Logistic Regression

Jeremy Teitelbaum

Logistic Regression

The logistic model

The log-odds of an event increase linearly with an independent variable.

$$\log \frac{p}{1-p} = ax + b$$

Example: The chance that a person buys a product depends on how many times they encounter advertising for that product.

The sigmoid function

$$\log \frac{p}{1-p} = ax + b$$

means that

$$p(x) = \frac{1}{1 + e^{-ax-b}}$$

The logistic curve

The function

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

is called the logistic function.

$$e^{-ax} = (e^{a})^{-x}$$

$$a = \ln 2$$

$$(e^{\ln 2})^{-x} = 2^{-x}$$

$$2^{x} = (e^{\ln 2})^{x}$$

$$(e^{\ln 2})^{x} \ln 2$$

$$= 2^{x} \ln 2$$

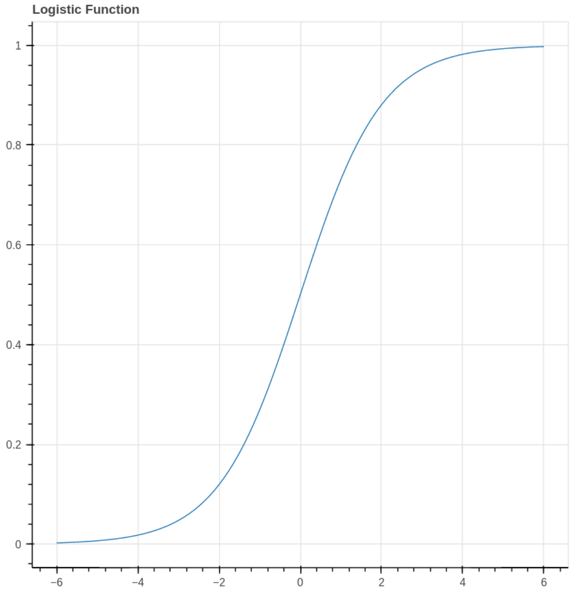


Figure 1: Logistic Curve

Sample data

Likelihood of event increases with x. Out of 100 tries:

x	-3	-2	-1	0	1	2	3
Occurrences (out of 100)	10	18	38	50	69	78	86

Two points of view

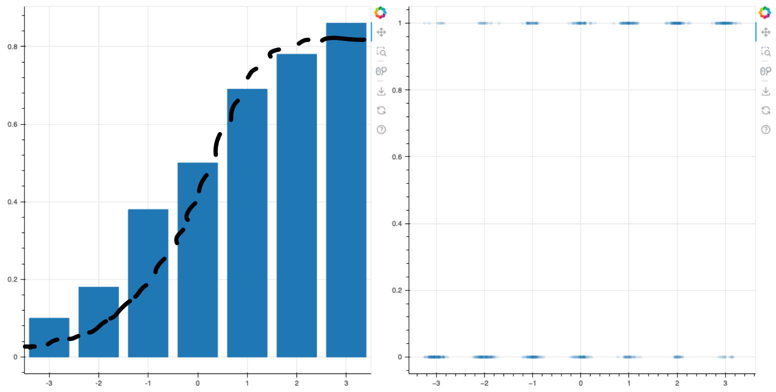


Figure 2: Logistic Data

$$\sim \frac{1}{1+e^{-ax-b}}$$

The Likelihood

The parameters a and b are unknown. But if we knew them, then we could figure out how likely our results were. For example, the chance of getting 10 positive outcomes is

$$p(10+|x=-3,a,b) = C(\sigma(a(-3)+b)^{10}(1-\sigma(a(-3)+b))^{90}$$

where C is a constant (it's a binomial coefficient).

$$p(a,b,x) = \frac{1}{1+e^{-ax-b}} = \sigma(x)$$

Sample data

Likelihood of event increases with x. Out of 100 tries:

x	-3	-2	-1	0	1	2	3
Occurrences (out of 100)	10	18	38	50	69	78	86

More on the likelihood

Assuming independence (given x, a, and b) the chance of our data is

$$P(\text{data}|a,b) = CP(10+|x=-3)P(18+|x=-2)\cdots P(86+|x=3)$$

Sample data

Likelihood of event increases with x. Out of 100 tries:

x	-3	-2	-1	0	1	2	3
Occurrences (out of 100)	10	18	38	50	69	78	86

$$P(N+ \,|x=x_0) \Rightarrow$$

Pe

$$\Rightarrow \binom{100}{N} \sigma(ax_0+b)^N (1-\sigma[ax_0+b])^{100-N}$$

Still more

Here each term is

$$P(y+|x) = \sigma(ax+b)^y(1-\sigma(ax+b)^{N(x)-y}$$

where $N(x)$ is the number of trials with that given x value. (this is a Binomial random variable).

Given some features

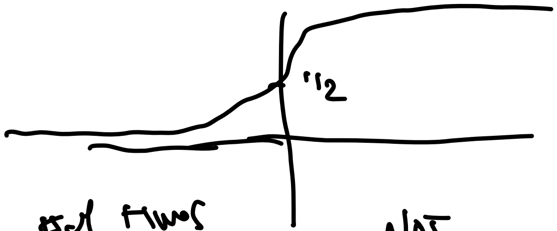
x

random event likelihood depends on x

p chance that event occurs

$$\boxed{\log \frac{p}{1-p} = ax+b}$$

$$p = \frac{1}{1+e^{-(ax+b)}} \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$



x # of ads you saw

1/0 bought or did not buy the thing

Find a,b given data

person → # of times & bought or NOT

$$\text{Data:} \quad \begin{array}{c} \begin{bmatrix} x_1 & 1 \\ x_2 & 0 \\ x_3 & 0 \\ x_4 & 1 \\ \vdots & \end{bmatrix} \\ \uparrow \quad\quad \uparrow \\ x_i \quad\quad y_i \end{array}$$

The log likelihood

We want to find the a and b that make our observed data *most likely*. To do this we need to find a, b that maximize P or, more simply $\log P$.

$$\log P = \sum_{i=0}^{6} \left[y_i \log P(y_i|x_i) + (100 - y_i) \log(1 - p(y_i|x_i)) \right]$$

We can drop the constant since it won't affect where the maximum occurs.

For each sample (person)

saw ad x_i times   Result is $y_i = 0, 1$ a, b unknown

$$p(+1|x_i; a, b) = \sigma(a x_i + b)$$

chance of 1: $\sigma(a x_i + b)$.   chance of 0 is $1 - \sigma(a x_i + b)$

$$L(a, b) = \prod_i \sigma(a x_i + b)^{y_i} (1 - \sigma(a x_i + b))^{1 - y_i}$$

$$\log L = \sum_i y_i \log \sigma(a x_i + b) + (1 - y_i) \log(1 - \sigma(a x_i + b))$$

Minimize $-\log L$ using gradient descent

Vector/Regression Form

Our data matrix consists of N rows (and 1 column), one for each person viewing the ads. The entry in each row is the number of times they saw the add.

The target matrix consists of 0 and 1 depending on whether they made a purchase or not.

We want to “fit” an equation that gives 0 or 1 as a function of x, but we can’t do this exactly, only in probabilistic terms.

This is why it’s called “regression.”

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 0 \\ \vdots \\ y_i \\ \vdots \\ 1 \end{bmatrix}$$

More on Vector/Regression Form

For each row of our matrix, the chance that y_i is 1 is $p(x_i)$ (given by the sigmoid function with parameters a, b) and the chance that $y_i = 0$ is $(1 - p(x_i))$. So our likelihood is

$$L(a,b) = C\prod_{i=0}^{N-1} p(x_i)^{y_i}(1-p(x_i))^{(1-y_i)} \qquad p(x_i) = \sigma(ax_i+b)$$

and

$$\log L(a,b) = C' + \overset{\sum}{\cancel{\prod}}_{i=0}^{N-1} y_i \log p(x_i) + (1-y_i)\log(1-p(x_i)).$$

Ignoring irrelevant constants this is

$$\boxed{\log L = \overset{Y^T \log p(X)}{Y} \cdot \log p(X) + (1-Y)\cdot \log(1-p(X))}$$

where for each row of X, $p(X)$ has $\sigma(ax_i + b)$ with (unknown) parameters a and b.

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad X^* = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} \qquad X^* \overset{W \downarrow}{\begin{bmatrix} a \\ b \end{bmatrix}} = \begin{bmatrix} ax_1+b \\ \vdots \\ ax_N+b \end{bmatrix}$$

$$\sigma(XW) = \begin{bmatrix} \sigma(ax_1+b) \\ \vdots \\ \sigma(ax_N+b) \end{bmatrix} \qquad \log\sigma(XW) = \begin{bmatrix} \log\sigma(ax_1+b) \\ \vdots \\ \log\sigma(ax_N+b) \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \qquad \cancel{0}\ Y^T \log(\sigma(XW)) = \sum_{y_i=1} \log\sigma(xw)$$

$$1 - Y = \begin{bmatrix} 1-y_1 \\ 1-y_2 \\ \vdots \\ 1-y_N \end{bmatrix} \qquad \log(1-\sigma(XW)) = \begin{bmatrix} \log(1-\sigma(ax_1+b)) \\ \vdots \\ \vdots \end{bmatrix}$$

$$(1-y)^T \log(1-\sigma(Xw)) = \sum_{y_i=0} \log(1-\sigma(ax_i+b))$$

The case of multiple features

In the case of multiple features, we have a set of k measurements for each sample (perhaps exposure to different types of ads) and a single outcome (buy/do not buy). This yields an $N \times k$ data matrix X. We seek a set of weights $m_1, \dots, m_k$ and an "intercept" b so that

$$\log \frac{p}{1-p} = \sum m_i x_i + b$$

relates the log-odds of our event occurring with the values of the features.

Note: Just as with linear regression, we can create a "fake" feature that is all 1, and then extend our data matrix to $N \times (k+1)$. Then $b = m_{k+1}$ and we can write

$$\log \frac{P}{1-P} = XM$$

where the right side is an $N \times 1$ matrix and the left side is an $N \times 1$ matrix whose entries are $\log \frac{p}{1-p}$.

The probability

From this we get the matrix equation

$$P = \sigma(XM)$$

The matrix P has the probability of getting a positive outcome for each sample given the features.

$$X = \begin{array}{c} \begin{array}{cc} \overbrace{\hphantom{x_{11} \quad x_{12} \quad \cdots \quad x_{1K}}}^{\text{features}} & \leftarrow \text{fake extra column} \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} & 1 \\ x_{21} & & & x_{2K} & 1 \\ & & \vdots & & \end{bmatrix} \end{array}$$

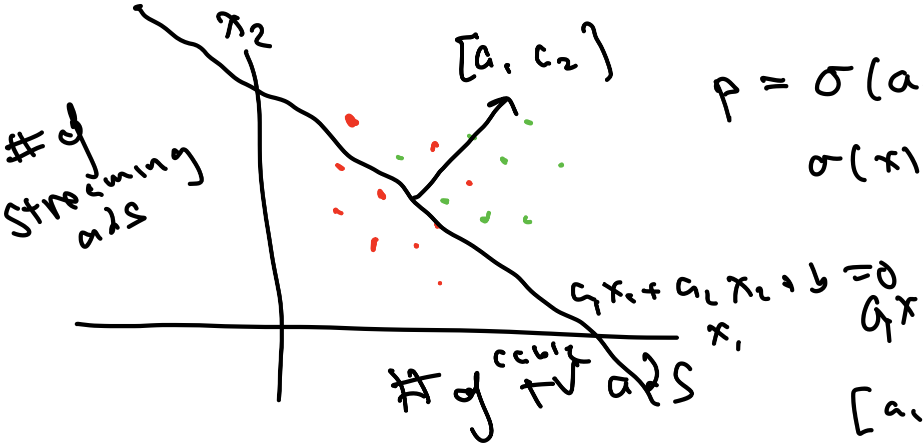
$$\sigma(XM) = \sigma\left(X \begin{bmatrix} m_1 \\ m_2 \\ m_K \\ b \end{bmatrix} \right) = \sigma\left(\begin{bmatrix} m_1 x_{11} + \cdots + m_K x_{1K} + b \\ \vdots \end{bmatrix} \right)$$

$$= \begin{bmatrix} \sigma(m_1 x_{11} + \cdots + m_K x_{1K} + b) \\ \vdots \end{bmatrix}$$

$$\log L = Y^T \log \sigma(XM) + (1-Y)^T \log(1-\sigma(XM))$$

A geometric remark

One way to think of this is that if the features (a row of X), thought of as a vector, points "more in the direction of the weight vector" M, then the probability of getting a positive outcome increases. If it's perpendicular, you get even odds. If it points oppositve the weight vector, you're unlikely to get what you want.



$$p = \sigma(a_1x_1+a_2x_2+b)$$

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

$$a_1x_1+a_2x_2+b=0$$

$$[a_1\ a_2]\cdot[x_1, x_2] = -b$$

The target

We have a vector Y which records when our event happened, and when it didn't.

The log-likelihood

$$L(M) = Y^\top \log(\sigma(XM)) + (1 - Y^\top)(1 - \log(\sigma(XM)))$$

Problem: Given X and Y, find M that maximizes this.

Credit card default

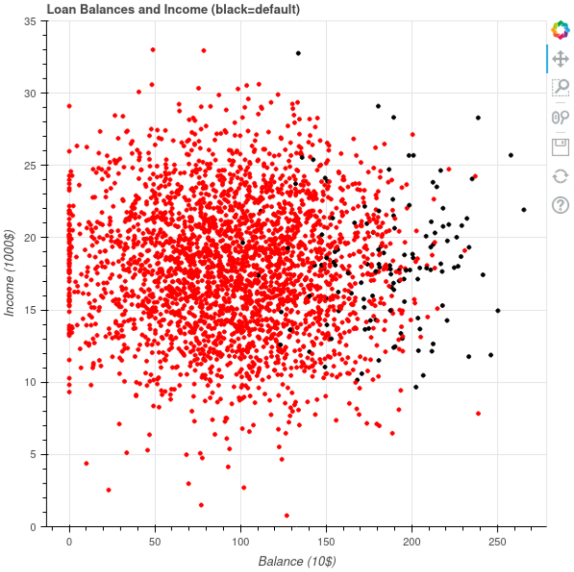


Figure 3: Default

Default with logistic line

$a x_1 + a_2 x_2 + b = 0$

$\updownarrow$

even-odds

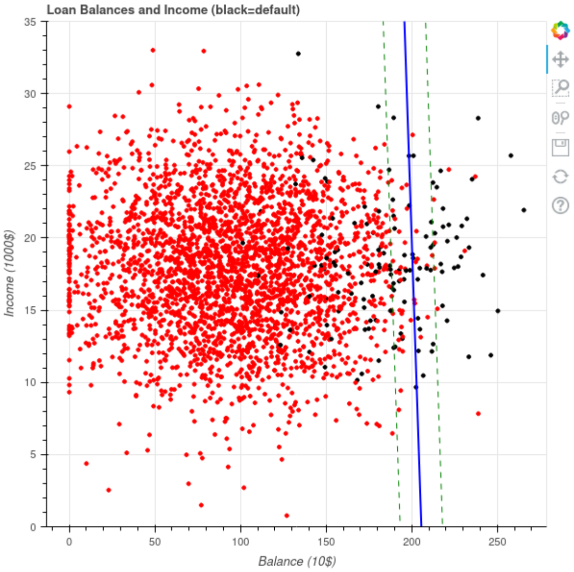


Figure 4: Default with line

Logistic regression for classification

We can use logistic regression for classification by fitting the logistic model and then saying that a point should be classified as 1 if the probability p given by the model says it is 1 with greater than .5 probability.

We could also set a more stringent requirement.

The “decision surface”

In the logistic regression model, for a given sample x with features x_i $(i = 1, \dots, k+1)$, the log odds of that sample yielding a “positive” result is

$$\log \frac{P}{1-P} = \sum x_i m_i$$

where the m_i are the weights. Notice that the equation $f(x) = \sum x_i m_i$ is a linear function of the features. The equation $f(x) = 0$ defines a “hyperplane” in feature space. (In the graph above, this is the blue line on the default data). On that line, it’s even odds if the target is 1 or 0.

If $f(x) > 0$, the odds are better than even that the target value for that point is 1; and if $f(x) < 0$ the odds are less than even.

If you are trying to classify points, you could say points where $f(x) > 0$ should be classified as 1 (because, more likely than not, the model says that they are a 1).

An example of classification

The sklearn digits dataset consists of a large number of 8×8 bitmap images together with labels from 0 to 9. For example:

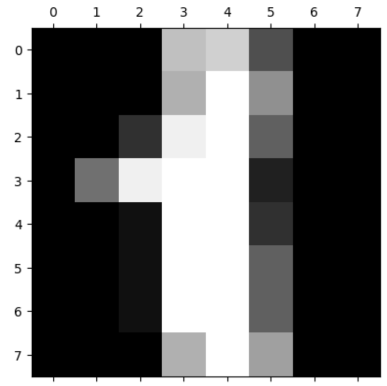


Figure 5: digit

MNIST

60K 28x28

gray scale

(0 -> 255)

handwritten digits

with labels

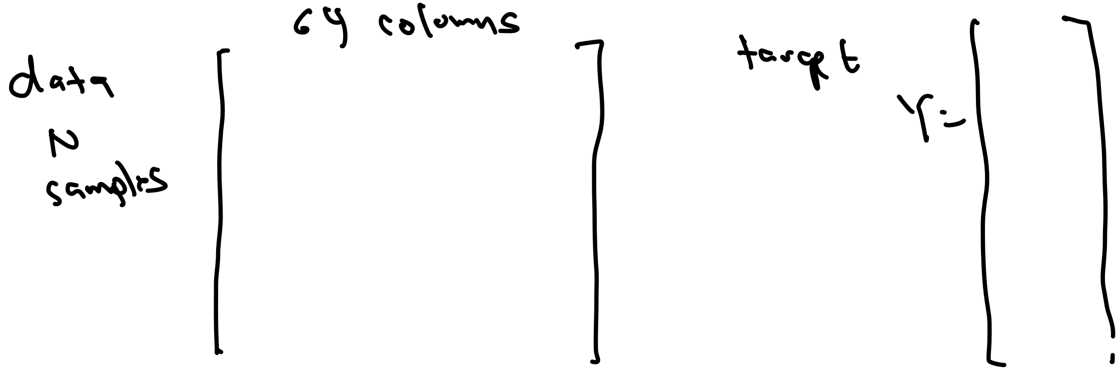
Digit recogntion (two-class)

We can view our images not as 8 × 8 arrays but as 64-entry vectors. Let's just focus on the zeros and ones. We fit a logistic regression model where the target is the value 0 or 1. This turns out to do an exceptionally good job at distinguishing these digits.

each sample is a 64 entry array

2 class classification

is it a zero? OR NOT.



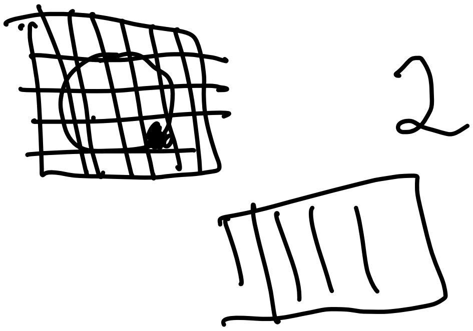
$y_i = 0$ if NOT a '0'

$= 1$ if it is a '0'

data is rescaled: ~~dg~~ divide by 255. so $0 \le$ data ≤ 1

$$\log L = Y^T \log \sigma(XM) + (1-Y)^T (1 - \log \sigma(XM))$$

$$M = \begin{bmatrix} m_1 \\ \vdots \\ m_{64} \end{bmatrix}$$



Multiclass logistic regression

Softmax

In multiclass logistic regression, we imagine not only that our data depends on multiple features but that there are several possible outcomes to our experiment.

Given numbers $z_1, \dots, z_n$, let

$$F(z_1, \dots, z_n) = \sum_{i=1}^{n} e^{z_i}$$

and define the "softmax" function by

$$\sigma(z_1, \dots, z_n) = \begin{bmatrix} \frac{e^{z_1}}{F} & \frac{e^{z_2}}{F} & \cdots & \frac{e^{z_n}}{F} \end{bmatrix}$$

$$\sigma(z_1, z_2) = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} , \frac{e^{z_2}}{e^{z_1} + e^{z_2}}$$

$$= \underset{p}{\frac{1}{1 + e^{z_2 - z_1}}} , \underset{1-p}{\frac{1}{1 + e^{z_1 - z_2}}}$$

Softmax

The softmax function is the higher dimensional generalization of the sigmoid function. Notice that it yields a vector of numbers between 0 and 1 whose sum is 1.

One-hot encoding

The second important element of multiclass regression is how to encode the labels. For example, in our digit problem, the labels run from 0 to 9. Given an image, we want to compute probabilities $p_0, \dots, p_9$ which add to 1 and, where, hopefully, the largest p_i corresponds to the true label.

To set this up we convert our labels to one hot encoding. In this picture, we replace our single vector with entries from 0 to 9 with a matrix with 10 columns. Each row of this matrix has a zero everywhere *except* in the column corresponding to the label is correct.

So if the label for an image is 2, the corresponding row of the target matrix is

$$[0, 0, 1, 0, 0, 0, 0, 0, 0, 0].$$

The model

If we have N samples, k features, and r classes, then our weight matrix M is $k \times r$, our target matrix Y is $N \times r$, and our data matrix X is $N \times k$. Our model says that

$$P = \sigma(XM)$$

where $\sigma(XM)$ means "apply the softmax function to each row of XM"; each row has r entries.

You can think of P as an attempt to "estimate" Y.

The rows of P give the probabilities of getting each possible label for that set of features.

Max Likelihood

As before we seek M so that the observed classification is most likely given M. To find the likelihood:

The probability that the i^{th} sample is in class j is $p_s(x_i; M)$ where p_s is the s^{th} entry in the i^{th} row of the matrix P. We can write this as

$$P(i, s) = \prod_{s=1}^{r} p_s(x_i; M)^{y_s}$$

since y_s is zero except at the correct class. Taking the logarithm makes this a sum:

Matrix form of max likelihood

$$\log P(i,s) = \sum_{s=1}^{r} y_s \log p_s(x_i; M).$$

By independence, the total log probability of this data is the sum of this over all samples and corresponding y-values.

$$\log L(M) = \sum_{i=1}^{N}\sum_{s=1}^{r} y_s \log p_s(x_i, M)$$

This can be written in matrix form as

$$\log L(M) = \mathrm{trace}(Y^T \log \sigma(XM)))$$

We need to maximize this; we'll consider that later.

End