Loss Functions in Logistic Regression

Why **Squared Error** is Problematic and How **Log Loss** Solves the Issue

Understanding Loss and Cost Functions

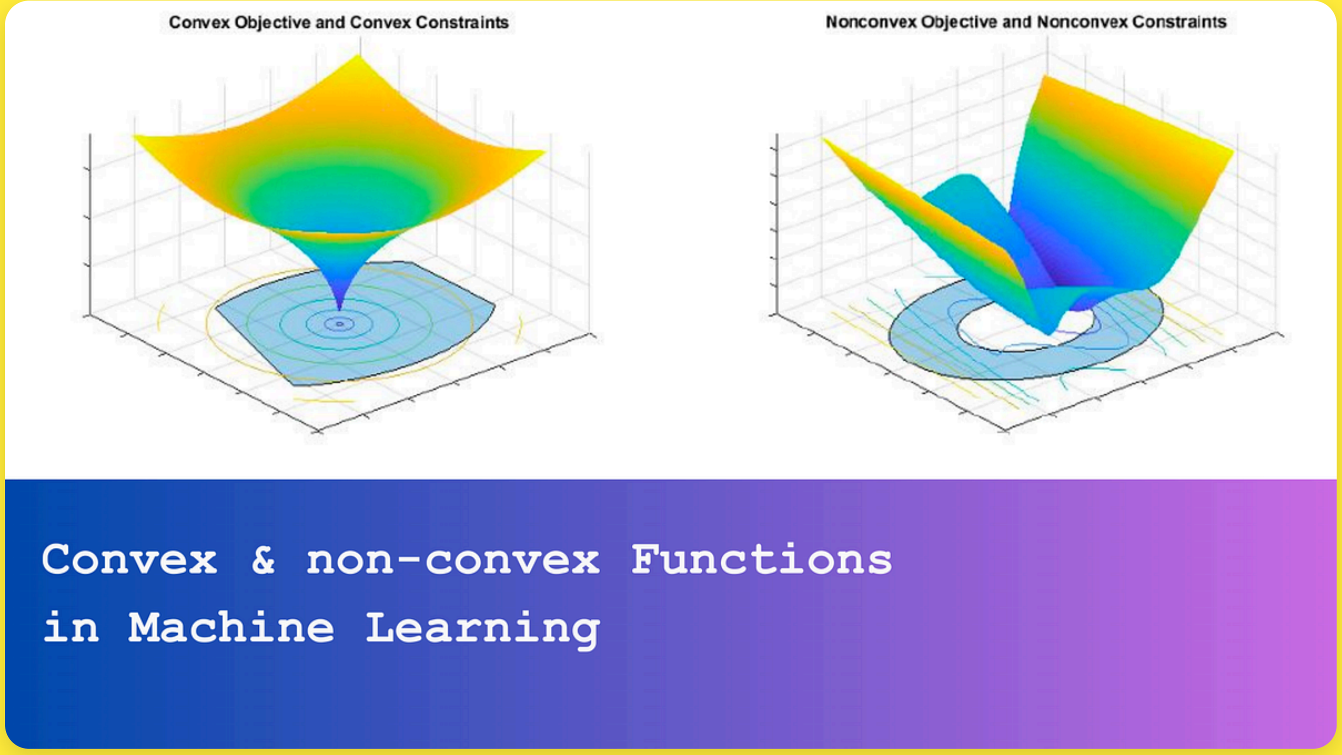
Loss Function

Measures the error for a single example

Cost Function

The average over all examples

Convex vs. Non-Convex Functions



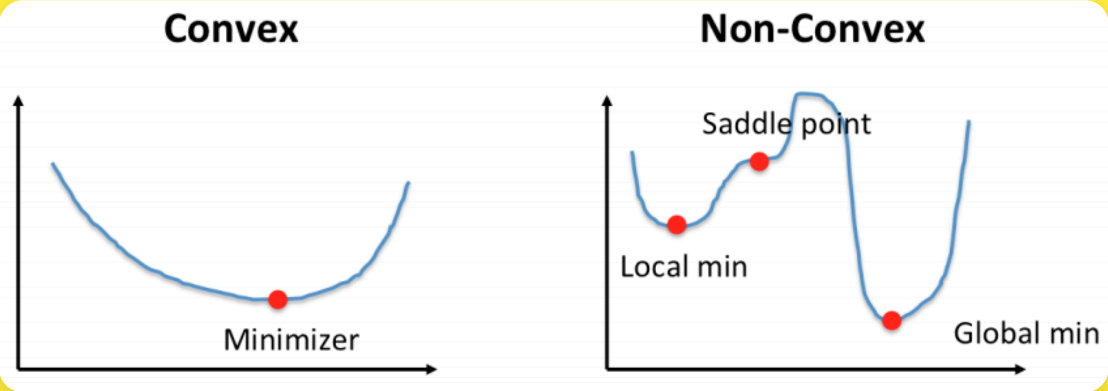
Convex Function

- Shape like a **"bowl"** or "basin"
- Single, unique **global optimum**

Non-Convex Function

- Shape with **irregularities**, peaks, and valleys
- Contains many **local optima**

Problems with Non-Convex Functions



Gradient Descent Algorithm

- Starts from a**random point**
- Moves**"downhill"**in the direction of the gradient
- Reaches a**minimum point**

Convex Function

Always reaches **global optimum**

Non-Convex Function

Might get stuck in **local optimum**

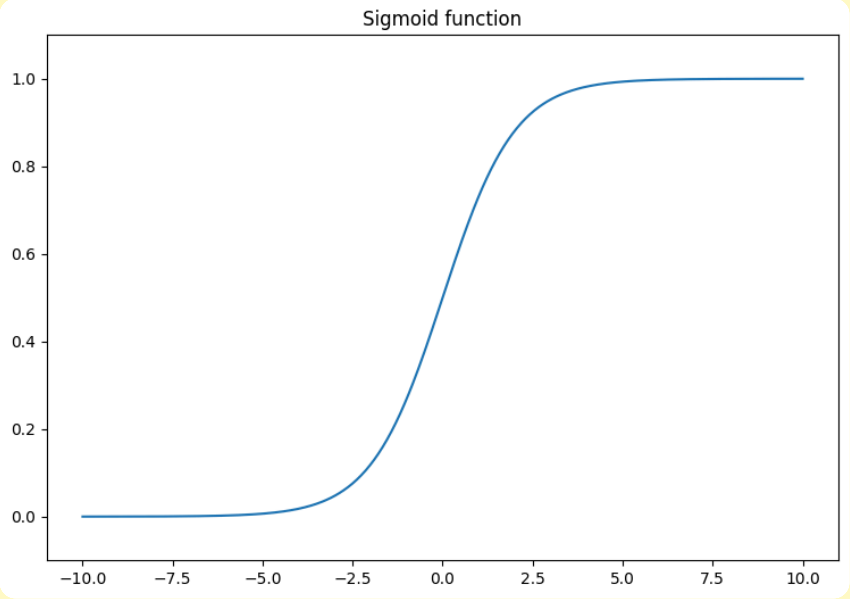
Why Squared Error is Problematic

The Problem with Squared Error

- Creates **complex mathematical expression** due to non-linear sigmoid function
- Makes the cost function **non-convex** (irregularities and valleys)

Gradient Descent might stop at a local solution instead of finding the best one

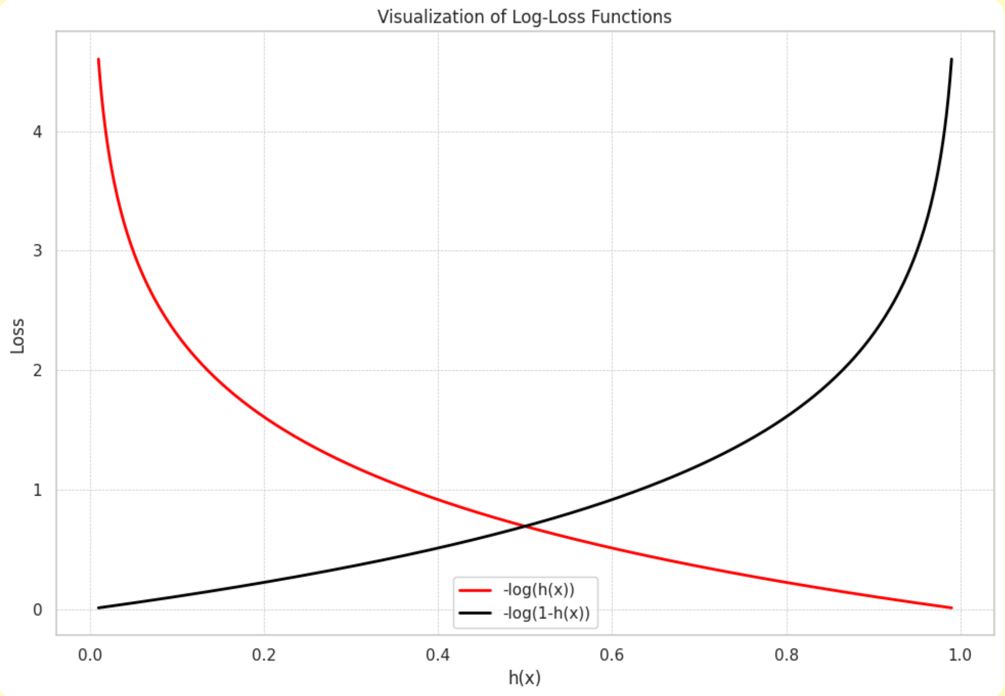
Sigmoid Function in Logistic Regression



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

Solution: Log Loss Function

Log Loss Function Visualization



The Solution

Use **Log Loss** function instead of Squared Error

Gives a **convex shape** with respect to W and B

Key Benefits

- Only**one global solution**
- Gradient Descent**always succeeds**in finding it
- Requires only appropriate**learning rate**

Mathematical Formulas

Σ Loss Function

$$L(\hat{y}, y) = -[y \log(\hat{y}) + (1-y) \log(1-\hat{y})]$$

Measures the error for a **single training example**

Variables:

- **y**: True label (0 or 1)
- **ŷ**: Predicted probability (σ(z))

Cost Function

$$J(w,b) = -1/m \sum[y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)})]$$

Average over the entire training set

Variables:

- **m**: Number of training examples
- **$y^{(i)}$**: True label for i^{th} example
- **$\hat{y}^{(i)}$**: Predicted probability for i^{th} example

Key Takeaways

Summary

- Squared Error causes**non-convexity**in logistic regression due to the sigmoid function
- Non-convex functions have**multiple local optima**where gradient descent can get stuck
- Log Loss provides a**convex function**with a single global optimum
- This ensures gradient descent can find the**optimal solution**for logistic regression models

Using the right loss function is crucial for the success of machine learning models