Logistic Regression

Logistic Regression is a binary classification model used to estimate the probability that an observation belongs to class 1 (or 0).

It models: $P(y = 1|x)$ where $y \in \{0,1\}$.

Despite the name, it is not a regression model for continuous output. It is a classification model that outputs probabilities.

1. Prove Convexity of Logistic Loss

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

Gradient: $\nabla L(\beta) = \sum_{i=1}^{n} x_i \left(\sigma(x_i^T \beta) - y_i\right)$

Hessian: $\nabla^2 L(\beta) = \sum_{i=1}^{n} x_i x_i^T \sigma(x_i^T \beta)(1 - \sigma(x_i^T \beta))$

Define: $W = \text{diag}(\sigma_i(1 - \sigma_i))$

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

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

W is positive semi-definite.

For any vector v:

$$v^T X^T W X v = (Xv)^T W (Xv) \geq 0$$

Therefore:

Logistic loss is convex

2. Why Logistic Regression Does Not Overfit Like High-Degree Polynomial Regression

Key differences:

1. Logistic regression is linear in parameters.
2. Polynomial regression increases model complexity exponentially.
3. Logistic decision boundary is linear unless features are expanded.
4. Logistic loss grows slowly for correctly classified points.
5. Squared loss in high-degree polynomial regression fits noise aggressively.

Thus:

- Logistic regression has controlled model capacity.
- Polynomial regression has high variance when degree is large.

Overfitting depends on feature dimension, not classification vs regression.

3. Derive IRLS from Newton's Method

Newton update: $\beta^{(t+1)} = \beta^{(t)} - [\nabla^2 L(\beta)]^{-1} \nabla L(\beta)$

We already have:

$$\nabla L = X^T(\sigma - y)$$
$$\nabla^2 L = X^T W X$$

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

Rewriting: $\beta^{(t+1)} = (X^T W X)^{-1} X^T W z$

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

This is equivalent to solving a weighted least squares problem.

This algorithm is called IRLS (Iteratively Reweighted Least Squares)

4. Logistic Loss vs Hinge Loss

For binary classification, let: $y \in \{-1, +1\}$

Let prediction score be: $z = x^T\beta$

Define margin: $m = yz$

If:

- $m > 0 \rightarrow$ correctly classified
- $m < 0 \rightarrow$ misclassified

Both losses depend on this margin.

Logistic Loss

Formula $\ell_{\text{logistic}}(m) = \log(1 + e^{-m})$

Key Points

- Smooth and differentiable.
- Always positive.
- Penalizes all points (even correctly classified ones).
- Used in Logistic Regression.
- Gives probability estimates.

Behavior

- If misclassified ($m < 0$) → loss is large.
- If correctly classified with large margin → loss becomes very small.
- Never exactly zero.

Hinge Loss

Formula $\ell_{\text{hinge}}(m) = \max(0, 1 - m)$

Key Points

- Convex but not differentiable at $m = 1$.
- Used in Support Vector Machines (SVM).
- Does NOT give probabilities.

- Enforces a margin.

Behavior

- If $m \geq 1 \rightarrow$ loss = 0.
- If $m < 1 \rightarrow$ penalized linearly.
- Ignores points far from boundary.

Main Differences

Property	**Logistic Loss**	**Hinge Loss**
Formula	log(1 + e^(-m))	max(0, 1 − m)
Smooth?	Yes	No
Gives probabilities?	Yes	No
Zero loss possible?	No	Yes
Used in	Logistic Regression	SVM

5. What Happens When Data is Perfectly Separable?

If there exists βsuch that: $y_i x_i^T \beta > 0 \forall i$

Then likelihood increases as: $\| \beta \| \rightarrow \infty$

Loss approaches zero.

There is no finite maximizer.

6. Why MLE Does Not Exist Under Perfect Separation

Log-likelihood: $\ell(\beta) = \sum y_i x_i^T \beta - \log(1 + e^{x_i^T \beta})$

If separable, we can scale: $\beta' = c\beta$

As $c \rightarrow \infty$: $\ell(\beta') \rightarrow 0$

Likelihood approaches 1.

But no finite maximizer exists.

MLE does not exist because solution diverges

Regularization fixes this.

7. Logistic Regression as Maximum Entropy Classifier

Maximum entropy principle:

Choose distribution that:

1. Satisfies constraints: $E[yx] = \text{observed}$
2. Maximizes entropy: $H(p) = -\sum p \log p$

Using Lagrange multipliers yields:

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

Thus logistic regression is the maximum entropy model under linear feature constraints.

8. Asymptotic Distribution of Estimator

Under regularity conditions: $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, I(\beta_0)^{-1})$

Where Fisher Information: $I(\beta) = X^T W X$

Thus:

$$\hat{\beta} \sim N(\beta_0, (X^T W X)^{-1}) \text{ asymptotically}$$

Estimator is:

- Consistent
- Asymptotically normal
- Efficient

9. Generative vs Discriminative

Generative (Naive Bayes):

Models: $P(x|y), P(y)$

Uses Bayes rule.

Discriminative (Logistic):

Models: $P(y|x)$

Comparison:

Property	**Naive Bayes**	**Logistic**
Assumptions	Strong independence	Fewer assumptions
Small data	Often better	Needs more data
Asymptotic accuracy	Lower	Higher

Naive Bayes converges faster but to worse solution if assumptions wrong.

10. When Does Logistic Regression Fail?

1. Perfect separation (no finite MLE)
2. High multicollinearity
3. p >> n without regularization
4. Nonlinear boundary unless features engineered
5. Severe class imbalance without correction
6. Heavy outliers in feature space

Logistic regression is linear in feature space.
If reality is highly nonlinear and features not transformed, it fails.