Mathematical Description of Logistic Regression

Logistic regression is a statistical model used for binary classification, extendable to multiclass problems via techniques like softmax regression. It predicts the probability that a given input belongs to a particular class. Below is the mathematical formulation.

1. Model Representation

For a binary classification problem, the goal is to predict the probability $P(y = 1|\mathbf{x})$, where $y \in \{0, 1\}$ is the class label, and $\mathbf{x} \in \mathbb{R}^n$ is the feature vector. The logistic regression model assumes this probability follows the logistic (sigmoid) function:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

- $\mathbf{w} \in \mathbb{R}^n$: Weight vector (parameters to be learned).
- $b \in \mathbb{R}$: Bias term (intercept).
- $\mathbf{w}^T\mathbf{x} + b$: Linear combination of features, often denoted as z.
- $\sigma(z) = \frac{1}{1+e^{-z}}$: Sigmoid function, mapping $z \in \mathbb{R}$ to $[0, 1]$.

The probability of the negative class is:

$$P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}) = \frac{e^{-(\mathbf{w}^T\mathbf{x}+b)}}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}.$$

2. Decision Rule

To classify an input $\mathbf{x}$, a threshold (typically 0.5) is applied to the predicted probability:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1|\mathbf{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Since $P(y = 1|\mathbf{x}) = \sigma(z)$, and $\sigma(z) = 0.5$ when $z = 0$, this is equivalent to:

$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{w}^T\mathbf{x} + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

3. Loss Function

The parameters $\mathbf{w}$ and b are learned by minimizing the **log-loss** (or **binary cross-entropy loss**). For a dataset of m samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, the log-loss is:

$$J(\mathbf{w}, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(P(y = 1|\mathbf{x}^{(i)})) + (1 - y^{(i)})\log(1 - P(y = 1|\mathbf{x}^{(i)}))\right]$$

Substituting $P(y = 1|\mathbf{x}^{(i)}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)} + b)$, the loss becomes:

$$J(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b)) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b)) \right]$$

4. Optimization

The loss function $J(\mathbf{w}, b)$ is convex, so optimization techniques like **gradient descent** are used to find the optimal parameters. The gradients with respect to $\mathbf{w}$ and b are:

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{1}{m} \sum_{i=1}^{m} \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b) - y^{(i)} \right) \mathbf{x}^{(i)}$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b) - y^{(i)} \right)$$

In gradient descent, the parameters are updated iteratively:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial J}{\partial \mathbf{w}}, \quad b \leftarrow b - \alpha \frac{\partial J}{\partial b}$$

where α is the learning rate.

5. Regularization (Optional)

To prevent overfitting, regularization terms (e.g., L2 or L1) can be added to the loss function. For **L2 regularization**, the loss becomes:

$$J(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b)) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b)) \right] + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

- λ: Regularization parameter controlling the strength of the penalty.
- $\|\mathbf{w}\|^2$: L2 norm of the weights (encourages smaller weights).

The gradient for $\mathbf{w}$ is modified to include the regularization term:

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{1}{m} \sum_{i=1}^{m} \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)} + b) - y^{(i)} \right) \mathbf{x}^{(i)} + \lambda \mathbf{w}$$

6. Multiclass Extension (Softmax Regression)

For K-class classification, logistic regression is generalized to **softmax regression**. The model outputs probabilities for each class using the softmax function:

$$P(y = k | \mathbf{x}) = \frac{e^{\mathbf{w}_k^T \mathbf{x} + b_k}}{\sum_{j=1}^{K} e^{\mathbf{w}_j^T \mathbf{x} + b_j}}$$

- $\mathbf{w}_k, b_k$: Parameters for class k.

The loss function is the **categorical cross-entropy**, and optimization proceeds similarly.

7. Summary

Logistic regression models the probability of a binary outcome using the sigmoid function, optimizing parameters via gradient descent on the log-loss. Regularization can be applied to improve generalization. The model is interpretable, computationally efficient, and widely used for binary classification tasks.