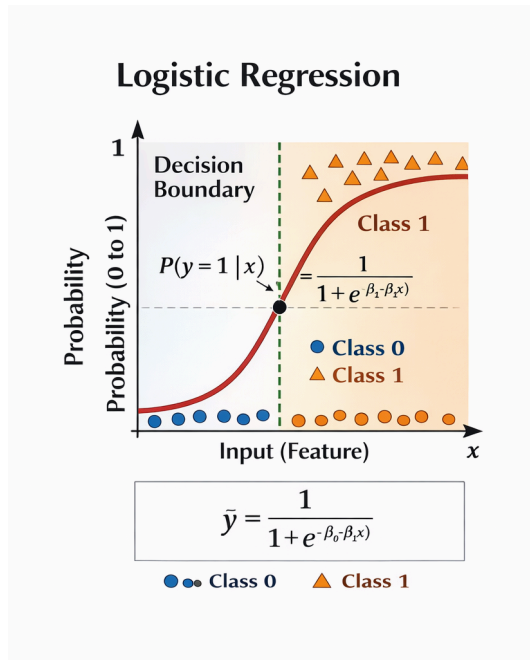


Documentation-Logistic Regression



1 Problem Formulation

1.1 Input Space

$$\mathbf{X} \subseteq \mathbb{R}^d$$

Each sample: $\mathbf{x} \in \mathbb{R}^d$

where each data point contains d numerical features.

1.2 Output Space

Binary classification: $\mathbf{Y} = \{0, 1\}$

1.3 Data Distribution

Assume data is drawn i.i.d. from unknown distribution:

$$(\mathbf{x}, y) \sim P(\mathbf{X}, \mathbf{Y})$$

1.4 Learning Objective (Expected Risk Minimization)

Minimize expected classification risk: $R(\mathbf{w}) = E_{(\mathbf{x}, y) \sim P} [\ell(y, \mathbf{f}_{\mathbf{w}}(\mathbf{x}))]$

Since distribution is unknown, minimize empirical risk.

2 Model Specification

2.1 Hypothesis Function

Logistic model: $P(y = 1 | \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$

Where sigmoid function: $\sigma(z) = 1 / (1 + e^{-z})$

Prediction Rule : $\hat{y} = 1$ if $\sigma(\mathbf{w}^T \mathbf{x}) \geq 0.5$
 $\hat{y} = 0$ otherwise

2.2 Parameter Space

$\mathbf{w} \in \mathbb{R}^d$

2.3 Structural Assumptions

- Log-odds are linear: $\log [P(y = 1 | \mathbf{x}) / P(y = 0 | \mathbf{x})] = \mathbf{w}^T \mathbf{x}$
- Classes approximately linearly separable
- Observations are independent

3 Loss Function

3.1 Explicit Mathematical Form

Binary Cross-Entropy (Log Loss): $\ell(y, \hat{p}) = - [y \log(\hat{p}) + (1 - y) \log(1 - \hat{p})]$

Where: $\hat{p} = \sigma(\mathbf{w}^T \mathbf{x})$

3.2 Why This Loss? (Statistical Reasoning)

Derived from Bernoulli likelihood: $P(y | \mathbf{x}, \mathbf{w}) = \hat{p}^y (1 - \hat{p})^{1-y}$

Maximizing likelihood

\Leftrightarrow minimizing negative log-likelihood

\Leftrightarrow minimizing cross-entropy.

3.3 Convexity Properties

Cross-entropy loss is convex in \mathbf{w} .

Hessian: $\mathbf{H} = \mathbf{X}^T \mathbf{W} \mathbf{X}$

Where \mathbf{W} is diagonal with positive entries.

Therefore:

- Convex objective
- Single global minimum

4 Objective Function

4.1 Empirical Risk Expression

$$\mathbf{J}(\mathbf{w}) = (1/n) \sum_{i=1}^n [-y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) - (1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i))]$$

4.2 Regularized Formulation (if used)

- Ridge (L2) Regularization: $\mathbf{J}(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$
- Lasso (L1) Regularization: $\mathbf{J}(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$

5 Optimization Method

5.1 Closed-form or iterative?

No closed-form solution.

Solved using:

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

5.2 Gradient expression

$$\nabla \mathbf{J}(\mathbf{w}) = (1/n) \mathbf{X}^T (\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y})$$

5.3 Convergence guarantees

Since objective is convex:

- Gradient descent converges to global minimum
- Newton's method converges quadratically near optimum

5.4 Computational complexity

- Each gradient iteration: $O(nd)$
- Newton's method: $O(nd^2 + d^3)$

6 Statistical Interpretation

6.1 MLE / MAP connection

- **MLE Connection:** Logistic regression = Maximum Likelihood Estimation under Bernoulli model.
- **MAP Interpretation** With Gaussian prior: $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I})$

Minimizing: $NLL + \lambda \|\mathbf{w}\|_2^2$

= MAP estimate.

6.2 Noise model assumption

$y \mid \mathbf{x} \sim \text{Bernoulli}(\sigma(\mathbf{w}^T \mathbf{x}))$

6.3 Probabilistic meaning of outputs

Model outputs calibrated probability: $P(y = 1 \mid \mathbf{x})$

Unlike SVM, logistic regression provides probabilistic output.

7 Regularization & Generalization

7.1 Bias–variance tradeoff

- Large $\lambda \rightarrow$ High bias, low variance
- Small $\lambda \rightarrow$ Low bias, high variance

7.2 Overfitting behavior

Occurs when:

- d large relative to n
- Perfect class separation

7.3 Capacity control mechanism

Controlled by:

- Norm constraint on w
- VC dimension $\approx d$

8 Theoretical Properties

8.1 Convexity / Global optimality

Loss is convex \rightarrow global optimum exists.

8.2 Consistency (if applicable)

Under correct model specification:

$$\hat{w} \rightarrow w^* \text{ as } n \rightarrow \infty$$

8.3 Stability considerations

- Stable under L2 regularization
- Sensitive to outliers in feature space

9 Computational Complexity

9.1 Training time complexity

- Gradient Descent: $O(Tnd)$
- Newton: $O(nd^2 + d^3)$

9.2 Inference time complexity

For one sample: $O(d)$

9.3 Memory complexity

$O(nd + d)$

10 Limitations

10.1 When it fails

- Non-linear decision boundaries
- Highly overlapping classes
- Extreme class imbalance

10.2 Assumption violations

- Log-odds not linear
- Non-independent samples

10.3 Sensitivity issues (outliers, scaling, etc.)

- Sensitive to multicollinearity
- Needs feature scaling
- Cannot handle extreme outliers well