CS109 Lecture 26 Worksheet: ANSWER KEY

Beyond Classification — Final-Exam Machine Learning

Summer 2026

Lecture 26 Worksheet: Answer Key

Instructor / TA copy.

The one idea for the day. All three problems are the same recipe: (1) assume a distribution for the data, (2) write the likelihood and take its log, (3) differentiate, (4) climb the gradient. Problem 1 is the recipe for a Bernoulli label, Problem 2 is the recipe for a two-parameter continuous distribution, and Problem 3 is the recipe applied to a *learned parameter of a fix-up model*. Emphasize at the board that the derivative $\frac{d}{dz}\sigma(z) = \sigma(z)(1-\sigma(z))$ and the "error times input" shape of the gradient recur in both logistic problems.

Problem 1: Logistic Regression *(Spring 2026 Final)*

(a) The linear term is

$$z = \theta_0 + \theta_1 x_1 x_2 = -1 + 2(1)(1) = 1.$$

Therefore

$$P(Y = 1 \mid \mathbf{x}) = \sigma(1) = \frac{1}{1+e^{-1}} = \frac{e}{e+1} \approx \boxed{0.731}.$$

Teaching note. The feature that multiplies θ_1 is the **product** $x_1 x_2$, not x_1 or x_2 separately — this is an interaction feature. Students should treat $s = x_1 x_2$ as a single engineered feature and the model is then ordinary logistic regression in s.

(b) Let $s^{(i)} = x_1^{(i)} x_2^{(i)}$ and $\hat{p}^{(i)} = \sigma\big(\theta_0 + \theta_1 s^{(i)}\big)$. Each label is Bernoulli($\hat{p}^{(i)}$), so the log-likelihood is

$$LL(\theta_0, \theta_1) = \sum_{i=1}^{n} \Big[y^{(i)} \log \hat{p}^{(i)} + \big(1 - y^{(i)}\big) \log\big(1 - \hat{p}^{(i)}\big) \Big].$$

Differentiate one term of LL with respect to θ_1 using the chain rule, going through the probability $\hat{p}^{(i)}$ and then through the argument $z^{(i)} = \theta_0 + \theta_1 s^{(i)}$:

$$\frac{\partial LL}{\partial \theta_1} = \sum_{i=1}^{n} \underbrace{\left[\frac{y^{(i)}}{\hat{p}^{(i)}} - \frac{1 - y^{(i)}}{1 - \hat{p}^{(i)}} \right]}_{\partial LL / \partial \hat{p}^{(i)}} \cdot \underbrace{\hat{p}^{(i)}\big(1 - \hat{p}^{(i)}\big)}_{\partial \hat{p}^{(i)} / \partial z^{(i)}} \cdot \underbrace{s^{(i)}}_{\partial z^{(i)} / \partial \theta_1},$$

where the middle factor uses $\dfrac{\partial}{\partial z}\sigma(z) = \sigma(z)\big(1 - \sigma(z)\big)$. Now combine the bracket over a common denominator and watch the sigmoid-derivative factor cancel it:

$$\left[\frac{y^{(i)}}{\hat{p}^{(i)}} - \frac{1 - y^{(i)}}{1 - \hat{p}^{(i)}} \right] = \frac{y^{(i)}\big(1 - \hat{p}^{(i)}\big) - \big(1 - y^{(i)}\big)\hat{p}^{(i)}}{\hat{p}^{(i)}\big(1 - \hat{p}^{(i)}\big)} = \frac{y^{(i)} - \hat{p}^{(i)}}{\hat{p}^{(i)}\big(1 - \hat{p}^{(i)}\big)},$$

where the numerator simplifies because $y^{(i)}(1-\hat{p}^{(i)})-(1-y^{(i)})\hat{p}^{(i)} = y^{(i)}-y^{(i)}\hat{p}^{(i)}-\hat{p}^{(i)}+y^{(i)}\hat{p}^{(i)} = y^{(i)}-\hat{p}^{(i)}$. Multiplying by the middle factor $\hat{p}^{(i)}(1-\hat{p}^{(i)})$ **cancels the denominator exactly**:

$$\frac{y^{(i)}-\hat{p}^{(i)}}{\hat{p}^{(i)}\big(1-\hat{p}^{(i)}\big)} \cdot \hat{p}^{(i)}\big(1-\hat{p}^{(i)}\big) \cdot s^{(i)} = \big(y^{(i)}-\hat{p}^{(i)}\big)\, s^{(i)}.$$

So the partial derivatives collapse to the clean **(error)** $\times$ **(input)** form (θ_0 is identical but with input 1, since $\partial z^{(i)}/\partial\theta_0 = 1$):

$$\frac{\partial LL}{\partial \theta_0} = \sum_{i=1}^{n}\Big(y^{(i)}-\hat{p}^{(i)}\Big), \qquad \frac{\partial LL}{\partial \theta_1} = \sum_{i=1}^{n}\Big(y^{(i)}-\hat{p}^{(i)}\Big)\, s^{(i)}.$$

Setting these to zero gives the MLE conditions, but they have no closed-form solution, so we find the optimal $\hat{\theta}_0, \hat{\theta}_1$ by gradient ascent:

$$\theta_0 \;\leftarrow\; \theta_0 + \eta \sum_{i=1}^{n}\Big(y^{(i)} - \sigma\big(\theta_0 + \theta_1 s^{(i)}\big)\Big),$$

$$\theta_1 \;\leftarrow\; \theta_1 + \eta \sum_{i=1}^{n}\Big(y^{(i)} - \sigma\big(\theta_0 + \theta_1 s^{(i)}\big)\Big)\, s^{(i)}.$$

Teaching note. The gradient is always **(error)** $\times$ **(input)**: the bias θ_0 has "input" 1, and θ_1 has input $s^{(i)}$. This is the exact same shape they will see again in Problem 3.

Problem 2: It's Gamma Time *(Winter 2026 Final)*

MLE procedure: Likelihood → Log-Likelihood → Gradient → Optimization. Because $\psi(\alpha)$ (the digamma function) has no simple inverse, we *cannot* set the derivative to zero and solve in closed form; we must use gradient-based optimization (gradient ascent).

Likelihood and log-likelihood.

$$L(\alpha,\beta) = \prod_{i=1}^{n} f(x_i \mid \alpha,\beta), \qquad \ell(\alpha,\beta) = \sum_{i=1}^{n} \log f(x_i \mid \alpha,\beta).$$

Substituting the Gamma PDF $f(x_i \mid \alpha,\beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}$ and expanding the log,

$$\ell(\alpha,\beta) = \sum_{i=1}^{n}\Big[\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log x_i - \beta x_i\Big].$$

Derivative with respect to α (this is where the digamma function appears, since $\frac{d}{d\alpha}\log\Gamma(\alpha) = \psi(\alpha)$):

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n}\Big[\log\beta - \psi(\alpha) + \log x_i\Big] = n\log\beta - n\,\psi(\alpha) + \sum_{i=1}^{n}\log x_i.$$

Derivative with respect to β:

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n}\left(\frac{\alpha}{\beta} - x_i\right) = \frac{\alpha n}{\beta} - \sum_{i=1}^{n} x_i.$$

Optimization. Since $\frac{\partial \ell}{\partial \alpha} = 0$ cannot be solved for α in closed form (no inverse digamma), run gradient ascent on both parameters simultaneously, calling `digamma(a)` to evaluate $\psi(\alpha)$ at each step:

$$\alpha \leftarrow \alpha + \eta\,\frac{\partial \ell}{\partial \alpha}, \qquad \beta \leftarrow \beta + \eta\,\frac{\partial \ell}{\partial \beta},$$

until the gradient is (approximately) zero.

Teaching note. Full credit requires: the log-likelihood expansion, **both** partial derivatives, the observation that $\frac{d}{d\alpha}\log\Gamma(\alpha) = \psi(\alpha)$, and the statement that we finish with gradient ascent because there is no closed form. (Aside: $\frac{\partial \ell}{\partial \beta} = 0$ alone *does* give $\hat{\beta} = \alpha n / \sum_i x_i$, so a common partial-credit path is solving β in terms of α and gradient-ascending only on α.)

Problem 3: Recalibrating an Uncalibrated Model *(Fall 2024 Final)*

(a)

$$P(Y = 1 \mid \hat{p} = 0.9,\ a = 2) = \sigma(2 \cdot 0.9 - 0.5) = \sigma(1.3) = \frac{1}{1 + e^{-1.3}} \approx \boxed{0.786}.$$

(b) This is the logistic-regression recipe with a single parameter a. We are still doing binary classification, so each $y^{(i)}$ is Bernoulli with success probability $P(Y = 1 \mid \hat{p}^{(i)}) = \sigma\big(a\hat{p}^{(i)} - 0.5\big)$. Write the likelihood and log-likelihood:

$$L(a) = \prod_{i=1}^{n} P\big(Y = 1 \mid \hat{p}^{(i)}\big)^{\,y^{(i)}} \Big(1 - P\big(Y = 1 \mid \hat{p}^{(i)}\big)\Big)^{\,1-y^{(i)}},$$

$$LL(a) = \sum_{i=1}^{n} \Big[y^{(i)} \log P\big(Y = 1 \mid \hat{p}^{(i)}\big) + \big(1 - y^{(i)}\big) \log\big(1 - P(Y = 1 \mid \hat{p}^{(i)})\big)\Big].$$

Differentiate with the chain rule, splitting into the log-likelihood's dependence on the probability times the probability's dependence on a:

$$\frac{\partial LL}{\partial a} = \sum_{i=1}^{n} \underbrace{\left[\frac{y^{(i)}}{P^{(i)}} - \frac{1 - y^{(i)}}{1 - P^{(i)}}\right]}_{\partial LL / \partial P^{(i)}} \cdot \underbrace{P^{(i)}\big(1 - P^{(i)}\big)\,\hat{p}^{(i)}}_{\partial P^{(i)} / \partial a},$$

where $P^{(i)} = \sigma\big(a\hat{p}^{(i)} - 0.5\big)$ and the second factor is exactly the derivative given in the problem. The $P^{(i)}(1 - P^{(i)})$ cancels the denominators, collapsing this to the familiar **error** $\times$ **input** form:

$$\frac{\partial LL}{\partial a} = \sum_{i=1}^{n} \Big(y^{(i)} - P^{(i)}\Big)\,\hat{p}^{(i)} = \sum_{i=1}^{n} \Big(y^{(i)} - \sigma\big(a\hat{p}^{(i)} - 0.5\big)\Big)\hat{p}^{(i)}.$$

Then estimate a by gradient ascent: $a \leftarrow a + \eta\,\dfrac{\partial LL}{\partial a}$.

Teaching note. This is the payoff of the day: Platt recalibration is *just logistic regression with one feature* ($\hat{p}$) and one weight (a). The gradient has the identical (error) $\times$ (input) shape as Problem 1(b). Point out that the "-0.5" is a fixed offset, not a learned intercept, so there is only one parameter to fit. The cancellation of $P^{(i)}(1 - P^{(i)})$ is the same sigmoid-derivative cancellation seen in Problem 1 — worth showing explicitly so students trust it.