

CS109 Lecture 20 Worksheet: Logistic Regression

Sigmoids, the Logistic Regression Assumption, Log-Likelihood, and Gradient Ascent

Summer 2026

Lecture 20 Worksheet: Logistic Regression

Topics: the sigmoid function, classification as $P(Y = 1 \mid X = x)$, the logistic regression assumption, the log-likelihood of labelled data, the gradient $\frac{\partial LL}{\partial \theta_j} = \sum_i x_j^{(i)} [y^{(i)} - \sigma(\theta^T x^{(i)})]$, and gradient ascent.

How to use this sheet. Work in small groups. A TA and the instructor will circulate to help. We will go over selected solutions on the board. You do not need to finish every part; aim for understanding over speed. Keep fractions exact where you can, and carry sigmoid values to 4 decimal places.

Useful facts. $\sigma(z) = \frac{1}{1 + e^{-z}}$, $\sigma'(z) = \sigma(z)[1 - \sigma(z)]$, $\sigma(0) = 0.5$, $\sigma(0.5) = 0.6225$, $\sigma(1) = 0.7311$, $\sigma(2) = 0.8808$.

Problem 1 (Review of Lecture 19): MLE Warm-Up

You collect the number of pages served per second by a web server over 5 seconds and model the counts as i.i.d. $X_i \sim \text{Poi}(\lambda)$. You observe

$$[3, 1, 4, 2, 5].$$

- (a) Write the log-likelihood $LL(\lambda)$, dropping terms that do not depend on λ .
- (b) Differentiate, set to 0, and find $\hat{\lambda}$.
- (c) In one sentence: why did dropping the $\ln(x_i!)$ terms not change the answer?

Problem 2: Sigmoids and a Single Prediction

A logistic regression model for “will this student like this movie?” uses two features: x_1 = number of sci-fi movies they rated highly, x_2 = number of romcoms they rated highly. As always we prepend $x_0 = 1$ so that θ_0 acts as an intercept. The trained parameters are

$$\theta = [\theta_0, \theta_1, \theta_2] = [-1, 2, 0.5].$$

- (a) For a student with $(x_1, x_2) = (1, 2)$, compute $z = \theta^T x$ and then $\hat{y} = P(Y = 1 \mid X = x) = \sigma(z)$.
- (b) If we classify by thresholding at 0.5, what class do we predict? Show that thresholding $\sigma(z)$ at 0.5 is exactly the same as thresholding z at 0.
- (c) Find *all* (x_1, x_2) for which the model is exactly undecided. Describe this set geometrically. What is it called?

- (d) The model has a third feature, $x_3 =$ number of documentaries rated highly, with $\theta_3 = 0$. What does a weight of exactly 0 say about that feature? What would a large *negative* weight say?

Problem 3: One Step of Gradient Ascent, by Hand

You are training a logistic regression with two binary features on this tiny dataset (again, $x_0 = 1$ is prepended):

i	x_0	x_1	x_2	y
1	1	1	0	1
2	1	0	1	0

Initialize $\theta = [0, 0, 0]$ and use step size $\eta = 1$.

- (a) Compute $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$ under the initial parameters. What is the log-likelihood $LL(\theta)$ of the data at initialization?
- (b) Recall the gradient for one training example is

$$\frac{\partial LL}{\partial \theta_j} = x_j [y - \sigma(\theta^T x)]$$

and that gradients from separate examples *add*. Compute the full gradient vector $\left[\frac{\partial LL}{\partial \theta_0}, \frac{\partial LL}{\partial \theta_1}, \frac{\partial LL}{\partial \theta_2} \right]$ over both datapoints.

- (c) Perform the update $\theta_j \leftarrow \theta_j + \eta \cdot \frac{\partial LL}{\partial \theta_j}$ and report the new θ .
- (d) Recompute $\hat{y}^{(1)}$, $\hat{y}^{(2)}$, and $LL(\theta)$. Did the log-likelihood go up? By how much?
- (e) θ_0 did not move. Explain why, using the structure of this particular dataset.

Problem 4: Where the Gradient Comes From

This problem walks through the derivation from lecture. Assume one training example (x, y) with $y \in \{0, 1\}$, and let $\hat{y} = \sigma(\theta^T x)$.

- (a) The logistic regression assumption says $Y \mid X = x \sim \text{Bern}(\hat{y})$. Write the PMF of Y in the single differentiable expression $\hat{y}^y (1 - \hat{y})^{1-y}$, then take the log to show

$$LL(\theta) = y \ln \hat{y} + (1 - y) \ln(1 - \hat{y}).$$

- (b) Differentiate that expression with respect to $\hat{y}$ and simplify to a single fraction. (Hint: your answer should be $\frac{y - \hat{y}}{\hat{y}(1 - \hat{y})}$.)
- (c) Show that $\frac{\partial \hat{y}}{\partial \theta_j} = \hat{y}(1 - \hat{y}) x_j$. You will need $\sigma'(z) = \sigma(z)[1 - \sigma(z)]$ and the chain rule.
- (d) Multiply (b) and (c) via the chain rule. Something beautiful cancels. What is $\frac{\partial LL}{\partial \theta_j}$?
- (e) Interpret the result in one sentence: what is $y - \hat{y}$, and why does multiplying it by x_j make sense as an update direction?

Problem 5: Logistic Regression Code (pset7: logistic_regression)

Implement Logistic Regression for binary input/output data using the gradient ascent algorithm described in class. Train on `simple-train.csv` with learning rate $\eta = 0.0001$ and 1,000 training steps, and test on `simple-test.csv`; you should be able to achieve 100% classification accuracy on the testing data. In this dataset there are two input features and the output class is determined entirely by the first feature (i.e. $y = x_1$).

Report the value of the weight associated with x_1 to 5 decimal places.

On the worksheet today, don't write the code — instead, sketch the pseudocode: what are the three nested loops, what gets reset at the top of each outer iteration, and where exactly does the parameter update happen relative to the loop over training examples?

Problem 6: Conceptual Rapid Fire

Answer each in one or two sentences.

- (a) In training a logistic regression model, why do we optimize the *log*-likelihood rather than the likelihood directly? (*problem from the 2025 Winter Final*)
- (b) Why is logistic regression a more principled model than linear regression for binary classification? (*problem from the 2025 Winter Final*)
- (c) If Y and a particular feature X_i are independent, what would you expect for the learned coefficient θ_i , and why? (*problem from the 2025 Winter Final*)
- (d) A student decides to add a *second* intercept to their logistic regression. To do so they add an extra feature whose value is 0 for every datapoint. How will this impact training? (*problem from the 2022 Fall Final*)
- (e) You want to model an interaction between two features, so your model becomes $P(Y = y \mid X = x) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 \cdot x_1 x_2)$. How would you change your logistic regression *code* to train this model? (*problem from the 2022 Fall Final*)

Challenge Problem (optional): Logistic Regression With a Prior

(*problem from the 2023 Winter Final*)

We would like to use MAP instead of MLE, because we don't have much training data. Use a simplified logistic regression with a single parameter θ and a single feature x (both numbers, not vectors, and no bias term):

$$P(Y = 1 \mid X = x) = \sigma(\theta \cdot x).$$

Our prior belief is $\theta \sim N(\mu = 0, \sigma = 1)$. Training data is $[(x_1, y_1), \dots, (x_n, y_n)]$.

- (a) Based **only** on the prior, how many times larger is the probability density that $\theta = 0$ compared to the density that $\theta = 2$?
- (b) The MAP objective is $\theta_{\text{MAP}} = \arg \max_{\theta} f(\theta \mid y_1, \dots, y_n)$. Derive, step by step, that this equals

$$\arg \max_{\theta} \left[\log f(\theta) + \sum_{i=1}^n \log f(y_i \mid \theta) \right].$$

- (c) Write an expression for the gradient $\frac{\partial}{\partial \theta}$ of that objective, so that we could run gradient ascent. In one sentence, what does the extra term coming from the prior *do* to the parameter during training?