}$$

$$\sigma = \frac{1}{1+e^{-z}}$$
$$\frac{\partial \sigma}{\partial z} = -(1+e^{-z})^{-2} - (e^{-z})$$
$$= \frac{e^{-z}}{(1+e^{-z})^2}$$
$$= \frac{1}{(1+e^{-z})} \cdot \frac{e^{-z}}{(1+e^{-z})}$$
$$= \frac{1}{(1+e^{-z})} \cdot \frac{e^{-z}+1-1}{(1+e^{-z})}$$
$$= \frac{1}{(1+e^{-z})} \cdot \frac{(1+e^{-z})}{(1+e^{-z})} - \frac{1}{(1+e^{-z})}$$
$$= \sigma(1-\sigma)$$

$$z = \sum_{i=1}^{N} x_i^T \beta$$
$$\frac{\partial z}{\partial \beta} = \sum_{i=1}^{N} x_i$$

First derivative

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \frac{y_i - \sigma}{\sigma(1-\sigma)} \cdot \sigma(1-\sigma) \cdot x_i$$
$$= \sum_{i=1}^{N} x_i(y_i - \sigma)$$
$$\frac{\partial^2 \ell}{} = -\sum^{N} x_i x_i^T \sigma(1-\sigma)$$

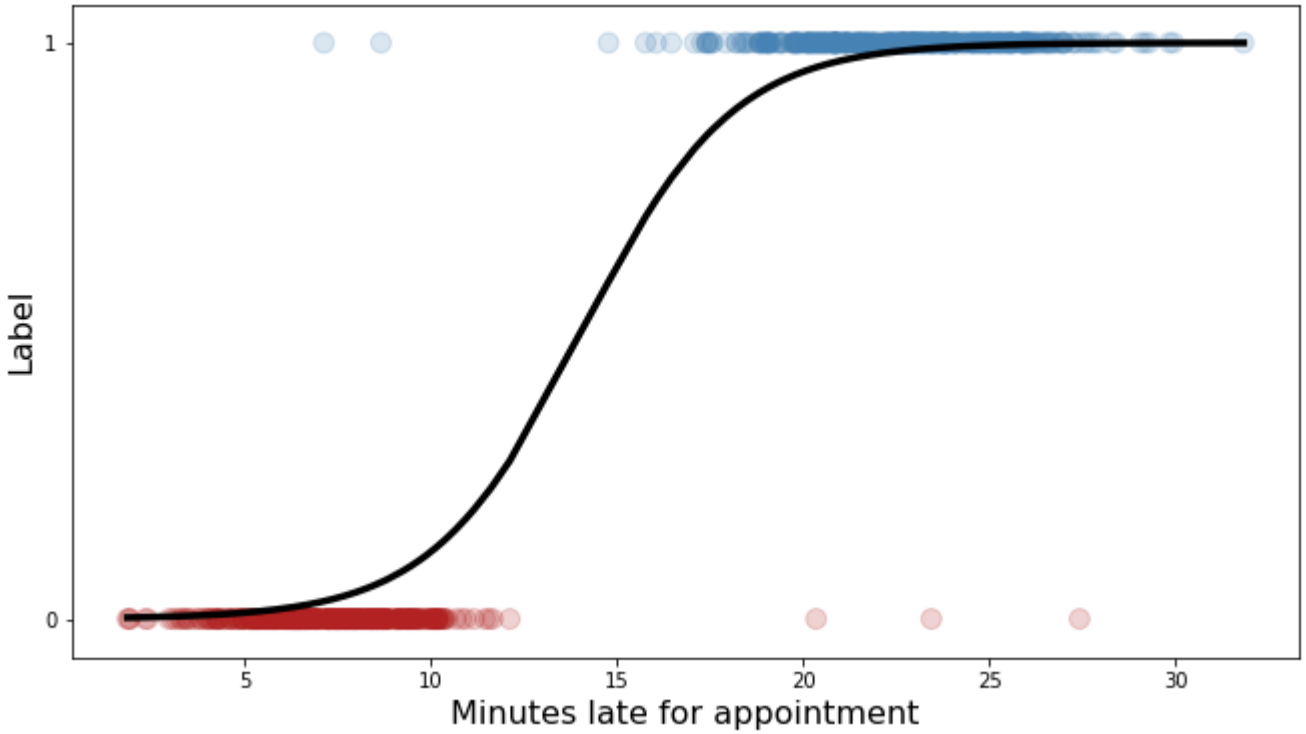
$$\frac{\partial}{\partial\beta\partial\beta^T} \sum_{i=1}$$

Second derivative

Now we apply this update until we converge to the maximum.

Fitted model

Our fitted model looks like



This is much better than the linear regression model at the start

Interpreting the parameters

odds and odds ratio

The odds tell us how likely an event is to happen.
To calculate the odds for a continuous feature we have:

$$when\ x_1 = k$$

$$\frac{p(x)}{1-p(x)} = \frac{e^{\beta_0+\beta_1 k}/(1+e^{\beta_0+\beta_1 k})}{1-e^{\beta_0+\beta_1 k}/(1+e^{\beta_0+\beta_1 k})}$$

$$= \frac{e^{\beta_0+\beta_1 k}/(1+e^{\beta_0+\beta_1 k})}{1/(1+e^{\beta_0+\beta_1 k})}$$

$$= e^{\beta_0+\beta_1 k}$$

$$when\ x_1 = k+1$$

$$\frac{p(x)}{1-p(x)} = \frac{e^{\beta_0+\beta_1 (k+1)}/(1+e^{\beta_0+\beta_1 (k+1)})}{1-e^{\beta_0+\beta_1 (k+1)}/(1+e^{\beta_0+\beta_1 (k+1)})}$$

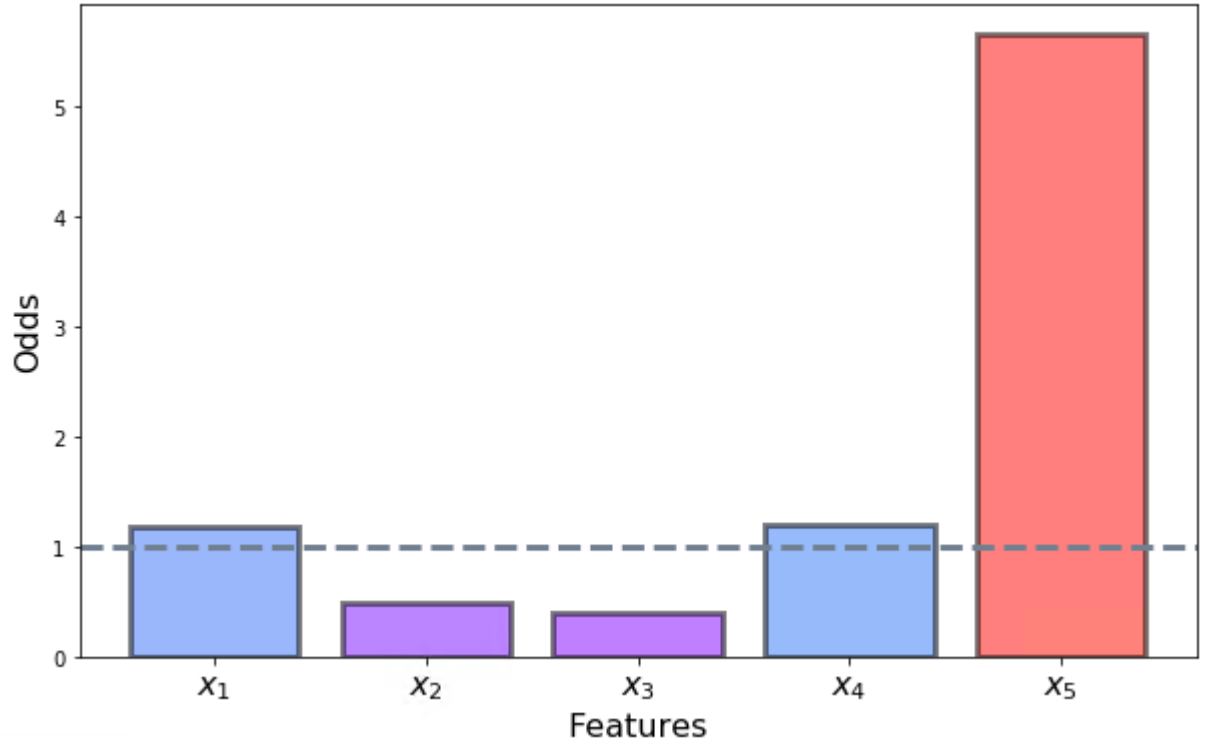
$$= \frac{e^{\beta_0+\beta_1 k+\beta_1}/(1+e^{\beta_0+\beta_1 k+\beta_1})}{1/(1+e^{\beta_0+\beta_1 k+\beta_1})}$$
$$= e^{\beta_0+\beta_1 k+\beta_1}$$

Now that we have the odds we can calculate the odds ratio.

$$\frac{odds\ when\ x_1 = k+1}{odds\ when\ x_1 = k} = \frac{e^{\beta_0+\beta_1 k+\beta_1}}{e^{\beta_0+\beta_1 k}} = e^{\beta_1}$$

So for a one unit increase in $x1$, with all other features held constant the chances of a positive outcome are increased by e to the power of our β parameter.

	β	e^{β}
x_1	0.161387	1.17514
x_2	-0.710869	0.491217
x_3	-0.908581	0.403096
x_4	0.18969	1.20887
x_5	1.73058	5.64395



So a one unit increase in $X5$ whilst all other features are held constant means it is 5.64 times more likely that there will be a customer complaint.

Conclusion

Hopefully this has provided you with a fairly comprehensive overview of logistic regression using maximum likelihood estimation.
If you spot any errors or have any questions please leave a comment :)

Resources

- https://www.statisticssolutions.com/assumptions-of-logistic-regression/
- Elements of Statistical Learning (https://web.stanford.edu/~hastie/ElemStatLearn/)
- https://en.wikipedia.org/wiki/Logit
- https://en.wikipedia.org/wiki/Bernoulli_distribution
- https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

Machine Learning | Logistic Regression | Maximum Likelihood | Statistics