The Loss Function in Logistic Regression

Classification Methods

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

Contents of This Video

In this video, we will cover:

- What is a loss function and why we need it
- Log loss (cross-entropy) formula
- Connection to probability and likelihood
- Visual understanding of the loss behavior
- How the model learns by minimizing loss
- Practical implications for student prediction

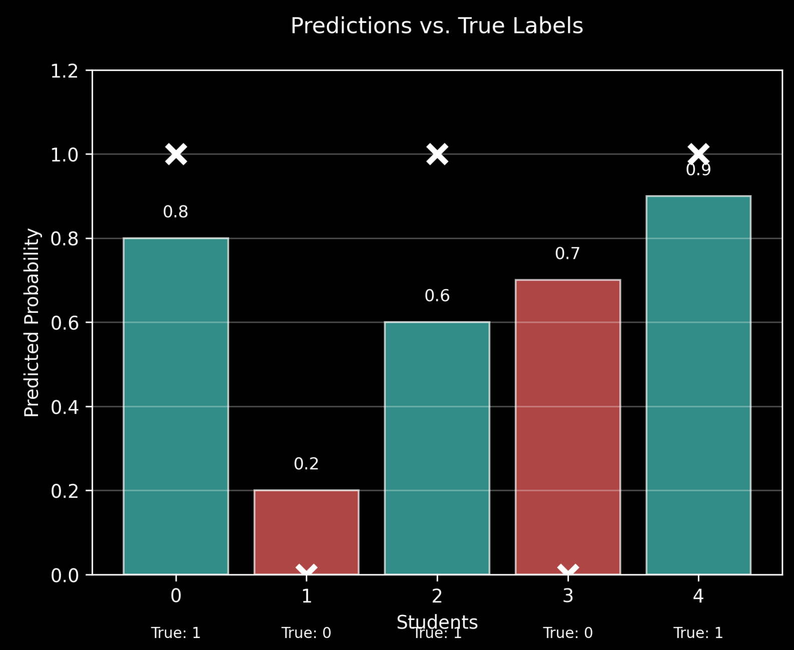
What is a Loss Function?

Definition: A function that measures how wrong model predictions are compared to actual outcomes

For Student Classification:

- True label: Pass (1) or Fail (0)
- Predicted probability: 0.7 (70% chance of passing)
- Loss function tells us how “wrong” this prediction is

Goal: Find model parameters that minimize total loss



The Log Loss Formula

For a single student prediction:

$$\text{Loss} = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})]$$

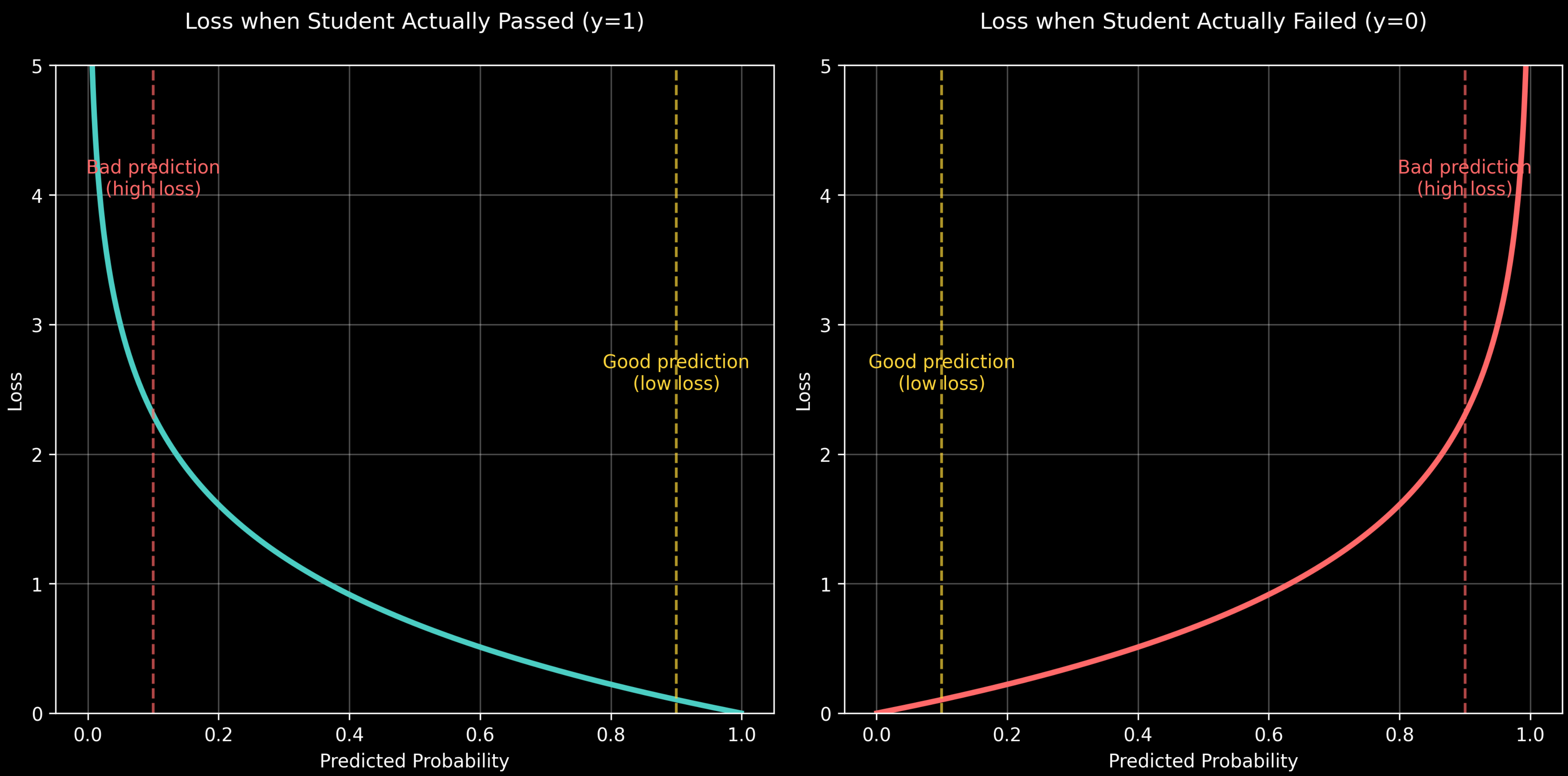
Where:

- y = true label (0 for fail, 1 for pass)
- $\hat{y}$ = predicted probability of passing
- $\log$ = natural logarithm

For the entire dataset:

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

Understanding Log Loss Behavior



Connection to Probability Theory

Why this specific formula?

Bernoulli Distribution: For binary outcomes, the probability of getting label y is:

$$P(y|\mathbf{x}) = \hat{y}^{y}(1 - \hat{y})^{1-y}$$

Likelihood for all students:

$$L(\mathbf{w}, b) = \prod_{i=1}^{m} [\hat{y}^{(i)}]^{y^{(i)}} [1 - \hat{y}^{(i)}]^{1-y^{(i)}}$$

Log-Likelihood (easier to work with):

$$\ell(\mathbf{w}, b) = \sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})\right]$$

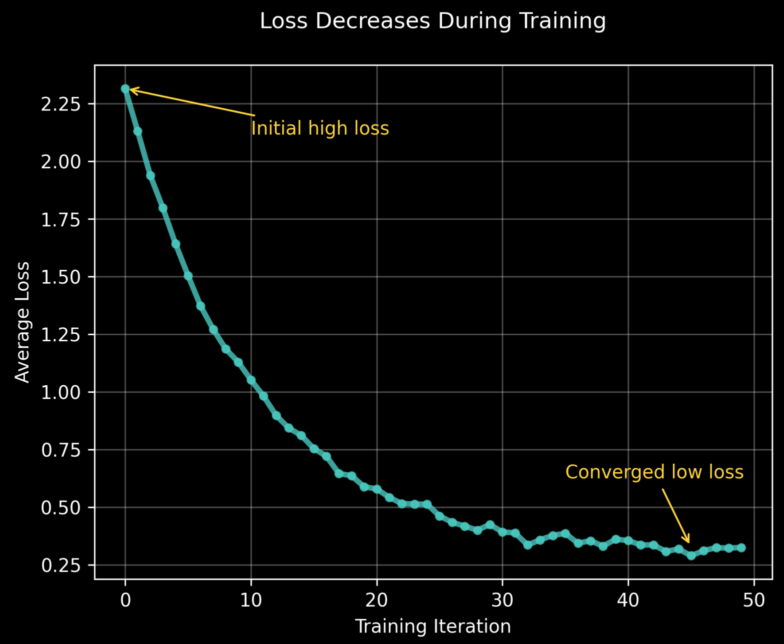
Minimize negative log-likelihood = minimize loss

How the Model Learns

Optimization Process:

1. Start with random coefficients $\mathbf{w}$ and bias b
2. Calculate predictions for all students
3. Compute total loss using log loss formula
4. Adjust coefficients to reduce loss
5. Repeat until loss stops decreasing

Gradient Descent: The most common algorithm used to minimize the loss



Practical Example

Student Success Prediction

Three students with predictions:

Student	True Label	Predicted Prob	Individual Loss
A	Pass (1)	0.9	0.11
B	Fail (0)	0.2	0.22
C	Pass (1)	0.3	1.20

Average Loss: $(0.11 + 0.22 + 1.20)/3 = 0.51$

Student C contributes most to the loss - this prediction needs improvement!

What We've Covered

In this video, we've explored:

- Loss functions measure prediction quality
- Log loss penalizes confident wrong predictions
- Mathematical connection to probability theory
- How gradient descent minimizes loss during training
- Practical examples with student predictions