Logistic Regression

Likelihood Functions

Many probability distributions have **unknown parameters**; we estimate these unknowns using sample data. The **likelihood function** gives us an idea of how well the data "supports" these parameters.

More formally:

Let $X_1, X_2, \ldots, X_n$ have a joint density function $f(X_1, X_2, \ldots, X_n|\theta)$. Given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is observed, the likelihood function is given by:

$$L(\theta) = L(\theta|x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta)$$

It'll be easier to understand with an example, but for now note that:

- In the probability density function f, $X_1, X_2, \ldots X_n$ are varying and θ is fixed.
- In the likelihood function, θ is varying but $x_1, x_2, \ldots, x_n$ (the observations) are fixed.

Consider a simple experiment involving a (potentially) biased coin. We can express the probability of flipping heads with a Bernoulli random variable:

$$p_X(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}$$

where $x = 1$ represents flipping heads and $x = 0$ represents flipping tails. θ is an **unknown parameter** that we would like to estimate.

Suppose we make four coin tosses independently of each other, and get the result $\{H, H, T, H\}$. Then our likelihood function looks like:

$$\begin{aligned} L(\theta) &= \theta \times \theta \times (1 - \theta) \times \theta \\ &= \theta^3(1 - \theta) \\ &= \theta^3 - \theta^4 \end{aligned}$$

Let's graph this function:

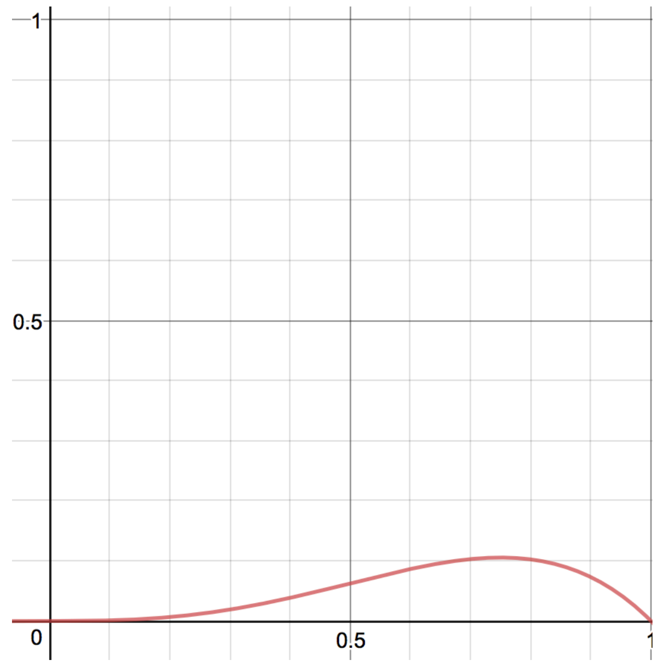


Figure 1: The graph of $L(\theta) = \theta^3 - \theta^4$.

So what does this likelihood function represent?

- The likelihood function **is not a probability density function**. It **does not** represent the probability that θ has a given value.
- Rather, it measures the support provided by the data for each possible value of the parameter.
  - Consider two possible values of theta, θ_1 and θ_2. If we find that $L(\theta_1) > L(\theta_2)$, all this means is that the sample **we have already observed** is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$.
  - **Essentially, this can be interpreted as θ_1 is a more plausible value for θ than θ_2.**

Maximum Likelihood Estimator (MLE)

Returning to the example above, in this case, it is clear that $L(\theta)$ has a critical point which is also a global maximum. (However, this global maximum may not always exist). We call this global maximum the **maximum likelihood estimator** and often notate it as $\hat{\theta}$.

To find the MLE, we use the same process as any other plain optimization problem: taking the first derivative and setting it to 0.

In our example,

$$
\begin{aligned}
L'(\theta) &= 3\theta^2 - 4\theta^3 \\
0 &= 3\hat{\theta}^2 - 4\hat{\theta}^3 \\
4\hat{\theta}^3 - 3\hat{\theta}^2 &= 0 \\
\hat{\theta}^2(4\hat{\theta} - 3) &= 0 \\
\therefore \hat{\theta} &= \frac{3}{4}. \quad (0 \text{ is an extraneous solution})
\end{aligned}
$$

This answer is intuitive: If I perform 4 independent coin tosses and 3 come up heads and 1 tails, the probability of getting a head **is mostly likely** to be 3/4, assuming it is unknown.

Back to Logistic Regression...

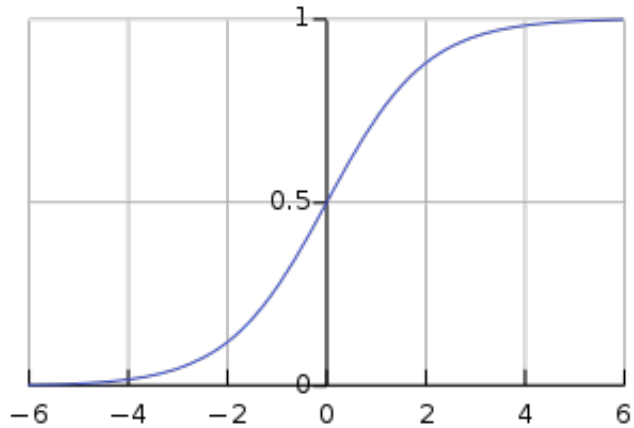
Let's first consider the simple case of a binary classifier. Formally speaking, for any training sample $\vec{x}$, $y \in \{0, 1\}$.

Our model will output a real number in the range of $0...1$ which will be interpreted as the probability that a given training sample $\vec{x}$ will belong to the class $y = 1$.

Sigmoid Function

To do this, the most common "activation function" used is called the **sigmoid function**. This function outputs a value between 0 and 1 as required and is defined as follows:

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



What will we use as our z? The same hypothesis we used for linear regression:

$$
z = \mathbf{X}\vec{\theta} = \begin{bmatrix} X_{1,1}\theta_1 + X_{1,2}\theta_2 + \cdots + X_{1,n}\theta_n \\ X_{2,1}\theta_1 + X_{2,2}\theta_2 + \cdots + X_{2,n}\theta_n \\ \cdots \\ X_{m,1}\theta_1 + X_{m,2}\theta_2 + \cdots + X_{m,n}\theta_n \end{bmatrix}
$$

So for a single training example $\vec{x}$, our prediction will be:

$$\sigma(\vec{\theta}) = \frac{1}{1 + e^{\vec{x} \cdot \vec{\theta}}}$$

Deriving an Objective Function

Given a set of observations $\{\mathbf{X}, \vec{y}\}$, how can we learn $\vec{\theta}$ such that we **maximize the likelihood** that these observations support $\vec{\theta}$? The MLE is the solution.

Likelihood Function

Just like in the example earlier, we can express the probability that a training sample $\vec{x}$ belongs to the class $y = 1$ using a Bernoulli random variable:

$$p_X(\vec{x}|\theta) = \begin{cases} \sigma(\vec{\theta}) & \text{if } y = 1 \\ 1 - \sigma(\vec{\theta}) & \text{if } y = 0 \end{cases}$$

If we have m training samples, then:

$$L(\vec{\theta}|\vec{x}) = \prod_{i=1}^{m} \sigma(\vec{\theta})^{y_i} (1 - \sigma(\vec{\theta}))^{1-y_i}$$

Here, we have used a "trick" involving the fact that $y_i \in \{0, 1\} \ \forall \ i$. For the i^{th} training sample:

- If $y_i = 1$, we have:

$$\begin{aligned} \sigma(\vec{\theta})^{y_i} (1 - \sigma(\vec{\theta}))^{1-y_i} &= \sigma(\vec{\theta})^{1} (1 - \sigma(\vec{\theta}))^{1-1} \\ &= \sigma(\vec{\theta}) \end{aligned}$$

- If $y_i = 0$, we have:

$$\begin{aligned} \sigma(\vec{\theta})^{y_i} (1 - \sigma(\vec{\theta}))^{1-y_i} &= \sigma(\vec{\theta})^{0} (1 - \sigma(\vec{\theta}))^{1-0} \\ &= 1 - \sigma(\vec{\theta}) \end{aligned}$$

as required by our random variable above.

Log Likelihood Function

We would like to maximize the value of $L(\vec{\theta})$. To make equations simpler, we instead equivalently maximize the **log likelihood**, $\log L(\vec{\theta})$. This will allow us to transform multiplications into additions and allows us to "bring down" exponents.

$$
\begin{aligned}
\log L(\vec{\theta}|\vec{x}) &= \log \prod_i \sigma(\vec{\theta})^{y_i}(1-\sigma(\vec{\theta}))^{1-y_i} \\
&= \sum_i \log\left(\sigma(\vec{\theta})^{y_i}(1-\sigma(\vec{\theta}))^{1-y_i}\right) \\
&= \sum_i \log \sigma(\vec{\theta})^{y_i} + \log(1-\sigma(\vec{\theta}))^{1-y_i} \\
&= \sum_i y_i \log \sigma(\vec{\theta}) + (1-y_i)\log(1-\sigma(\vec{\theta}))
\end{aligned}
$$

Maximum Likelihood Estimator

To calculate the MLE, we need one additional lemma:

Proposition: The first derivative of $\sigma(z)$ is $\sigma(z)(1-\sigma(z))$.

Proof:

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

Let $x = 1 + e^{-z}$. Then:

$$
\begin{aligned}
\sigma(z) &= \frac{1}{x} \\
\frac{d\sigma}{dz} &= -\frac{1}{x^2}\frac{dx}{dz} \\
&= -\frac{1}{(1+e^{-z})^2} \times -e^{-z} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}-1}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}}{(1+e^{-z})^2} - \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} - \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) \\
&= \sigma(z)(1-\sigma(z)).
\end{aligned}
$$

We now have all we need to calculate the gradient of our objective function:

$$
\begin{aligned}
\frac{\partial L}{\partial \vec{\theta}} &= \sum_i y_i \left(\frac{1}{\sigma(z)} \big(\sigma(z)(1-\sigma(z))\big) \frac{\partial z}{\partial \vec{\theta}} \right) + (1-y_i) \left(\frac{1}{1-\sigma(z)} \big(-(\sigma(z)(1-\sigma(z)))\big) \frac{\partial z}{\partial \vec{\theta}} \right) \\
&= \sum_i y_i (1-\sigma(z)) \overrightarrow{x_i} + (1-y_i)(-\sigma(z)) \overrightarrow{x_i} \\
&= \sum_i \big(y_i - y_i \sigma(z) - \sigma(z) + y_i \sigma(z) \big) \overrightarrow{x_i} \\
&= \sum_i \big(y_i - \sigma(z) \big) \overrightarrow{x_i}
\end{aligned}
$$

Matrix Version

$$
J(\vec{\theta}) = \vec{y}^T \log \sigma(\mathbf{X}\vec{\theta}) + (\vec{1} - \vec{y})^T \log(\vec{1} - \sigma(\mathbf{X}\vec{\theta}))
$$

$$
\nabla J(\vec{\theta}) = \mathbf{X}^T (\vec{y} - \sigma(\mathbf{X}\vec{\theta}))
$$

$$
\begin{aligned}
\nabla J(\vec{\theta}) &= \vec{y}^T \frac{1}{\sigma(\mathbf{X}\vec{\theta})} \sigma(\mathbf{X}\vec{\theta})(\vec{1} - \sigma(\mathbf{X}\vec{\theta}))\mathbf{X} + (\vec{1} - \vec{y})^T \frac{1}{1 - \sigma(\mathbf{X}\vec{\theta})} (-(\sigma(\mathbf{X}\vec{\theta})(\vec{1} - \sigma(\mathbf{X}\vec{\theta}))))(\mathbf{X}) \\
&= \vec{y}^T (\vec{1} - \sigma(\mathbf{X}\vec{\theta}))\mathbf{X} + (\vec{1} - \vec{y})^T (-\sigma(\mathbf{X}\vec{\theta}))(\mathbf{X}) \\
&= (\vec{y}^T \vec{1} - \vec{y}^T \sigma(\mathbf{X}\theta))\mathbf{X} + (-\vec{1}^T \sigma(\mathbf{X}\vec{\theta}) + \vec{y}^T \sigma(\mathbf{X}\vec{\theta}))(\mathbf{X}) \\
&= \vec{y}^T \vec{1}\mathbf{X} - \vec{y}^T \sigma(\mathbf{X}\vec{\theta})\mathbf{X} - \vec{1}^T \sigma(\mathbf{X}\vec{\theta})\mathbf{X} + \vec{y}^T \sigma(\mathbf{X}\vec{\theta})\mathbf{X} \\
&= \vec{y}^T \vec{1}\mathbf{X} - \vec{1}^T \sigma(\mathbf{X}\vec{\theta})\mathbf{X} \\
&= \mathbf{X}^T \big(\vec{y} - \sigma(\mathbf{X}\vec{\theta})\big)
\end{aligned}
$$