Introduction to Linear and Logistic Regression

C. Caruvana

5th August 2020

Abstract

We introduce some theory and applications of linear regression and logistic regression.

Contents

1 Linear Regression

Suppose we wish to determine if there is a linear relationship between two variables; e.g. weight and blood pressure. We would collect data in the form of ordered pairs (x, y); in this example, x measures weight and y measures blood pressure. Suppose we've collected n data points, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We can plot these observations in the plane and visually inspect if they display a linear shape. However, eyeballing is neither particular efficient nor generally effective. So let's investigate a formulaic approach.

1.1 Least Squares

We will start in the context where we have one independent variable. Once we've developed the theory corresponding to that case, we'll move on to several independent variables.

Suppose we have the observed data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and a proposed model

$$L(x) = mx + b.$$

The standard method of least squares to to compute

$$\sum_{j=1}^{n} (mx_j + b - y_j)^2$$

which can be visualized as the sum of the square distance between the predicted values and the observed values. As a point of comparison, notice that the actual distance between the predicted values and the observed values is given by

$$|mx_j + b - y_j| = \sqrt{(mx_j + b - y_j)^2}.$$

One reason to consider the square distances is to facilitate optimization. Note that

$$\frac{\partial}{\partial m}|mx_j + b - y_j| = \pm x_j$$

and

$$\frac{\partial}{\partial b}|mx_j + b - y_j| = \pm 1.$$

This doesn't quite offer any information to minimize the distances. However, observe that

$$\frac{\partial}{\partial m}\sum_{j=1}^{n}(mx_j + b - y_j)^2 = \sum_{j=1}^{n} 2x_j(mx_j + b - y_j) = 2m\Sigma(x^2) + 2b\Sigma x - 2\Sigma(xy)$$

and

$$\frac{\partial}{\partial b}\sum_{j=1}^{n}(mx_j + b - y_j)^2 = \sum_{j=1}^{n} 2(mx_j + b - y_j) = 2m\Sigma x + 2nb - 2\Sigma y.$$

Since our goal is to optimize, we can set up the system

$$\begin{cases} 2m\Sigma(x^2) + 2b\Sigma x - 2\Sigma xy &= 0 \\ 2m\Sigma x + 2nb - 2\Sigma y &= 0 \end{cases} \iff \begin{cases} m\Sigma(x^2) + b\Sigma x &= \Sigma(xy) \\ m\Sigma x + nb &= \Sigma y \end{cases}$$

which we could rewrite as a matrix equation:

$$\begin{bmatrix} \Sigma(x^2) & \Sigma x \\ \Sigma x & n \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \Sigma(xy) \\ \Sigma y \end{bmatrix}$$

The left-most matrix has an inverse if and only if its determinant, $n\Sigma(x^2) - (\Sigma x)^2$, is non-zero. In particular, the inverse of the left-most matrix would be

$$\frac{1}{n\Sigma(x^2) - (\Sigma x)^2} \begin{bmatrix} n & -\Sigma x \\ -\Sigma x & \Sigma(x^2) \end{bmatrix}$$

which we can then left-multiply to find solutions for m and b. One can verify that these solutions are

$$m = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{n\Sigma(x^2) - (\Sigma x)^2}$$

and

$$b = \frac{\Sigma y - m\Sigma x}{n} = \bar{y} - m\bar{x}.$$

1.1.1 Perpendicular Approach

One may wonder if minimizing the distances in the "vertical" direction is really the best way. Why not minimize against the shortest distance from a point to a line, a "perpendicular" approach? Let's explore this route for the sake of being thorough. For an effectively identical exposition on this topic, please see Wolfram MathWorld.

Though we explore this approach in the case of one independent variable, we will be opting for the standard optimization when we extend to several independent variables in Section 1.2.1.

Given our proposed linear model $L(x) = mx + b$, consider the observed data point (x_j, y_j). Note that the line passing through (x_j, y_j) that is perpendicular to $L(x)$ is given by

$$m(y - y_j) = -(x - x_j).$$

Assuming $m \neq 0$, we have

$$y = -\frac{x}{m} + \frac{x_j + my_j}{m}$$

and use this to solve for the intersection between the two lines:

$$\begin{aligned} mx + b &= -\frac{x}{m} + \frac{x_j + my_j}{m} \\ (m^2 + 1)x &= x_j + my_j - bm \\ x &= \frac{x_j + my_j - bm}{m^2 + 1} \end{aligned}$$

and then

$$\begin{aligned} y &= m \cdot \frac{x_j + my_j - bm}{m^2 + 1} + b \\ &= \frac{mx_j + m^2 y_j + b}{m^2 + 1}. \end{aligned}$$

Observe that

$$\frac{x_j + my_j - bm}{m^2 + 1} - x_j = \frac{my_j - bm - m^2 x_j}{m^2 + 1}$$

and that

$$\frac{mx_j + m^2 y_j + b}{m^2 + 1} - y_j = \frac{mx_j + b - y_j}{m^2 + 1}.$$

Then the square distance is given by

$$\begin{aligned} \left(\frac{my_j - bm - m^2 x_j}{m^2 + 1}\right)^2 + \left(\frac{mx_j + b - y_j}{m^2 + 1}\right)^2 &= m^2 \cdot \frac{(mx_j + b - y_j)^2}{(m^2 + 1)^2} + \frac{(mx_j + b - y_j)^2}{(m^2 + 1)^2} \\ &= \frac{(mx_j + b - y_j)^2}{m^2 + 1}. \end{aligned}$$

The partial derivative with respect to b has only changed by a multiplicative constant so we still have that $b = \bar{y} - m\bar{x}$. Explicitly,

$$\frac{\partial}{\partial b} \sum_{j=1}^{n} \frac{(mx_j + b - y_j)^2}{m^2 + 1} = \sum_{j=1}^{n} \frac{2}{m^2 + 1} \cdot (mx_j + b - y_j) = \frac{2}{m^2 + 1} \cdot [m\Sigma x + nb - \Sigma y].$$

To address the partial derivative with respect to m, note that

$$\begin{aligned}\frac{\partial}{\partial m}\frac{(mx_j+b-y_j)^2}{m^2+1} &= \frac{2x_j(m^2+1)(mx_j+b-y_j)-2m(mx_j+b-y_j)^2}{(m^2+1)^2}\\ &= 2\cdot\frac{mx_j+b-y_j}{(m^2+1)^2}\cdot(x_j-bm+my_j)\,.\end{aligned}$$

Then

$$\sum_{j=1}^{n} 2\cdot\frac{mx_j+b-y_j}{(m^2+1)^2}\cdot(x_j-bm+my_j)=0$$

if and only if

$$\sum_{j=1}^{n}(mx_j+b-y_j)(x_j-bm+my_j)=0.$$

Observe that

$$(mx_j+b-y_j)(x_j-bm+my_j)=m^2x_jy_j-m^2bx_j+mx_j^2-my_j^2+2mby_j-mb^2+bx_j-x_jy_j$$

which means

$$m^2\Sigma xy-m^2b\Sigma x+m\Sigma x^2-m\Sigma y^2+2mb\Sigma y-nmb^2+b\Sigma x-\Sigma xy=0.$$

We collect the terms with factors of b and simplify. First, we simplify the terms with a single power of b:

$$\begin{aligned}-m^2b\Sigma x+2mb\Sigma y+b\Sigma x &= -m^2(\bar{y}-m\bar{x})\Sigma x+2m(\bar{y}-m\bar{x})\Sigma y+(\bar{y}-m\bar{x})\Sigma x\\ &= -m^2\bar{y}\Sigma x+m^3\bar{x}\Sigma x+2m\bar{y}\Sigma y-2m^2\bar{x}\Sigma y+\bar{y}\Sigma x-m\bar{x}\Sigma x\\ &= \frac{1}{n}\left[m^3\left(\Sigma x\right)^2+2m\left(\Sigma y\right)^2-3m^2\Sigma x\Sigma y+\Sigma x\Sigma y-m\left(\Sigma x\right)^2\right]\end{aligned}$$

Next we simplify the term with a square b:

$$\begin{aligned}-nmb^2 &= -nm(\bar{y}-m\bar{x})^2\\ &= -nm((\bar{y})^2-2m\bar{x}\bar{y}+m^2(\bar{x})^2)\\ &= -nm(\bar{y})^2+2nm^2\bar{x}\bar{y}-nm^3(\bar{x})^2\\ &= \frac{1}{n}\left[-m(\Sigma y)^2+2m^2\Sigma x\Sigma y-m^3(\Sigma x)^2\right]\end{aligned}$$

Hence,

$$-m^2b\Sigma x+2mb\Sigma y+b\Sigma x-nmb^2=\frac{1}{n}\left[m\left(\Sigma y\right)^2-m^2\Sigma x\Sigma y+\Sigma x\Sigma y-m\left(\Sigma x\right)^2\right].$$

Multiplying

$$m^2\Sigma xy-m^2b\Sigma x+m\Sigma x^2-m\Sigma y^2+2mb\Sigma y-nmb^2+b\Sigma x-\Sigma xy=0$$

by n obtains

$$m^2\left(n\Sigma xy-\Sigma x\Sigma y\right)+m\left(n\Sigma x^2-n\Sigma y^2+\left(\Sigma y\right)^2-\left(\Sigma x\right)^2\right)-\left(n\Sigma xy-\Sigma x\Sigma y\right)=0$$

or

$$m^2 + m \cdot \frac{n\Sigma x^2 - n\Sigma y^2 + (\Sigma y)^2 - (\Sigma x)^2}{n\Sigma xy - \Sigma x \Sigma y} - 1 = 0.$$

Let

$$a = -\frac{1}{2} \cdot \frac{n\Sigma x^2 - n\Sigma y^2 + (\Sigma y)^2 - (\Sigma x)^2}{n\Sigma xy - \Sigma x \Sigma y}$$

and notice that the solutions to

$$m^2 + m \cdot \frac{n\Sigma x^2 - n\Sigma y^2 + (\Sigma y)^2 - (\Sigma x)^2}{n\Sigma xy - \Sigma x \Sigma y} - 1 = 0$$

are $m = a \pm \sqrt{a^2 + 1}$.

Though this method does yield results, the computational overhead is greater than the method presented in Section 1.1. One may also imagine how tedious this approach would become when there is more than one independent variable. We will provide some examples to examine potential differences between the methods.

1.2 Examples

Example 1.
The following tables contains observed data and some auxiliary quantities. All values were computed using a spreadsheet program so some rounding errors may occur.

X	Y
1.912947521	23.65814
1.759317921	22.94887519
3.311278431	31.76267182
1.093271003	18.98024112
2.175468756	24.89669254
2.457015287	26.67881422
0.2588599	14.24211082
2.648008717	27.91796173
2.397334955	26.54132455
2.735146011	28.24804238
2.824624841	28.87492603
2.149242246	24.89958201
1.866919364	23.28936564
1.436861463	20.96720411
2.202221549	25.36419647
1.324341565	20.4037108
3.387594944	32.0846165
1.003282658	18.45854908
0.997801647	18.39084681
2.186349182	25.03012205

n	20
ΣX	40.12788796
ΣY	483.6379939
ΣX^2	93.00063303
ΣY^2	12102.0174
ΣXY	1041.618954

Vertical

m	5.705388374
b	12.73464042

Perpendicular

m_1	5.708394706
b_1	12.72860853
m_2	−0.175180598
b_2	24.53338106

In the values underneath the 'Perpendicular' label, $m_1 = a + \sqrt{a^2 + 1}$ and $m_2 = a - \sqrt{a^2 + 1}$ where

$$a = -\frac{1}{2} \cdot \frac{n\Sigma x^2 - n\Sigma y^2 + (\Sigma y)^2 - (\Sigma x)^2}{n\Sigma xy - \Sigma x \Sigma y}$$

as in the derivations. Then $b_1 = \bar{y} - m_1\bar{x}$ and $b_2 = \bar{y} - m_2\bar{x}$. We see that the values m_1 and b_1 are the appropriate linear model for the observed data and that the difference between the standard least squares model is minimal.

Example 2.
We now consider an example where the data are not approximately linearly related as they were in Example 1. The format follows the format in Example 1. Again, all values were computed using a spreadsheet program so some rounding errors may occur.

X	Y
1.912947521	54.03677792
1.759317921	43.08854573
3.311278431	74.00096191
1.093271003	28.76392401
2.175468756	38.93272068
2.457015287	45.88985983
0.2588599	52.27123894
2.648008717	25.41396484
2.397334955	48.45117487
2.735146011	37.80523453
2.824624841	29.72250069
2.149242246	45.55706305
1.866919364	50.00450035
1.436861463	48.97881454
2.202221549	22.65972283
1.324341565	46.92592202
3.387594944	59.66573502
1.003282658	46.83425797
0.997801647	36.29521649
2.186349182	62.50839378

n	20
ΣX	40.12788796
ΣY	897.80653
ΣX^2	93.00063303
ΣY^2	43408.48419
ΣXY	1833.13123

Vertical

m	2.544568191
b	39.78491914

Perpendicular

m_1	97.34937002
b_1	-150.4309042
m_2	-0.01027228
b_2	44.91093675

Example 3.
In this example, we use the same X values as in Example 2 but use $Y = (X-2)^2$. Particularly, there is a dependence relationship between X and Y but it's nonlinear. As usual, all values were computed using a

spreadsheet program so some rounding errors may occur.

X	Y
1.912947521	0.007578134
1.759317921	0.057927863
3.311278431	1.719451124
1.093271003	0.822157474
2.175468756	0.030789284
2.457015287	0.208862973
0.2588599	3.031568849
2.648008717	0.419915297
2.397334955	0.157875066
2.735146011	0.540439657
2.824624841	0.680006128
2.149242246	0.022273248
1.866919364	0.017710456
1.436861463	0.317125012
2.202221549	0.040893555
1.324341565	0.456514321
3.387594944	1.925419729
1.003282658	0.99344546
0.997801647	1.00440154
2.186349182	0.034726018

n	20
ΣX	40.12788796
ΣY	12.48908119
ΣX^2	93.00063303
ΣY^2	19.84224531
ΣXY	22.79170937

Vertical

m	-0.181475445
b	0.988565376

Perpendicular

m_1	1.102954752
b_1	-1.588508176
m_2	-0.906655507
b_2	2.44356259

1.2.1 Generalization to Several Variables

Now suppose we have k independent variables $x_1, x_2, \ldots, x_k$ which we can represent as an k-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. We wish to find a linear model for n observations $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$. Suppose our linear model is

$$L(\mathbf{x}) = \mathbf{m} \bullet \mathbf{x} + b = m_1 x_1 + m_2 x_2 + \cdots + m_k x_k + b.$$

Then the method of least squares asks us to minimize

$$\sum_{j=1}^{n} (\mathbf{m} \bullet \mathbf{x}_j + b - y_j)^2$$

where $\mathbf{x}_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,k})$. Then, for $\ell = 1, 2, \ldots, k$,

$$\frac{\partial}{\partial m_\ell} \sum_{j=1}^{n} (\mathbf{m} \bullet \mathbf{x}_j + b - y_j)^2 = 2 \sum_{j=1}^{n} x_{j,\ell} (\mathbf{m} \bullet \mathbf{x}_j + b - y_j)$$

and

$$\frac{\partial}{\partial b} \sum_{j=1}^{n} (\mathbf{m} \bullet \mathbf{x}_j + b - y_j)^2 = 2 \sum_{j=1}^{n} (\mathbf{m} \bullet \mathbf{x}_j + b - y_j).$$

Setting these equal to zero allows us to phrase the problem in terms of a matrix equation as before when we only had one independent variable. Then the problem of the linear regression hinges on finding an inverse for a square matrix.

As an example, let's consider the case when we have two independent variables, x and y. We wish to find constants a, b, and c so that $L(x, y) = ax + by + c$ best fits the set of observations

$$\{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\}.$$

The equations we have now are

$$\frac{\partial}{\partial a}\sum_{j=1}^{n}(ax_j + by_j + c - z_j)^2 = 2\sum_{j=1}^{n} x_j(ax_j + by_j + c - z_j) = 0,$$

$$\frac{\partial}{\partial b}\sum_{j=1}^{n}(ax_j + by_j + c - z_j)^2 = 2\sum_{j=1}^{n} y_j(ax_j + by_j + c - z_j) = 0,$$

and

$$\frac{\partial}{\partial c}\sum_{j=1}^{n}(ax_j + by_j + c - z_j)^2 = 2\sum_{j=1}^{n}(ax_j + by_j + c - z_j) = 0.$$

After some simplification, we obtain

$$\begin{cases} a\Sigma x^2 + b\Sigma xy + c\Sigma x &= \Sigma xz \\ a\Sigma xy + b\Sigma y^2 + c\Sigma y &= \Sigma yz \\ a\Sigma x + b\Sigma y + nc &= \Sigma z \end{cases}$$

which is equivalent to the matrix equation

$$\begin{bmatrix} \Sigma x^2 & \Sigma xy & \Sigma x \\ \Sigma xy & \Sigma y^2 & \Sigma y \\ \Sigma x & \Sigma y & n \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \Sigma xz \\ \Sigma yz \\ \Sigma z \end{bmatrix}$$

As an additional example, the matrix equation corresponding to three independent variables is

$$\begin{bmatrix} \Sigma x^2 & \Sigma xy & \Sigma xz & \Sigma x \\ \Sigma xy & \Sigma y^2 & \Sigma yz & \Sigma y \\ \Sigma xz & \Sigma yz & \Sigma z^2 & \Sigma z \\ \Sigma x & \Sigma y & \Sigma z & n \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} \Sigma xw \\ \Sigma yw \\ \Sigma zw \\ \Sigma w \end{bmatrix}$$

1.3 Coefficient of Determination

For a set $\{y_1, y_2, \ldots, y_n\}$ of observations and a corresponding set of predictions $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, we define

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} = \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2 - \sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}.$$

Notice that, in this computation, if $y_j \approx \hat{y}_j$, then $R^2 \approx 1$. Also, $R^2 \leqslant 1$ for any possible collection of predictions $\hat{y}_j$.

Example 4.
We compute the R^2 values for the data in Examples 1, 2, and 3.

	Example 1		
	Vertical	Perpendicular	
	m, b	m_1, b_1	m_2, b_2
R^2	0.999457187	0.99945691	−0.062317741

	Example 2		
	Vertical	Perpendicular	
	m, b	m_1, b_1	m_2, b_2
R^2	0.026036158	−36.11573083	−0.000210637

	Example 3		
	Vertical	Perpendicular	
	m, b	m_1, b_1	m_2, b_2
R^2	0.034149875	−1.67655224	−0.511162167

2 Logistic Regression

We'll motivate the scenario with a single independent variable. Consider a data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where $y_j \in \{0, 1\}$ for all $j = 1, 2, \ldots, n$. One can think of this as a two-coloring of the real numbers $x_1, x_2, \ldots, x_n$. In particular, let x_j be colored red if $y_j = 1$ and blue otherwise. The goal of logistic regression is to find constants m and b so that

$$p(x) = \frac{\exp(mx + b)}{1 + \exp(mx + b)}$$

is a "good fit" for the data. We'll discuss what we mean by a good fit shortly.

The resulting model produces a way to decide whether new observations should be classified as 0 or 1 in the following way. First, notice that $0 < p(x) < 1$ for any x. We can decide on a threshold p_0 and then, for any observation x, say that x is red if $p(x) > p_0$ and blue otherwise. The choice of threshold will depend on the potential consequences of labeling something as red/blue. However, given the threshold p_0, consider

$$\begin{aligned} p(x) > p_0 &\iff \exp(mx + b) > p_0 + p_0 \exp(mx + b) \\ &\iff (1 - p_0) \exp(mx + b) > p_0 \\ &\iff \exp(mx + b) > p_0/(1 - p_0) \\ &\iff mx + b > \ln(p_0/(1 - p_0)) \\ &\iff mx + b - \ln(p_0/(1 - p_0)) > 0 \end{aligned}$$

That is, depending on the chosen threshold, our decision boils down to checking the sign of a linear function in terms of x.

Now, suppose we have $\{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\}$ where $z_j \in \{0, 1\}$. Like before, we frame this as a two-coloring where (x_j, y_j) is red if $z_j = 1$ and blue otherwise. Recall that in the single variable case, we arrived at a single point which separated the real line into two regions. Here, we will be finding a line in the plane that aims to separate the blue from the red. Note that we can represent such a line with $ax+by+c = 0$. Then, given a point (x, y), we can make a prediction as to red or blue based on the sign of $ax + by + c$. In a similar way to the single variable case, we consider

$$p(x, y) = \frac{\exp(ax + by + c)}{1 + \exp(ax + by + c)}.$$

Notice that if the point (x, y) lies on the line, $ax + by + c = 0$ which means $p(x, y) = 0.5$. If $ax + by + c > 0$, then $p(x, y) > 0.5$ and if $ax + by + c < 0$, then $p(x, y) < 0.5$. The larger $ax + by + c$ is, the closer $p(x, y)$ is to 1. In the discussion that follows, we'll extend to k-many variables.

In one interpretation, the function $p(x)$ is intended to reflect probabilities; that is, $p(x)$ is supposed to be $P(y = 1|x)$. One reason to choose the form of $p(x)$ is for its relative ease in terms of computation. The complication with using the cumulative density for a Gaussian, or normal, distribution is that it's not expressible as a closed form function. Some other candidates are translations of the arctangent and the hyperbolic tangent. Nevertheless, we'll presently entertain the standard choice.

2.1 The Objective

Assume we've chosen a prediction function

$$p(x_1, x_2, \cdots, x_k) = \frac{\exp(m_1x_1 + m_2x_2 + \cdots + m_kx_k + b)}{1 + \exp(m_1x_1 + m_2x_2 + \cdots + m_kx_k + b)}$$

given a set

$$\{(x_{1,1}, x_{1,2}, \ldots, x_{1,k}, y_1), (x_{2,1}, x_{2,2}, \ldots, x_{2,k}, y_2), \ldots, (x_{n,1}, x_{n,2}, \ldots, x_{n,k}, y_n)\}$$

of observations where $y_j \in \{0, 1\}$. Let $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ and $\mathbf{x}_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,k})$ and notice that we can rewrite the prediction function as

$$p(\mathbf{x}) = \frac{\exp(\mathbf{m} \bullet \mathbf{x} + b)}{1 + \exp(\mathbf{m} \bullet \mathbf{x} + b)}.$$

We call the quantities $m_1, m_2, \ldots, m_k, b$ the *weights*.

Now we compare the prediction values against the observed values. For any $y_j = 1$, notice that the prediction model says that $P(y_j = 1|\mathbf{x}_j) = p(\mathbf{x}_j) = p(\mathbf{x}_j)^{y_j}$ and that $(1 - p(\mathbf{x}_j))^{1-y_j} = 1$. For any $y_j = 0$, notice that the prediction model says that

$$P(y_j = 0|\mathbf{x}_j) = 1 - P(y_j = 1|\mathbf{x}_j) = 1 - p(\mathbf{x}_j) = (1 - p(\mathbf{x}_j))^{1-y_j}$$

and that $p(\mathbf{x}_j)^{y_j} = 1$. Assuming all observations are independent, the probability we obtain the observed data assuming the probabilities given by $p(\mathbf{x})$ is

$$J(\mathbf{m}, b) = \prod_{j=1}^{n} p(\mathbf{x}_j)^{y_j} (1 - p(\mathbf{x}_j))^{1-y_j}.$$

Now, this is the function we wish to optimize, in particular, to maximize it since that would mean we maximize the likelihood of the observed data given the predictions determined by $p(\mathbf{x})$. To simplify matters a bit, let's find the optima of the corresponding logarithm:

$$\ln\left(J(\mathbf{m}, b)\right) = \sum_{j=1}^{n} y_j \ln(p(\mathbf{x}_j)) + (1 - y_j) \ln(1 - p(\mathbf{x}_j))$$

Observe that

$$\ln(p(\mathbf{x}_j)) = \ln\left(\frac{\exp(\mathbf{m} \bullet \mathbf{x}_j + b)}{1 + \exp(\mathbf{m} \bullet \mathbf{x}_j + b)}\right) = \mathbf{m} \bullet \mathbf{x}_j + b - \ln(1 + \exp(\mathbf{m} \bullet \mathbf{x}_j + b))$$

and

$$\begin{aligned}
\ln(1 - p(\mathbf{x}_j)) &= \ln\left(1 - \frac{\exp(\mathbf{m} \bullet \mathbf{x}_j + b)}{1 + \exp(\mathbf{m} \bullet \mathbf{x}_j + b)}\right) \\
&= \ln\left(\frac{1}{1 + \exp(\mathbf{m} \bullet \mathbf{x}_j + b)}\right) \\
&= -\ln(1 + \exp(\mathbf{m} \bullet \mathbf{x}_j + b)).
\end{aligned}$$

Hence,

$$\begin{aligned}\ln(J(\mathbf{m},b)) &= \sum_{j=1}^{n} y_j \ln(p(\mathbf{x}_j)) + (1-y_j)\ln(1-p(\mathbf{x}_j)) \\ &= \sum_{j=1}^{n} y_j(\mathbf{m}\bullet\mathbf{x}_j + b - \ln(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b))) + (1-y_j)(-\ln(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b))) \\ &= \sum_{j=1}^{n} \mathbf{m}\bullet\mathbf{x}_j y_j + b y_j - \ln(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b)) \\ &= \left[\sum_{j=1}^{n}\sum_{\ell=1}^{k} m_\ell x_{j,\ell} y_j\right] + b\Sigma y - \sum_{j=1}^{n}\ln(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b)).\end{aligned}$$

Then, for $\ell = 1, 2, \ldots, k$,

$$\frac{\partial}{\partial m_\ell}\ln(J(\mathbf{m},b)) = \left[\sum_{j=1}^{n} x_{j,\ell} y_j\right] - \sum_{j=1}^{n}\frac{x_{j,\ell}\exp(\mathbf{m}\bullet\mathbf{x}_j+b)}{1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b)} = \sum_{j=1}^{n} x_{j,\ell}(y_j - p(\mathbf{x}_j))$$

and

$$\frac{\partial}{\partial b}\ln(J(\mathbf{m},b)) = \Sigma y - \sum_{j=1}^{n}\frac{\exp(\mathbf{m}\bullet\mathbf{x}_j+b)}{1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b)} = \sum_{j=1}^{n} y_j - p(\mathbf{x}_j).$$

Setting these equal to zero doesn't yield closed form solutions, in general. We will require numerical techniques.

Before we discuss a method to approximate solutions, let's check concavity by computing the second derivatives. We check concavity since it communicates information concerning local extrema. First, check that, for $\ell = 1, 2, \ldots, k$,

$$\begin{aligned}\frac{\partial p}{\partial m_\ell} &= \frac{x_\ell(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))\exp(\mathbf{m}\bullet\mathbf{x}+b) - x_\ell\exp(\mathbf{m}\bullet\mathbf{x}+b)^2}{(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))^2} \\ &= \frac{x_\ell\exp(\mathbf{m}\bullet\mathbf{x}+b)}{(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))^2}\end{aligned}$$

and that

$$\begin{aligned}\frac{\partial p}{\partial b} &= \frac{(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))\exp(\mathbf{m}\bullet\mathbf{x}+b) - \exp(\mathbf{m}\bullet\mathbf{x}+b)^2}{(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))^2} \\ &= \frac{\exp(\mathbf{m}\bullet\mathbf{x}+b)}{(1+\exp(\mathbf{m}\bullet\mathbf{x}+b))^2}.\end{aligned}$$

It follows that

$$\frac{\partial^2}{\partial m_\ell^2}\ln(J(\mathbf{m},b)) = -\sum_{j=1}^{n}\frac{x_{j,\ell}^2\exp(\mathbf{m}\bullet\mathbf{x}_j+b)}{(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b))^2}$$

for $\ell = 1, 2, \ldots, k$ and that

$$\frac{\partial^2}{\partial b^2}\ln(J(\mathbf{m},b)) = -\sum_{j=1}^{n}\frac{\exp(\mathbf{m}\bullet\mathbf{x}_j+b)}{(1+\exp(\mathbf{m}\bullet\mathbf{x}_j+b))^2}.$$

Unless all $\mathbf{x}_j$ are zero, the second derivative with respect to any of the weights are negative. Hence, the function $\ln(J(\mathbf{m},b))$, when restricted to one weight, would attain a unique maximum, if a maximum exists.

2.2 Gradient Descent

Recall that the direction of most rapid ascent is the gradient. This section is titled "Gradient Descent" due to the most common name of this technique which is used to find minima. However, since we're trying to find a maximum here, we will be using gradient ascent.

The basic idea is the following.

1. Initialize $(m_1, m_2, \ldots, m_k, b)$ to be arbitrary.
2. Adjust the $(m_1, m_2, \ldots, m_k, b)$ by moving along the gradient by some scaling constant ρ. That is, set the new $(m_1, m_2, \ldots, m_k, b)$ by

$$m_\ell^{\text{new}} = m_\ell + \rho \cdot \sum_{j=1}^{n} x_{j,\ell}(y_j - p(\mathbf{x}_j))$$

and

$$b^{\text{new}} = b + \rho \cdot \sum_{j=1}^{n} y_j - p(\mathbf{x}_j)$$

The task of finding an efficient scaling constant ρ will not be addressed here.

2.3 Examples

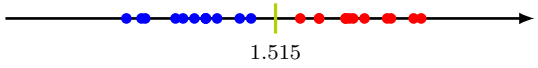
Example 5.
Suppose we are given the following observations:

x	y	x	y	x	y	x	y
1.17	0	1.27	0	1.16	0	1.45	0
1.33	0	1.12	0	1.36	0	1.25	0
1.33	0	1.30	0	1.42	0	1.33	0
1.72	1	1.70	1	1.88	1	1.71	1
1.58	1	1.81	1	1.70	1	1.63	1
1.82	1	1.75	1	1.63	1	1.90	1

We use a scaling coefficient of 1 in the gradient ascent and apply 100 iterations. The equation produced is

$$0 = 50.54714424495551x - 76.56101204614309$$

which translates to $x \approx 1.5146456479345755$. Graphically,



Example 6.
Suppose we are given the following observations:

x	y	z	x	y	z	x	y	z	x	y	z	x	y	z
0.94	0.99	0	1.41	0.62	1	1.05	1.19	0	0.79	1.43	0	1.51	0.64	1
1.42	1.28	1	1.19	1.48	0	0.67	1.21	0	0.72	0.62	0	1.39	0.67	1
0.89	0.87	0	0.91	0.74	0	1.24	1.13	1	0.81	0.95	0	1.11	1.55	0
1.29	1.11	1	1.54	0.81	1	0.98	0.97	0	1.23	0.88	1	0.61	0.60	0
1.26	0.82	1	1.38	0.76	1	1.32	1.20	1	1.06	1.14	0	1.14	1.39	0

We use a scaling coefficient of 1 in the gradient ascent and apply 100 iterations. The equation produced is

$$0 = 47.49556536544975x - 28.459417864950904y - 24.237777075778855.$$

Graphically,