Maximum Likelihood Estimation (MLE) for Logistic Regression

Model Setup

We model the probability of a binary target $y_i \in \{0, 1\}$ given features $\mathbf{x}_i \in \mathbb{R}^d$:

$$P(y_i = 1 \mid \mathbf{x}_i; \boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}) = \frac{1}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}}$$

where $\sigma(z)$ is the sigmoid function, and $\boldsymbol{\beta}$ is the parameter vector.

Likelihood Function

Each y_i is a Bernoulli random variable:

$$P(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} \cdot (1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}))^{1-y_i}$$

The likelihood for n i.i.d. samples is:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\sigma(\mathbf{x}_i^\top \boldsymbol{\beta})^{y_i} (1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}))^{1-y_i} \right]$$

Log-Likelihood

To simplify computation, take the log of the likelihood:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})) \right]$$

Gradient of the Log-Likelihood

Let $\sigma_i = \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})$. The gradient is:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (y_i - \sigma_i) \mathbf{x}_i$$

Vectorized form:

$$\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = X^\top (\mathbf{y} - \boldsymbol{\sigma})$$

Where:

- $X \in \mathbb{R}^{n \times d}$ is the feature matrix,
- $\mathbf{y} \in \mathbb{R}^n$ is the label vector,
- $\boldsymbol{\sigma} \in \mathbb{R}^n$ is the vector of predicted probabilities.

MLE Optimization

There is no closed-form solution for $\boldsymbol{\beta}$, so we use iterative optimization methods:

- **Gradient Ascent**:
$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \eta \cdot \nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$
- **Newton-Raphson**
- **Quasi-Newton** methods (e.g., BFGS)

Summary

- Logistic regression models $P(y = 1 \mid \mathbf{x})$ using the sigmoid function.
- The likelihood is based on the Bernoulli distribution.
- MLE aims to maximize the log-likelihood.
- The gradient can be computed analytically.
- Optimization is done numerically (no closed-form solution).