Logistic Regression Q&A

1. Prove convexity of logistic loss.

Logistic loss (binary case):

For one sample:

$\ell(w) = \log(1 + e^{-yw^T x})$

where **y** $\in$ **$\{-1, 1\}$**

Let **$z = w^T x$**

First Derivative: $d\ell/dz = -y / (1 + e^{yz})$

Second Derivative: $d^2\ell/dz^2 = e^{yz} / (1 + e^{yz})^2 \geq 0$

Since the second derivative is always non-negative:

✔ Logistic loss is convex
✔ Strictly convex if X has full rank

Thus, optimization has a unique global minimum.

2. Why does logistic regression not overfit as easily as high-degree polynomial regression?

Logistic Regression

- Linear decision boundary
- Limited capacity
- Few parameters
- Implicit smoothness

High-Degree Polynomial Regression

- Very flexible hypothesis space
- Can interpolate noise
- High variance
- Large parameter magnitudes

Core Reason: Overfitting depends on model capacity.

Logistic regression is a low-capacity linear classifier, whereas high-degree polynomial regression greatly increases VC dimension.

3. Derive IRLS from Newton's method.

Log-Likelihood: $\ell(w) = \sum_i [\, y_i \log p_i + (1 - y_i) \log(1 - p_i) \,]$

Where: $p_i = \sigma(w^T x_i)$

Gradient: $\nabla \ell = X^T(y - p)$

Hessian: $H = -X^T W X$

Where: $W = \mathrm{diag}(p_i (1 - p_i))$

Newton Update

$w\_new = w\_old - H^{-1} \nabla \ell$

Substituting:

$w\_new = (X^T W X)^{-1} X^T W z$

Where:

$z = Xw + W^{-1} (y - p)$

This method is called: Iteratively Reweighted Least Squares (IRLS)

Because each iteration solves a weighted least squares problem.

4. Compare logistic loss vs hinge loss mathematically.

Logistic Loss: $\log(1 + e^{-yf}(x))$

- Smooth
- Differentiable
- Probabilistic

Hinge Loss (SVM): $\max(0, 1 - y\, f(x))$

- Not differentiable at margin
- Margin-based

Comparison table

Logistic	Hinge
Smooth	Non-smooth
Probabilistic	Margin-based
Penalizes all points	Penalizes only misclassified/margin points
Used in Logistic Regression	Used in SVM

5. What happens when data is perfectly separable?

If there exists **w** such that:

$\mathbf{y_i\, w^T x_i > 0}$ for all i

Then:

- Log-likelihood increases indefinitely
- $\|w\| \to \infty$
- Decision boundary becomes infinitely steep

6. Why does MLE not exist under perfect separation?

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

If data is separable:

As $\|\mathbf{w}\| \to \infty$,

$\ell(w) \rightarrow 0$

There is no finite maximizer.

Therefore:

❌ MLE does not exist
✔ Regularization is required

7. Derive logistic regression as maximum entropy classifier.

Maximum entropy principle:

Choose the distribution with:

- Maximum entropy
- Subject to feature expectation constraints

Constraint:

$E[yx]$ = empirical mean

Solving the constrained optimization yields:

$P(y \mid x) = 1 / (1 + e^{-w^T x})$

Thus: Logistic regression is the maximum entropy distribution under linear constraints.

8. Analyze asymptotic distribution of estimator.

Under regularity conditions: $\sqrt{n}\,(\hat{w} - w) \rightarrow N(0, I^{-1})$*

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

Thus:

- Estimator is consistent
- Asymptotically normal
- Variance shrinks as $n \rightarrow \infty$

9. Compare generative (Naive Bayes) vs discriminative (logistic).

Generative Model

Models: **$P(x \mid y)\, P(y)$**

Example: Naive Bayes

Pros:

- Works well with small datasets
- Faster training

Discriminative Model

Models: **P(y | x)**

Example: Logistic Regression

Pros:

- Better asymptotic performance
- Fewer assumptions

Key Difference:

Naive Bayes	Logistic Regression
Strong independence assumption	No independence assumption
Estimates joint distribution	Estimates conditional directly
Can be biased	Lower asymptotic error

10. When does logistic regression fail?

1. **Nonlinear decision boundaries**
   Cannot model complex patterns without feature engineering.
2. **Perfect separation**
   MLE does not exist.

3. **Multicollinearity**
   Causes variance inflation.
4. **High dimensional, small sample ($p \gg n$)**
   Unstable without regularization.
5. **Severe class imbalance**
   Biased decision boundary.
6. **Outliers in feature space**
   Affect coefficient estimates.