Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary classification problems. It models the probability that an input belongs to a particular class using a linear function combined with a sigmoid transformation. Unlike linear regression, it predicts probabilities instead of continuous values.

1. Problem Formulation

This section defines the classification problem mathematically.

Input Space $\mathcal{X} \subseteq \mathbb{R}^p$ Where Each input vector: $x \in \mathbb{R}^p$

Output Space $\mathcal{Y} = \{0,1\}$ Binary classification output.

Data Distribution

Training samples:

$$(x_i, y_i) \sim P_{X,Y}$$

are assumed i.i.d.

Learning Objective (Expected Risk Minimization)

True objective:

$$\min_{\beta} \mathbb{E}[\ell(y, x^T\beta)]$$

Since distribution is unknown, we minimize empirical risk.

2. Model Specification

This section defines the mathematical model.

Hypothesis Function

Linear function:

$$z = x^T\beta$$

Probability model:

$$P(y = 1|x) = \sigma(x^T\beta)$$

where sigmoid function is:

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

Parameter Space

$$\beta \in \mathbb{R}^p$$

Structural Assumptions

1. Log-odds are linear:

$$\log \frac{P(y = 1|x)}{P(y = 0|x)} = x^T\beta$$

2. Conditional distribution is Bernoulli.
3. Observations are independent.

3. Loss Function

This section defines how classification error is measured.

Explicit Mathematical Form

Negative log-likelihood (logistic loss):

$$\ell(y, x^T\beta) = \log\left(1 + e^{x^T\beta}\right) - yx^T\beta$$

Alternative form (for $y \in \{-1,1\}$):

$$\ell(y, z) = \log\left(1 + e^{-yz}\right)$$

Why This Loss?

1. Derived from Bernoulli likelihood.

2. Produces probabilistic outputs.
3. Convex and smooth.
4. Penalizes wrong confident predictions heavily.

Convexity Properties

Hessian:

$$\nabla^2 L(\beta) = X^T W X$$

where:

$$W = \mathrm{diag}(\sigma_i(1-\sigma_i))$$

Since $\sigma_i(1-\sigma_i) > 0$,

$$X^T W X \text{ is positive semi-definite}$$

Thus logistic loss is convex.

4. Objective Function

This section defines what is minimized during training.

Empirical Risk

$$R(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(1+e^{x_i^T\beta}\right) - y_i x_i^T\beta\right]$$

Regularized Formulation

L2 regularization:

$$R(\beta) = \frac{1}{n}\Sigma \ell_i + \lambda \parallel \beta \parallel^2$$

L1 regularization:

$$R(\beta) = \frac{1}{n}\sum \ell_i + \lambda \parallel \beta \parallel_1$$

Regularization prevents overfitting and ensures stability.

5. Optimization Method

This section explains parameter estimation.

Closed-form or Iterative?

No closed-form solution exists.

Optimization is done using:

- Gradient Descent
- Newton's Method
- IRLS (Iteratively Reweighted Least Squares)

Gradient Expression

$$\nabla L(\beta) = X^T(\sigma - y)$$

Newton Update

$$\beta^{(t+1)} = \beta^{(t)} - (X^T W X)^{-1} X^T(\sigma - y)$$

Convergence Guarantees

Since objective is convex:

- Any local minimum is global minimum.
- Newton's method converges quadratically near optimum.

Computational Complexity

Per iteration:

- Gradient: $O(np)$
- Newton step: $O(np^2 + p^3)$

6. Statistical Interpretation

MLE Connection

Assume:

$$y|x \sim \text{Bernoulli}(\sigma(x^T\beta))$$

Maximizing likelihood equals minimizing logistic loss.

Thus:

$$\text{Logistic Regression} = \text{Maximum Likelihood Estimator}$$

MAP Interpretation

With Gaussian prior:

$$\beta \sim N(0, \tau^2 I)$$

We obtain L2-regularized logistic regression.

Probabilistic Meaning

Output represents:

$$P(y = 1|x)$$

Provides calibrated probabilities.

7. Regularization & Generalization

This section discusses model capacity.

Bias–Variance Tradeoff

- Without regularization: low bias, potentially high variance.
- With L2: slightly higher bias, lower variance.

Overfitting Behavior

Logistic regression does not overfit easily unless:

- High-dimensional data

- No regularization
- Strong multicollinearity

Capacity Control

Controlled by:

- Number of features
- Regularization parameter λ

8. Theoretical Properties

Convexity / Global Optimality

Objective is convex → unique global optimum.

Consistency

Under regularity conditions:

$$\widehat{\beta} \to \beta$$

as $n \to \infty$.

Asymptotic Normality

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to N(0, I(\beta_0)^{-1})$$

where:

$$I(\beta) = X^T W X$$

Stability

Sensitive to:

- Multicollinearity
- Perfect separation

9. Computational Complexity

Training Time

- Gradient descent: $O(np)$per iteration

- Newton method: $O(np^2 + p^3)$

Inference Time

Prediction per sample:

$$O(p)$$

Memory Complexity

- Data storage: $O(np)$
- Parameter storage: $O(p)$

10. Limitations

This section highlights weaknesses.

When It Fails

1. Perfectly separable data (MLE diverges)
2. Highly nonlinear boundaries
3. Severe multicollinearity
4. Extreme class imbalance

Assumption Violations

- Incorrect linear log-odds assumption
- Non-independent observations

Sensitivity Issues

- Sensitive to feature scaling
- Can be unstable in high dimensions
- Influenced by outliers in feature space