LOGISTIC REGRESSION

1. Prove That the Log-Likelihood of Logistic Regression Is Concave

"Show that the log-likelihood function of logistic regression is concave in the parameter vector β. Why does this matter practically?"

Answer

The log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} \Big[y_i \ln P_i + (1 - y_i) \ln(1 - P_i) \Big], \qquad P_i = \sigma(x_i^\top \beta)$$

Concavity is shown via the **Hessian**. The gradient is:

$$\nabla \ell(\beta) = \sum_{i=1}^{n} (y_i - P_i)\, x_i = X^\top (y - P)$$

Differentiating again gives the Hessian:

$$H = \nabla^2 \ell(\beta) = -\sum_{i=1}^{n} P_i(1 - P_i)\, x_i x_i^\top = -X^\top W X, \qquad W = \operatorname{diag}\big(P_i(1 - P_i)\big)$$

Since $P_i(1-P_i) \geq 0$ for all i, W is positive semi-definite, and so is X^TWX for any X:

$$z^\top X^\top W X z = \sum_{i=1}^{n} w_i \big(x_i^\top z\big)^2 \;\geq\; 0 \quad \Rightarrow \quad H \preceq 0 \;\; (\ell \text{ concave})$$

Therefore H is negative semi-definite, so ℓ(β) is concave.

Why it matters: concavity guarantees any stationary point found by gradient ascent or Newton–Raphson is a **global** maximum, not a local one — optimization is reliable and solver choice affects only speed, not correctness. This also explains why IRLS (Newton's method here) converges quickly: it exploits the curvature given by W.

2. What Happens Under Perfect Separation, and Why?

"Suppose your training data is perfectly linearly separable by class. What happens to the MLE estimates? Explain why, not just what."

Answer

If a hyperplane perfectly separates the two classes, the MLE **does not exist** — it diverges to infinity rather than converging to a finite value.

Why: the likelihood strictly increases as $\|\beta\| \to \infty$ in the separating direction. If β already classifies every point correctly, scaling it by $c > 1$ pushes every P_i closer to its correct extreme (0 or 1), strictly increasing ℓ(β) with no bound. Since ℓ increases monotonically along that ray with no maximum, iterative solvers either fail to converge or report huge, meaningless coefficients with near-zero standard errors.

This is a direct consequence of Question 1: the Hessian $-X^TWX$ becomes **singular** as $P_i \to 0$ or 1 (since $W \to 0$), so the curvature that would normally bound the maximum disappears.

Practical fix: L2 (Ridge) regularization adds a penalty that is strictly concave and bounded above, guaranteeing a finite MLE even under separation:

$$\ell_{\text{ridge}}(\beta) = \ell(\beta) - \lambda \|\beta\|^2, \qquad \lambda > 0 \;\Rightarrow\; \text{strictly concave, bounded above}$$

This is the theoretical justification for regularizing logistic regression by default in near-separable or high-dimensional settings — not merely an overfitting heuristic.

3. Derive Logistic Regression as a Member of the Exponential Family / GLM Framework

"Show that logistic regression is a special case of the Generalized Linear Model (GLM) framework, and identify its link function and canonical parameter."

Answer

A Bernoulli response Y ~ Bernoulli(P) has probability mass function:

$$f(y;P) = P^y(1-P)^{1-y} = \exp\left\{y\ln\left(\frac{P}{1-P}\right) + \ln(1-P)\right\}$$

This is in **exponential family form**:

$$f(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}, \qquad \theta = \ln\left(\frac{P}{1-P}\right), \quad b(\theta) = \ln(1+e^\theta)$$

with natural (canonical) parameter θ = the logit itself. A GLM links the linear predictor Z = Xβ to E[Y] via a link function g: g(E[Y]) = Z. For logistic regression E[Y] = P, and the **canonical link** sets g(P) equal to the natural parameter:

$$g(P) = \ln\left(\frac{P}{1-P}\right) = Z = X\beta \qquad \text{(canonical logit link)}$$

Because the canonical link is used, several elegant properties follow: the sufficient statistic for β is X^TY, the score equations simplify to $X^T(Y-P) = 0$ (matching Question 1's gradient), and Fisher scoring coincides **exactly** with Newton–Raphson (expected and observed Hessians are equal in the canonical case) — which is why IRLS converges so cleanly for logistic regression specifically, compared to GLMs with non-canonical links.

4. Why Does the Asymptotic Variance of the MLE Depend on P(1−P), and What Does This Imply for Experimental Design?

"Using the Fisher Information, explain why logistic regression estimates are least precise near P = 0.5, and what this implies for data collection strategy."

Answer

The Fisher Information matrix is the same matrix as the negative Hessian in Question 1 (observed and expected information coincide, since this is a canonical-link GLM):

$$I(\beta) = X^\top W X, \qquad W = \mathrm{diag}\big(P_i(1-P_i)\big)$$

The asymptotic covariance of the MLE is:

$$\mathrm{Var}(\hat{\beta}) \;\approx\; I(\beta)^{-1} = (X^\top W X)^{-1}$$

Each observation's contribution is weighted by $P_i(1-P_i)$, which is **maximised at $P_i = 0.5$** and shrinks toward 0 as $P_i \to 0$ or 1. This seems counterintuitive — one might expect points near P = 0.5 (most "uncertain") to be least informative — but it is the opposite: an observation the model already predicts confidently (P near 0 or 1) barely moves the likelihood surface when β changes slightly, so it contributes little curvature/information. An observation near the decision boundary is maximally sensitive to β and sharpens the estimate most.

Implication for experimental design: this is the theoretical basis for **optimal/adaptive experimental design** in binary-response settings (dose–response studies, A/B testing near a threshold) — deliberately sampling covariate values where the current model predicts P ≈ 0.5 yields the most information per observation, shrinking standard errors fastest, rather than uniform or random sampling. It also explains why standard errors blow up under class imbalance or near-separation (Question 2): most points end up with $P_i(1-P_i) \approx 0$, starving the Fisher Information of any signal.

5. Show That Gaussian Naive Bayes and Logistic Regression Are a Generative–Discriminative Pair

"Derive the posterior P(Y=1|X) for a Gaussian Naive Bayes classifier with shared class variance, and show it takes exactly the logistic regression form. What does this reveal about the relationship between generative and discriminative classifiers?"

Answer

Assume a generative model: class prior P(Y=1) = π, and class-conditional densities with a **shared** variance σ²:

$$X \mid Y{=}k \;\sim\; N(\mu_k, \sigma^2), \qquad k = 0, 1 \quad \text{(shared variance } \sigma^2\text{)}$$

By Bayes' theorem, the posterior is:

$$P(Y{=}1 \mid X{=}x) = \frac{\pi\, f_1(x)}{\pi\, f_1(x) + (1-\pi) f_0(x)} = \frac{1}{1 + \dfrac{1-\pi}{\pi} \cdot \dfrac{f_0(x)}{f_1(x)}}$$

Expanding the density ratio $f_0(x)/f_1(x)$ and simplifying the quadratic terms (the x^2 terms cancel exactly because σ^2 is shared across classes):

$$\frac{f_0(x)}{f_1(x)} = \exp\left\{ \frac{(x-\mu_1)^2 - (x-\mu_0)^2}{2\sigma^2} \right\} = \exp\left\{ \frac{x(\mu_0 - \mu_1)}{\sigma^2} + \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} \right\}$$

Substituting back into the posterior gives exactly the logistic sigmoid form:

$$P(Y{=}1 \mid x) = \frac{1}{1 + e^{-(b_0 + b_1 x)}}, \qquad b_1 = \frac{\mu_1 - \mu_0}{\sigma^2}, \quad b_0 = \ln \frac{\pi}{1-\pi} + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}$$

What this reveals: Gaussian Naive Bayes with a shared covariance and logistic regression have **identical functional forms** for P(Y|X) — both are members of the same hypothesis class. They differ only in **how they are fit**: Naive Bayes estimates μ_0, μ_1, σ^2, π by maximising the **joint** likelihood P(X,Y) (generative), while logistic regression estimates b_0, b_1 directly by maximising the **conditional** likelihood P(Y|X) (discriminative). This is the Ng & Jordan (2001) result: generative and discriminative classifiers can share the same asymptotic hypothesis class, but Naive Bayes converges faster with less data (lower variance, higher bias if the Gaussian/independence assumptions are wrong), while logistic regression is asymptotically more accurate as $n \to \infty$ because it makes no distributional assumption on X — only on the form of P(Y|X). A candidate who can produce this derivation on the board demonstrates genuine command of the generative–discriminative distinction, not just familiarity with the logistic regression formula.