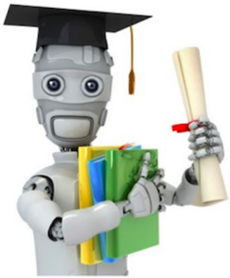
Logistic Regression

Collin Prather

April 25th, 2018



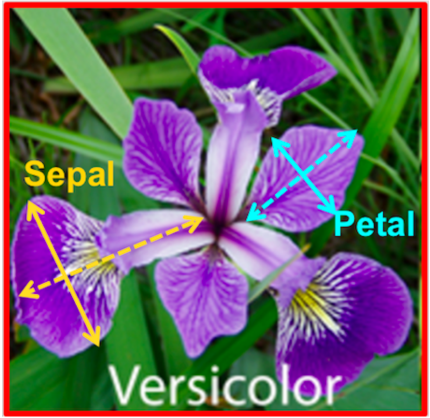
Machine Learning

Arthur Samuel

Machine learning is “Field of study that gives computers the ability to learn without being explicitly programmed”.

Class 0

Class 1



Iris Flower Data

	Sepal Length (cm)	Sepal Width (cm)	Target
0	5.1	3.5	0
1	4.9	3.0	0
2	4.7	3.2	0
3	4.6	3.1	0
4	5.0	3.6	0
5	5.4	3.9	0
6	4.6	3.4	0
7	5.0	3.4	0
8	4.4	2.9	0
9	4.9	3.1	0

Logistic regression as a linear classifier

- The linear combination z,

$$z = \theta_1 x_1 + \theta_2 x_2,$$

Logistic regression as a linear classifier

- The linear combination z,

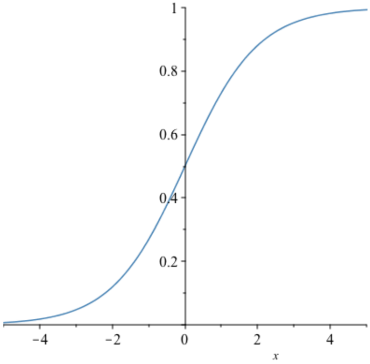
$$z = \theta_1 x_1 + \theta_2 x_2,$$

- can be represented as the matrix equation

$$\underbrace{\begin{bmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ \vdots & \vdots \\ x_1^{150} & x_2^{150} \end{bmatrix}}_{150\times 2} \cdot \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}}_{2\times 1} = \underbrace{\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}}_{150\times 1}$$

Hypothesis Representation

$$\text{Sigmoid/Logistic Function: } g(z) = \frac{1}{1 + e^{-z}}$$



Hypothesis Interpretation

Predictions interpreted as Probabilities

$$h_\theta(x) = g(z) = \frac{1}{1 + e^{-z}}$$

Hypothesis Interpretation

Predictions interpreted as Probabilities

$$h_\theta(x) = g(z) = \frac{1}{1 + e^{-z}}$$
$$= p(y = 1 | x; \theta)$$

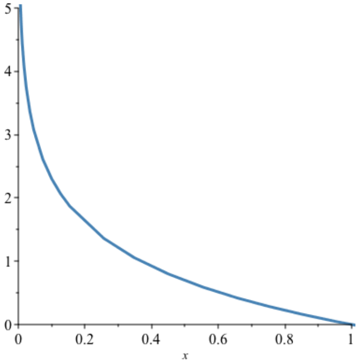
1. Logistic Regression in the Context of Machine Learning
2. Logistic Regression Introduction
3. Linear Algebra Orientation
4. Hypothesis Function
   - Representation
   - Interpretation
5. Cost Function
   - When $y = 1$
   - When $y = 0$
6. Gradient Descent

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

- the domain of $\mathrm{Cost}(h_\theta(x), y)$ is $[0, 1]$.

When y = 1

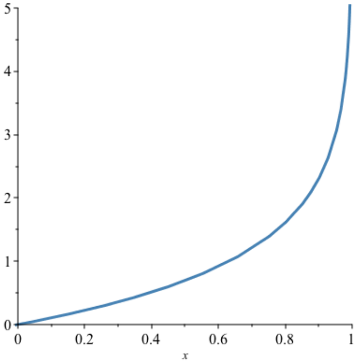
Figure: Cost($h_\theta(x), y$) when y = 1



$$\text{as } h_\theta(x) \to 1, \ \text{Cost}(h_\theta(x), y) \to 0$$
$$\text{as } h_\theta(x) \to 0, \ \text{Cost}(h_\theta(x), y) \to \infty$$

When y = 0

Figure: $\text{Cost}(h_\theta(x), y)$ when y = 0



$$\text{as} \quad h_\theta(x) \to 1, \ \text{Cost}(h_\theta(x), y) \to \infty$$
$$\text{as} \quad h_\theta(x) \to 0, \ \text{Cost}(h_\theta(x), y) \to 0$$

Simplified Cost Function

Previously, we had defined $\text{Cost}(h_\theta(x), y)$ by cases like so:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Simplified Cost Function

Previously, we had defined $\text{Cost}(h_\theta(x), y)$ by cases like so:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

However, it can be compressed:

$$\text{Cost}(h_\theta(x), y) = \underbrace{-y\log(h_\theta(x))}_{=0 \text{ when } y=0} \underbrace{-(1-y)\log(1-h_\theta(x))}_{=0 \text{ when } y=1}$$

Gradient Descent

Cost Over Entire Dataset

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Gradient Descent

Cost Over Entire Dataset

$$
\begin{aligned}
J(\theta) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\right]
\end{aligned}
$$

Gradient Descent

Iterative Gradient Descent Algorithm

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

The Partial Derivative Term

- When we calculate the partial derivative term $\frac{\partial}{\partial\theta_j}J(\theta)$, we find that

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\left(\frac{1}{m}\left[\sum_{i=1}^{m}\text{Cost}(h_\theta(x^{(i)}), y^{(i)})\right]\right)\\ &= \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x^{(i)}\end{aligned}$$

The Partial Derivative Term

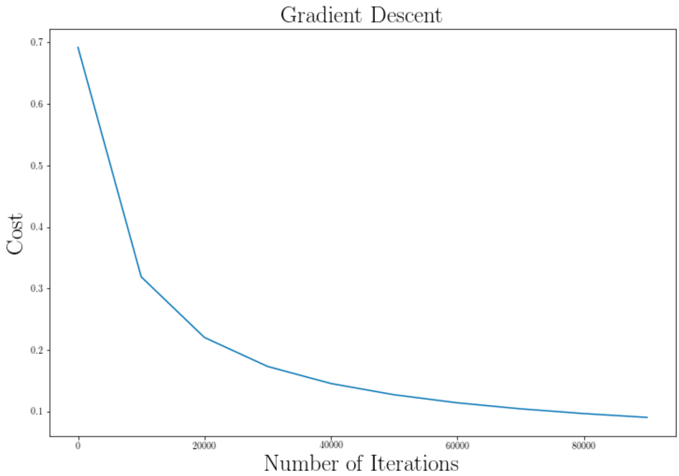
- When we calculate the partial derivative term $\frac{\partial}{\partial \theta_j} J(\theta)$, we find that

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \left(\frac{1}{m} \left[\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) \right] \right)$$
$$= \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x^{(i)}$$

Updated Gradient Descent Algorithm

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

The Learning Rate α



Batch Gradient Descent

$$\theta_1 := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_1^{(i)}$$

$$\theta_2 := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_2^{(i)}$$