Logistic regression: log-odds, decision boundaries and thresholds

Ten problems on what a logistic regression model says, rather than how it is trained: probabilities from a logit, weights read as odds ratios, the decision boundary as a line with its normal and a point's distance from it, a threshold other than one half, the scale of the weights, separable data where the weights grow without bound and a ridge penalty that stops them, the linear boundary of two Gaussian classes with unequal priors, XOR and the feature that fixes it, a circular boundary from squared features, and the boundaries between three softmax classes, with worked solutions and the mistakes of multiplying a probability by an odds ratio, leaving out the norm in a distance, putting the threshold on the logit, trusting a small gradient on separable data and placing the boundary halfway between the classes when the priors differ.

BEFORE YOU START

Logistic regression turns a linear score into a probability. The score $z = w^\top x + b$ is the log-odds of the positive class, the sigmoid turns it into a probability, and the model predicts the positive class where the probability is high enough. The set of points where it is exactly on the fence is the decision boundary, and because z is linear in x that boundary is a line, a plane or a hyperplane, with w as its normal. The regression-gradients page trains the model; these ten problems read it. They compute probabilities and odds ratios, draw the boundary and measure distances to it, move it with a threshold, separate the role of the direction of w from its size, follow the weights on separable data, derive the boundary from two Gaussian classes, and then bend the boundary with new features and split the plane among three classes. The five mistakes at the end are the ones that give a believable number: a probability multiplied by an odds ratio, a distance without the norm of w, a threshold applied to the logit, a small gradient taken as convergence, and a boundary placed halfway between the classes when one class is rarer.

- **The model.** For features $x \in \mathbb{R}^d$, weights $w \in \mathbb{R}^d$ and bias b, the logit is $z = w^\top x + b$ and the model's probability of class 1 is $p = \sigma(z)$, with $\sigma(t) = 1/(1 + e^{-t})$. Labels are $y \in \{0, 1\}$.
- **Log-odds.** The odds of class 1 are $p/(1-p)$, and $\ln \dfrac{p}{1-p} = z$: the logit is the log-odds. $\ln$ is the natural log.
- **Decision rule.** With threshold τ in $(0, 1)$, predict 1 when $p \geq \tau$. Since σ is increasing, this is $z \geq \ln \dfrac{\tau}{1-\tau}$; for $\tau = \frac{1}{2}$ it is $z \geq 0$.
- **Decision boundary.** The set $\{x : w^\top x + b = c\}$ for the cutoff c on the logit is a hyperplane with normal vector w, pointing toward the side where class 1 is predicted. The signed distance from a point x to $\{w^\top x + b = 0\}$ is $(w^\top x + b)/\|w\|$.

- **Loss.** The training loss is the mean binary cross-entropy $L = -\frac{1}{N}\sum_n \left[y_n \ln p_n + (1 - y_n)\ln(1 - p_n)\right]$, as on the regression-gradients page. For one example, $-\ln \sigma(z) = \ln(1 + e^{-z})$ and $-\ln(1 - \sigma(z)) = \ln(1 + e^{z})$.
- **Softmax.** With k classes, each class has its own w_i and b_i, the logits are $z_i = w_i^\top x + b_i$, and the probabilities are $\text{softmax}(z)_i = e^{z_i} / \sum_j e^{z_j}$.

Builds on: Regression gradients: linear, logistic and softmax

Problems

1. A model has $w = (2, -1)$ and $b = -1$. Find the logit, the probability of class 1 and the prediction at $\tau = \frac{1}{2}$ for $x = (1, 0.5)$, $x = (0, 1)$ and $x = (0.5, 0)$. difficulty ·

2. For the model of Problem 1, show that the logit is the log-odds, and say what happens to the odds when x_1 increases by 1 and when x_2 increases by 1. Check with the odds at $(1, 0.5)$ and $(2, 0.5)$. difficulty ·

3. For the model of Problem 1, write the decision boundary at $\tau = \frac{1}{2}$ as a line in the (x_1, x_2) plane, find where it crosses the axes, say which side is class 1, and find the signed distances of $(1, 0.5)$ and $(0, 1)$ from it. difficulty ··

4. False positives are expensive, so the model of Problem 1 is used with threshold $\tau = 0.8$. Find the new boundary, how far it moves and in which direction, and the new prediction for $(1, 0.5)$. difficulty ··

5. Multiply w and b of Problem 1 by $c > 0$. Show that the decision boundary does not change, and find p at $(1, 0.5)$ for $c = 3$ and $c = 10$. Over what distance across the boundary does p go from $\sigma(-1)$ to $\sigma(1)$? difficulty ··

6. Train a one-feature model without bias on two points: $x = -1$ with $y = 0$ and $x = 1$ with $y = 1$. Show that the mean loss is $L(w) = \ln(1 + e^{-w})$, that it has no minimizer, and what gradient descent does. Then add the penalty $\frac{\lambda}{2}w^2$ and find the minimizer for $\lambda = 0.1$. difficulty ···

7. Class 0 has features $x \sim \text{Normal}(\mu_0, \sigma^2)$ and class 1 has $x \sim \text{Normal}$ difficulty ···
(μ_1, σ^2), with prior probabilities π_0 and π_1. Show that $\ln \dfrac{P(y=1 \mid x)}{P(y=0 \mid x)}$ is linear in x, and find w and b. Find the boundary for $\mu_0 = 0$, $\mu_1 = 2$, $\sigma = 1$ with equal priors, and with $\pi_1 = 0.2$.

8. The four points $(0,0)$ and $(1,1)$ of class 0 and $(0,1)$ and $(1,0)$ of class difficulty ··
1 are the XOR pattern. Show that no line separates them. Then add the feature $x_3 = x_1 x_2$ and check that $z = x_1 + x_2 - 2x_3 - 0.5$ separates them.

9. A model uses the features x_1^2 and x_2^2, with logit $z = x_1^2 + x_2^2 - 4$. difficulty ··
Describe the decision boundary at $\tau = \frac{1}{2}$ and the class-1 region, and find p at $(0,0)$, $(1,1)$ and $(2,2)$.

10. A three-class softmax model has $w_1 = (1,0)$, $w_2 = (0,1)$, $w_3 = (0,0)$ difficulty ···
and all biases 0. Find the boundary between each pair of classes, describe the region of each class, and find the probabilities at $x = (2,1)$.

Answers

1. $(1, 0.5)$: $z = 0.5$, $p \approx 0.6225$, class 1; $(0, 1)$: $z = -2$, $p \approx 0.1192$, class 0; $(0.5, 0)$: $z = 0$, $p = 0.5$, on the boundary
2. The logit is the log-odds; a unit step in x_1 multiplies the odds by $e^2 \approx 7.39$, and a unit step in x_2 by $e^{-1} \approx 0.37$
3. Boundary $x_2 = 2x_1 - 1$, through $(0.5, 0)$ and $(0, -1)$, with class 1 on the side $w = (2, -1)$ points to; signed distances 0.2236 and -0.894
4. Boundary $2x_1 - x_2 - 1 = \ln 4$, moved 0.620 along w; $(1, 0.5)$ is now predicted class 0
5. The boundary is the same for every $c > 0$; $p \approx 0.8176$ for $c = 3$ and 0.9933 for $c = 10$; p goes from 0.269 to 0.731 over a band of width $2/(c\|w\|)$, which is $2/\sqrt{5} \approx 0.894$ for $c = 1$
6. $L(w) = \ln(1 + e^{-w})$ has no minimizer, and gradient descent sends w to infinity; with $\lambda = 0.1$ the minimizer is $w \approx 1.634$
7. $P(y = 1 \mid x) = \sigma(wx + b)$ with $w = (\mu_1 - \mu_0)/\sigma^2$ and $b = (\mu_0^2 - \mu_1^2)/(2\sigma^2) + \ln(\pi_1/\pi_0)$; the boundary is $x = 1$ for equal priors and $x = 1 + \frac{\ln 4}{2} \approx 1.693$ for $\pi_1 = 0.2$
8. No linear boundary in (x_1, x_2) separates XOR; with the product feature, $z = x_1 + x_2 - 2x_1x_2 - 0.5$ does
9. The boundary is the circle $x_1^2 + x_2^2 = 4$, with class 1 outside it; $p \approx 0.018$, 0.119 and 0.982
10. Boundaries $x_1 = x_2$, $x_1 = 0$ and $x_2 = 0$, meeting at the origin; class 3 owns the negative quadrant and classes 1 and 2 split the rest along $x_1 = x_2$; at $(2, 1)$ the probabilities are $(0.665, 0.245, 0.090)$

Worked solutions

Problem 1

> A model has $w = (2, -1)$ and $b = -1$. Find the logit, the probability of class 1 and the prediction at $\tau = \frac{1}{2}$ for $x = (1, 0.5)$, $x = (0, 1)$ and $x = (0.5, 0)$.

$x = (1, 0.5)$: $z = 2 - 0.5 - 1 = 0.5$, $p = \sigma(0.5) \approx 0.6225$	$w^\top x = 2 \cdot 1 + (-1) \cdot 0.5$.
$x = (0, 1)$: $z = 0 - 1 - 1 = -2$, $p = \sigma(-2) \approx 0.1192$	A negative logit gives $p < \frac{1}{2}$.
$x = (0.5, 0)$: $z = 1 - 0 - 1 = 0$, $p = \frac{1}{2}$	A zero logit is even odds.
$(1, 0.5)$: $z = 0.5$, $p \approx 0.6225$, **class** 1; $(0, 1)$: $z = -2$, $p \approx 0.1192$, **class** 0; $(0.5, 0)$: $z = 0$, $p = 0.5$, **on the boundary**	The sign of the logit decides the class at $\tau = \frac{1}{2}$; its size says how sure the model is. With the rule $p \geq \tau$, a point on the boundary goes to class 1.

Problem 2

> For the model of Problem 1, show that the logit is the log-odds, and say what happens to the odds when x_1 increases by 1 and when x_2 increases by 1. Check with the odds at $(1, 0.5)$ and $(2, 0.5)$.

$1 - \sigma(z) = \dfrac{e^{-z}}{1 + e^{-z}}$, so $\dfrac{\sigma(z)}{1 - \sigma(z)} = e^{z}$	Divide the two fractions; the denominators cancel.
$\ln \dfrac{p}{1 - p} = z$	Take logs.
Increasing x_1 by 1 adds $w_1 = 2$ to z, so it multiplies the odds by $e^{2} \approx 7.389$	Adding to a log multiplies the quantity.

Increasing x_2 by 1 multiplies the odds by $e^{-1} \approx 0.368$	$w_2 = -1$.
Odds at $(1, 0.5)$: $e^{0.5} \approx 1.649$; at $(2, 0.5)$: $e^{2.5} \approx 12.18$	Their ratio is e^2.
The logit is the log-odds; a unit step in x_1 multiplies the odds by $e^2 \approx 7.39$, and a unit step in x_2 by $e^{-1} \approx 0.37$	e^{w_j} is the odds ratio of feature j, the same at every starting point. The probability itself changes by different amounts at different points: from 0.6225 to $\sigma(2.5) \approx 0.9241$ here (Mistake 1).

Problem 3

> For the model of Problem 1, write the decision boundary at $\tau = \frac{1}{2}$ as a line in the (x_1, x_2) plane, find where it crosses the axes, say which side is class 1, and find the signed distances of $(1, 0.5)$ and $(0, 1)$ from it.

$2x_1 - x_2 - 1 = 0$, that is $x_2 = 2x_1 - 1$	The points where $z = 0$.
It crosses the x_1 axis at $(0.5, 0)$ and the x_2 axis at $(0, -1)$	Set $x_2 = 0$, then $x_1 = 0$.
The normal $w = (2, -1)$ points to the side with $z > 0$, below and to the right of the line	Moving along w raises z by $\|w\|^2 = 5$ per unit of w.
$\|w\| = \sqrt{4 + 1} = \sqrt{5}$	The length of the normal.
$(1, 0.5)$: $\dfrac{0.5}{\sqrt{5}} \approx 0.2236$; $(0, 1)$: $\dfrac{-2}{\sqrt{5}} \approx -0.894$	The logit divided by $\|w\|$.

Boundary $x_2 = 2x_1 - 1$, through $(0.5, 0)$ and $(0, -1)$, with class 1 on the side $w = (2, -1)$ points to; signed distances 0.2236 and -0.894	The logit is $\|w\|$ times the signed distance, so a point's logit grows linearly as it moves away from the boundary. The point $(0, 1)$ has the larger logit in size, and it is also the farther of the two from the line.

Problem 4

> False positives are expensive, so the model of Problem 1 is used with threshold $\tau = 0.8$. Find the new boundary, how far it moves and in which direction, and the new prediction for $(1, 0.5)$.

$p \geq 0.8$ exactly when $z \geq \ln \dfrac{0.8}{0.2} = \ln 4 \approx 1.386$	σ is increasing, so invert it at 0.8.
New boundary: $2x_1 - x_2 - 1 = \ln 4$	A line parallel to the old one, with the same normal.
It is $\dfrac{\ln 4}{\|w\|} = \dfrac{1.386}{\sqrt{5}} \approx 0.620$ from the old one	Two parallel lines $w^\top x + b = 0$ and $= c$ are $\|c\|/\|w\|$ apart.
It moves along w, into the old class-1 side	The class-1 region shrinks.
$(1, 0.5)$ has $z = 0.5 < 1.386$	Problem 1.
Boundary $2x_1 - x_2 - 1 = \ln 4$, moved 0.620 along w; $(1, 0.5)$ is now predicted class 0	A threshold never tilts a logistic boundary, it only slides it along w. Raising τ trades recall for precision, and sweeping it from 0 to 1 traces the model's ROC curve.

Problem 5

Multiply w and b of Problem 1 by $c > 0$. Show that the decision boundary does not change, and find p at $(1, 0.5)$ for $c = 3$ and $c = 10$. Over what distance across the boundary does p go from $\sigma(-1)$ to $\sigma(1)$?

$cw^\top x + cb = 0$ exactly when $w^\top x + b = 0$	Divide by $c > 0$.
$c = 3$: $z = 1.5$, $p \approx 0.8176$; $c = 10$: $z = 5$, $p \approx 0.9933$	The logit scales with c.
$z = \pm 1$ at signed distances $\pm 1/(c\|w\|)$	The logit is $c\|w\|$ times the signed distance (Problem 3).
The boundary is the same for every $c > 0$; $p \approx 0.8176$ for $c = 3$ and 0.9933 for $c = 10$; p goes from 0.269 to 0.731 over a band of width $2/(c\|w\|)$, which is $2/\sqrt{5} \approx 0.894$ for $c = 1$	The direction of w and the ratio $b/\|w\|$ fix where the boundary is; the size of w fixes how sharp it is. A larger $\|w\|$ is a more confident model with the same predictions at $\tau = \frac{1}{2}$.

Problem 6

Train a one-feature model without bias on two points: $x = -1$ with $y = 0$ and $x = 1$ with $y = 1$. Show that the mean loss is $L(w) = \ln(1 + e^{-w})$, that it has no minimizer, and what gradient descent does. Then add the penalty $\frac{\lambda}{2} w^2$ and find the minimizer for $\lambda = 0.1$.

At $x = 1$: $-\ln \sigma(w) = \ln(1 + e^{-w})$	Class 1 with logit w.
At $x = -1$: $-\ln(1 - \sigma(-w)) = -\ln \sigma(w) = \ln(1 + e^{-w})$	Class 0 with logit $-w$, and $1 - \sigma(-w) = \sigma(w)$.
$L(w) = \ln(1 + e^{-w})$ and $L'(w) = -\sigma(-w) < 0$	Both points give the same loss; differentiate with the chain rule.
L decreases for every w and tends to 0 without reaching it	No finite w is a minimizer, and gradient descent increases w forever, more and more slowly.

Penalized: $L'(w) + \lambda w = 0$ gives $\sigma(-w) = \lambda w$	The gradient of $\frac{\lambda}{2}w^2$ is λw.
For $\lambda = 0.1$: $w \approx 1.634$	Solve numerically; the left side falls and the right side rises, so there is one root.
$L(w) = \ln(1 + e^{-w})$ **has no minimizer, and gradient descent sends** w **to infinity; with** $\lambda = 0.1$ **the minimizer is** $w \approx 1.634$	On separable data the boundary is already right, and the loss rewards only confidence, so $\|w\|$ grows without bound (Problem 5). A ridge penalty, early stopping or label smoothing keeps it finite; with $\lambda = 0.01$ the minimizer is $w \approx 3.36$.

Problem 7

> Class 0 has features $x \sim \text{Normal}(\mu_0, \sigma^2)$ and class 1 has $x \sim \text{Normal}(\mu_1, \sigma^2)$, with prior probabilities π_0 and π_1. Show that $\ln \dfrac{P(y=1 \mid x)}{P(y=0 \mid x)}$ is linear in x, and find w and b. Find the boundary for $\mu_0 = 0$, $\mu_1 = 2$, $\sigma = 1$ with equal priors, and with $\pi_1 = 0.2$.

$\ln \dfrac{P(y=1 \mid x)}{P(y=0 \mid x)} = \ln \dfrac{\pi_1}{\pi_0} + \dfrac{(x-\mu_0)^2 - (x-\mu_1)^2}{2\sigma^2}$	Bayes' rule; the normalizing constants of the two densities are equal and cancel.
$(x-\mu_0)^2 - (x-\mu_1)^2 = 2x(\mu_1 - \mu_0) + \mu_0^2 - \mu_1^2$	The x^2 terms cancel because the variances are equal.
$w = \dfrac{\mu_1 - \mu_0}{\sigma^2}$ and $b = \dfrac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln \dfrac{\pi_1}{\pi_0}$	Collect the coefficient of x and the constant.
Equal priors: $w = 2$, $b = -2$, boundary $x = 1$	Halfway between the means.
$\pi_1 = 0.2$: $b = -2 + \ln \frac{0.2}{0.8} = -2 - \ln 4 \approx -3.386$	The log prior odds shift the bias.

$P(y=1 \mid x) = \sigma(wx+b)$ **with** $w = (\mu_1-\mu_0)/\sigma^2$ **and** $b = (\mu_0^2-\mu_1^2)/(2\sigma^2) + \ln(\pi_1/\pi_0)$; **the boundary is** $x=1$ **for equal priors and** $x = 1 + \frac{\ln 4}{2} \approx 1.693$ **for** $\pi_1 = 0.2$	Two Gaussian classes with a shared variance give exactly a logistic posterior, which is one reason the model works well beyond this case. A rarer class pushes the boundary toward its own mean (Mistake 5). With unequal variances the x^2 terms survive and the boundary is quadratic.

Problem 8

> The four points $(0,0)$ and $(1,1)$ of class 0 and $(0,1)$ and $(1,0)$ of class 1 are the XOR pattern. Show that no line separates them. Then add the feature $x_3 = x_1 x_2$ and check that $z = x_1 + x_2 - 2x_3 - 0.5$ separates them.

Suppose $z = w_1x_1 + w_2x_2 + b$ separates them, with $z > 0$ on class 1 and $z < 0$ on class 0	A separating line.
Class 0: $b < 0$ and $w_1 + w_2 + b < 0$, which add to $w_1 + w_2 + 2b < 0$	$(0,0)$ and $(1,1)$.
Class 1: $w_2 + b > 0$ and $w_1 + b > 0$, which add to $w_1 + w_2 + 2b > 0$	$(0,1)$ and $(1,0)$.
Steps 2 and 3 contradict each other, so no line works	The two diagonals of the square cross, and a line cannot put both ends of one diagonal on one side and both ends of the other on the other side.
With $x_3 = x_1x_2$: $z(0,0) = -0.5$ and $z(1,1) = 1 + 1 - 2 - 0.5 = -0.5$	Class 0, both negative.
$z(0,1) = z(1,0) = 0.5$	Class 1, both positive.

No linear boundary in (x_1, x_2) separates XOR; with the product feature, $z = x_1 + x_2 - 2x_1x_2 - 0.5$ does	The boundary is linear in (x_1, x_2, x_3) and curved, a hyperbola, in the original plane. A hidden layer learns features like x_3 instead of having them supplied.

Problem 9

A model uses the features x_1^2 and x_2^2, with logit $z = x_1^2 + x_2^2 - 4$. Describe the decision boundary at $\tau = \frac{1}{2}$ and the class-1 region, and find p at $(0, 0)$, $(1, 1)$ and $(2, 2)$.

$z = 0$ on $x_1^2 + x_2^2 = 4$	A circle of radius 2 about the origin.
Class 1 where $x_1^2 + x_2^2 > 4$, outside the circle	z grows with the distance from the origin.
$p(0, 0) = \sigma(-4) \approx 0.0180$	The center is the most confidently class 0.
$p(1, 1) = \sigma(-2) \approx 0.1192$ and $p(2, 2) = \sigma(4) \approx 0.9820$	$1 + 1 - 4 = -2$ and $4 + 4 - 4 = 4$.
The boundary is the circle $x_1^2 + x_2^2 = 4$, with class 1 outside it; $p \approx 0.018$, 0.119 and 0.982	In the feature space (x_1^2, x_2^2) the boundary is the straight line $u + v = 4$ with normal $(1, 1)$. A model linear in its features can draw any boundary those features can express.

Problem 10

A three-class softmax model has $w_1 = (1, 0)$, $w_2 = (0, 1)$, $w_3 = (0, 0)$ and all biases 0. Find the boundary between each pair of classes, describe the region of each class, and find the probabilities at $x = (2, 1)$.

Classes i and j tie where $z_i = z_j$, that is $(w_i - w_j)^\top x + (b_i - b_j) = 0$	Equal logits give equal probabilities; the softmax denominator is shared.

1 vs 2: $x_1 = x_2$; 1 vs 3: $x_1 = 0$; 2 vs 3: $x_2 = 0$	$w_1 - w_2 = (1, -1)$, $w_1 - w_3 = (1, 0)$, $w_2 - w_3 = (0, 1)$.
Class 1 where $x_1 > x_2$ and $x_1 > 0$; class 2 where $x_2 > x_1$ and $x_2 > 0$; class 3 where $x_1 < 0$ and $x_2 < 0$	The predicted class has the largest logit, so it must beat both others.
At $(2, 1)$ the logits are $(2, 1, 0)$	$z_i = w_i^\top x$.
$\text{softmax}(2, 1, 0) = \dfrac{(e^2, e, 1)}{e^2 + e + 1}$	The denominator is about 11.107.
Boundaries $x_1 = x_2$, $x_1 = 0$ and $x_2 = 0$, meeting at the origin; class 3 owns the negative quadrant and classes 1 and 2 split the rest along $x_1 = x_2$; at $(2, 1)$ the probabilities are $(0.665, 0.245, 0.090)$	Only differences of weights matter: adding the same vector to every w_i changes no boundary and no probability. With two classes the one boundary is $(w_1 - w_2)^\top x + (b_1 - b_2) = 0$, which is logistic regression with $w = w_1 - w_2$.

Where this goes wrong

1. Multiplying a probability by an odds ratio

Problem 2 says a unit step in x_1 multiplies something by e^2, and the probability is the number on the screen.

At $(1, 0.5)$, $p \approx 0.6225$, and $w_1 = 2$	Right so far.
"One more unit of x_1 multiplies the probability by e^2."	The slip that causes the mistake: e^{w_j} multiplies the odds $p/(1-p)$, not p.
$p(2, 0.5) \approx 0.6225 \cdot 7.389 \approx 4.60$ ✗	A probability above 1 is the clue. The odds go from 1.649 to 12.18, and the probability from 0.6225 to 0.9241.

2. Leaving out the norm in a distance

The logit is zero on the boundary and grows away from it, so it looks like a distance.

$(1, 0.5)$ has logit $z = 0.5$	Right so far: Problem 1.
"So it lies 0.5 from the boundary."	The slip that causes the mistake: the logit is $\|w\|$ times the distance, and $\|w\| = \sqrt{5}$ here.
distance $= 0.5$ ✗	The distance is $0.5/\sqrt{5} \approx 0.2236$. Doubling w and b would double this "distance" without moving the point or the line (Problem 5).

3. Putting the threshold on the logit

A threshold of 0.8 is a probability, and the boundary equation has a number on its right side.

Predict class 1 when $p \geq 0.8$	Right so far: Problem 4.
"The boundary moves to $2x_1 - x_2 - 1 = 0.8$."	The slip that causes the mistake: the right side is a value of the logit, so it is $\ln(0.8/0.2) = \ln 4 \approx 1.386$.
predict class 1 when $z \geq 0.8$ ✗	At $x = (1, 0)$, $z = 1$ and $p = \sigma(1) \approx 0.731$, below 0.8, yet this rule predicts class 1. The two scales differ even at $\tau = \frac{1}{2}$, whose cutoff on the logit is 0, not 0.5.

4. Taking a small gradient as convergence

Training stops when the gradient is tiny, and on the separable data of Problem 6 it gets tiny quickly.

$L'(w) = -\sigma(-w)$	Right so far: Problem 6, step 3.
"At $w = 10$ the gradient is -4.5×10^{-5}, so w has converged."	The habit that causes the mistake: reading a small gradient as a nearby minimum, when this loss has no minimum.

$w \approx 10$ is the fitted weight ✗	At $w = 20$ the loss falls from 4.5×10^{-5} to 2.1×10^{-9}, and it keeps falling forever. The reported weight depends only on when training stopped, and so do the model's confidences.

5. Placing the boundary halfway between the classes

With equal variances the boundary of Problem 7 sits halfway between the means, and the priors are easy to forget.

Means 0 and 2, variance 1, $\pi_1 = 0.2$	Right so far.
"The boundary is at the midpoint, $x = 1$."	The term that is missing: the log prior odds $\ln(\pi_1/\pi_0) = -\ln 4$ in the bias.
predict class 1 for $x \geq 1$ ✗	At $x = 1$ the posterior is $P(y = 1 \mid x) = \sigma(-\ln 4) = 0.2$, not $\frac{1}{2}$. The even-odds point is $x \approx 1.693$, pushed toward the rarer class's mean.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 8 of *Deep Learning Math by Hand*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/logistic-regression-and-decision-boundaries