



On Optimizing a Logistic Regressor

Alexander G. Ororbia II
Introduction to Machine Learning
CSCI-335
10/21/2026

Cost Function for Logistic Regression

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

**But, where does
it come from?**



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

- If $y = 1$: $\text{Cost}(h_{\theta}(x), y) = -\log(h_{\theta}(x))$
- If $y = 0$: $\text{Cost}(h_{\theta}(x), y) = -\log(1 - h_{\theta}(x))$

Logistic Regression

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

Learning: fit parameter θ

$$\min_{\theta} J(\theta)$$

Prediction: given new x

$$\text{Output } h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where Does the **Cost** Come From?

- Training set with m examples

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

- Maximum likelihood estimate for parameter θ

$$\begin{aligned}\theta_{\text{MLE}} &= \underset{\theta}{\operatorname{argmax}} P_{\theta} \left((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)}) \right) \\ &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^m P_{\theta} \left((x^{(i)}, y^{(i)}) \right)\end{aligned}$$

- Maximum conditional likelihood estimate for parameter θ

Remember the Bernoulli Distribution?

$$P(\mathbf{x} = 1) = \phi$$

$$P(\mathbf{x} = 0) = 1 - \phi$$

$$P(\mathbf{x} = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathbf{x}}[\mathbf{x}] = \phi$$

$$\text{Var}_{\mathbf{x}}(\mathbf{x}) = \phi(1 - \phi)$$

Remember the Bernoulli Distribution?

$$P(x = 1) = \phi$$

$$P(x = 0) = 1 - \phi$$

Bernoulli probability
mass function
(likelihood)

$$P(x = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_x[x] = \phi$$

$$\text{Var}_x(x) = \phi(1 - \phi)$$

- **Goal:** choose θ to maximize conditional likelihood of training data

- $P_{\theta}(Y = 1|X = x) = h_{\theta}(x) = \frac{1}{1+e^{-\theta^T x}}$
 - $P_{\theta}(Y = 0|X = x) = 1 - h_{\theta}(x) = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}$

- **Training data** $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
- **Data likelihood** $= \prod_{i=1}^m P_{\theta}((x^{(i)}, y^{(i)}))$
- **Data conditional likelihood** $= \prod_{i=1}^m P_{\theta}(y^{(i)}|x^{(i)})$

$$\theta_{\text{MCLE}} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^m P_{\theta}(y^{(i)}|x^{(i)})$$

Expressing Conditional Log-Likelihood

$$L(\theta) = \log \prod_{i=1}^m P_{\theta}(y^{(i)} | x^{(i)}) = \sum_{i=1}^m \log P_{\theta}(y^{(i)} | x^{(i)})$$

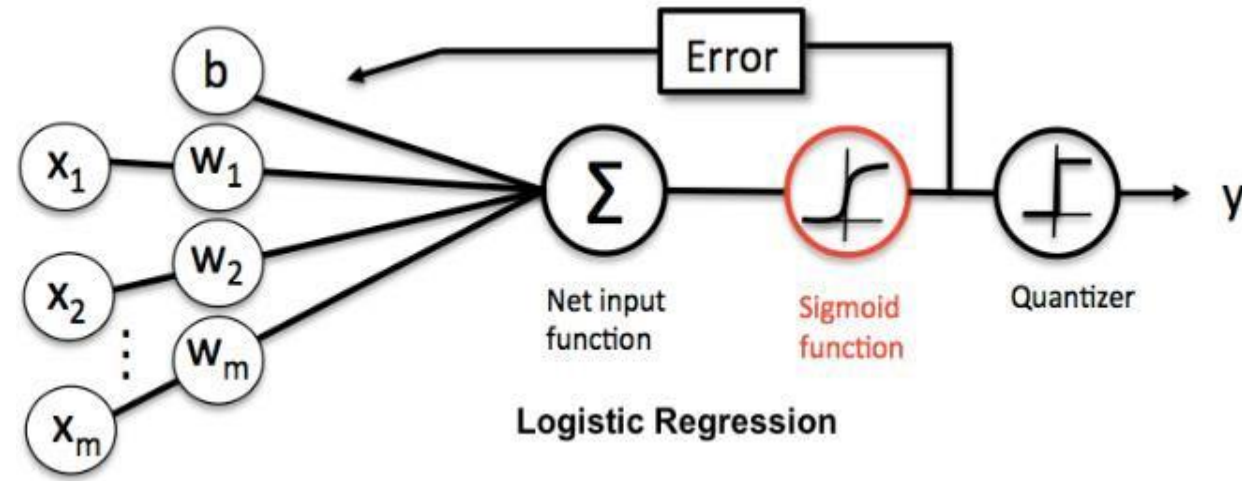
The logarithm is
your *friend*!

$$= \sum_{i=1}^m y^{(i)} \log P_{\theta}(y^{(i)} = 1 | x^{(i)}) + (1 - y^{(i)}) \log P_{\theta}(y^{(i)} = 0 | x^{(i)})$$

$$= \sum_{i=1}^m y^{(i)} \log (h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))$$

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

Logistic Regression



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

Bernoulli log likelihood!

Learning: fit parameter θ

$$\min_{\theta} J(\theta)$$

Prediction: given new x

$$\text{Output } h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

(Evaluation!)

Gradient Descent

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

Goal: $\min_{\theta} J(\theta)$

Good news: Convex function!

Bad news: No analytical solution

Repeat {

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

}

What do we still need?

Derive gradient of $J(\theta_j)$ with respect to each θ_j

(Optimization!)



Derivation of gradient of $J(\theta)$:

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

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

Gradient Descent

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

Goal: $\min_{\theta} J(\theta)$

Good news: Convex function!

Bad news: No analytical solution

Repeat {

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

}

(Simultaneously update all θ_j)

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Gradient Descent

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \right]$$

Goal: $\min_{\theta} J(\theta)$

Repeat { ***(Simultaneously*** update all θ_j)

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

}

(Optimization!)

Gradient Descent for *Linear* Regression

Repeat {

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

}

Gradient Descent for *Logistic* Regression

Repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad \boxed{h_{\theta}(x) = \frac{1}{1 + e^{-\theta^{\top} x}}}$$

}

Gradient Descent for *Linear* Regression

Repeat {

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

}

$$h_{\theta}(x) = \theta^{\top} x$$

Woah, they're the same!!

Gradient D

Logistic Regression

Repeat {

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

}

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

There Are More Forms of Regression at Your Fingertips

- Poisson regression $\lambda := E(Y \mid x) = e^{\theta'x}$

$$p(y_1, \dots, y_m \mid x_1, \dots, x_m; \theta) = \prod_{i=1}^m \frac{e^{y_i \theta' x_i} e^{-e^{\theta' x_i}}}{y_i!}$$

- Exponential regression $\lambda_i = \exp(x_i' \beta)$

$$\prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i} \quad l(\lambda) = n \ln(\lambda) - \lambda \sum x_i$$

- Other distributions and likelihoods/densities
 - You just need to be able to write out the density/likelihood, take the log-likelihood, and take derivatives of that log-likelihood with respect to your regression parameters Θ

Questions?

Deep robots!

Deep questions?!

