

Logistic Regression — Practice Questions

CM52054: Foundational Machine Learning
Practice set with fully worked answers

Conceptual Questions

1) **Binary classification setup.**

Formally define the binary classification setting used for logistic regression. Specify:

- a) The form of the dataset D .
- b) The domain and codomain of inputs x and outputs y .
- c) The learning goal.

Answer:

- a) The dataset is

$$D = \{(x_1, y_1), \dots, (x_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}.$$

- b) Each input $x_i \in \mathbb{R}^M$ is a feature vector; each output $y_i \in \{0, 1\}$ is a binary label.
- c) The goal is to learn parameters w of a model such that, given a new feature vector x_{N+1} , we can predict the probability of the Boolean outcome and then the class label $y_{N+1} \in \{0, 1\}$.

2) **Logistic (sigmoid) function.**

Give the mathematical definition of the logistic (sigmoid) function $\sigma(x)$. What is its range and why is that range useful for classification?

Answer: The logistic function is

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}.$$

Its range is $(0, 1)$, which makes the output interpretable as a probability, i.e. $p \in (0, 1)$ for a binary event such as “class 1” vs “class 0”. This is why it is used in logistic regression instead of a linear output.

3) **Derivative of the sigmoid.**

Show that the derivative of the logistic function satisfies

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

Answer: Start from

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

Differentiate:

$$\begin{aligned}\sigma'(x) &= \frac{d}{dx}(1 + e^{-x})^{-1} \\ &= -(1 + e^{-x})^{-2} \cdot \frac{d}{dx}(1 + e^{-x}) \\ &= -(1 + e^{-x})^{-2} \cdot (-e^{-x}) \\ &= \frac{e^{-x}}{(1 + e^{-x})^2}.\end{aligned}$$

Now rewrite numerator and denominator using $\sigma(x)$:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \Rightarrow \quad 1 - \sigma(x) = 1 - \frac{1}{1 + e^{-x}} = \frac{e^{-x}}{1 + e^{-x}}.$$

Hence

$$\sigma(x)(1 - \sigma(x)) = \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} = \frac{e^{-x}}{(1 + e^{-x})^2} = \sigma'(x).$$

4) **Logit (inverse sigmoid) function.**

Derive the inverse of the sigmoid function, i.e. express x as a function of $p = \sigma(x)$. This inverse is called the logit function.

Answer: Let $p = \sigma(x) = \frac{1}{1 + e^{-x}}$. Then

$$p(1 + e^{-x}) = 1 \quad \Rightarrow \quad p + pe^{-x} = 1 \quad \Rightarrow \quad pe^{-x} = 1 - p.$$

So

$$e^{-x} = \frac{1 - p}{p} \quad \Rightarrow \quad -x = \ln\left(\frac{1 - p}{p}\right) \quad \Rightarrow \quad x = \ln\left(\frac{p}{1 - p}\right).$$

Thus the logit function is

$$\text{logit}(p) = \sigma^{-1}(p) = \ln \frac{p}{1 - p},$$

mapping $p \in (0, 1)$ back to \mathbb{R} .

5) **Logistic regression model.**

Write down the logistic regression model for a feature vector $x \in \mathbb{R}^M$ using the sigmoid function. How is it related to linear regression?

Answer: First define a linear score

$$z = w^\top x = w_0x_0 + w_1x_1 + \cdots + w_Mx_M,$$

with $x_0 = 1$ as the bias term. In linear regression, this z itself would be the output.

In logistic regression we pass this through a sigmoid:

$$f_w(x) = \sigma(w^\top x) = \frac{1}{1 + e^{-w^\top x}},$$

which yields a probability in $(0, 1)$. So logistic regression is a linear model in feature space followed by the nonlinear sigmoid activation.

6) **Class probabilities in logistic regression.**

Explain how logistic regression models class probabilities $p(y = 1 \mid x; w)$ and $p(y = 0 \mid x; w)$.

Answer: Define $f_w(x) = \sigma(w^\top x)$. Then

$$p(y = 1 \mid x; w) = f_w(x), \quad p(y = 0 \mid x; w) = 1 - f_w(x).$$

In compact form, for $y \in \{0, 1\}$,

$$p(y \mid x; w) = [f_w(x)]^y [1 - f_w(x)]^{1-y}.$$

7) **Likelihood for logistic regression.**

Given a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, write down the likelihood $L(w)$ of the logistic regression model parameters w . Assume samples are i.i.d.

Answer: Using

$$p(y_i \mid x_i; w) = [f_w(x_i)]^{y_i} [1 - f_w(x_i)]^{1-y_i},$$

the likelihood is

$$L(w) = \prod_{i=1}^N p(y_i \mid x_i; w) = \prod_{i=1}^N [f_w(x_i)]^{y_i} [1 - f_w(x_i)]^{1-y_i}.$$

8) **Log-likelihood and its advantages.**

Derive the log-likelihood $\log L(w)$ from the likelihood in the previous question and explain why we prefer to work with log-likelihood.

Answer: Take the logarithm:

$$\begin{aligned} \log L(w) &= \log \prod_{i=1}^N [f_w(x_i)]^{y_i} [1 - f_w(x_i)]^{1-y_i} \\ &= \sum_{i=1}^N \left(y_i \log f_w(x_i) + (1 - y_i) \log [1 - f_w(x_i)] \right). \end{aligned}$$

We use the log-likelihood because:

- It converts a product of probabilities into a sum, which is numerically more stable and simpler to differentiate.
- Maximising $\log L(w)$ is equivalent to maximising $L(w)$ since the logarithm is monotonic.

Gradients & Optimisation

9) **Gradient for a single data point.**

For a single data point (x, y) , show that the gradient of the log-likelihood with respect to w has the form

$$\nabla_w \log p(y \mid x; w) = (y - f_w(x)) x.$$

Answer: For one sample,

$$\log p(y \mid x; w) = y \log f_w(x) + (1 - y) \log(1 - f_w(x)).$$

Let $f = f_w(x) = \sigma(w^\top x)$. Then

$$\nabla_w \log p(y | x; w) = \left(\frac{y}{f} - \frac{1-y}{1-f} \right) \nabla_w f.$$

Compute $\nabla_w f$. Since $f = \sigma(z)$ with $z = w^\top x$,

$$\frac{\partial f}{\partial z} = f(1-f), \quad \nabla_w z = x,$$

so

$$\nabla_w f = f(1-f)x.$$

Thus

$$\begin{aligned} \nabla_w \log p(y | x; w) &= \left(\frac{y}{f} - \frac{1-y}{1-f} \right) f(1-f)x \\ &= (y(1-f) - (1-y)f)x \\ &= (y - yf - f + yf)x \\ &= (y - f)x. \end{aligned}$$

Entry-wise, for the j -th component,

$$\frac{\partial}{\partial w_j} \log p(y | x; w) = (y - f_w(x))x_j.$$

10) **Gradient of the total log-likelihood.**

Using the previous result, derive the gradient of the total log-likelihood $\log L(w)$ over all N samples.

Answer: We have

$$\log L(w) = \sum_{i=1}^N \log p(y_i | x_i; w).$$

Using linearity of the gradient and the single-sample result:

$$\nabla_w \log L(w) = \sum_{i=1}^N \nabla_w \log p(y_i | x_i; w) = \sum_{i=1}^N (y_i - f_w(x_i))x_i.$$

11) **Gradient ascent update rule.**

State the gradient ascent update rule for logistic regression when we maximise the log-likelihood. Explain the roles of w_t , α_t , and p_t .

Answer: The generic update is

$$w_{t+1} = w_t + \alpha_t p_t,$$

where

- w_t is the parameter vector at iteration t .
- $\alpha_t > 0$ is the step size (learning rate).
- $p_t = \nabla_w \log L(w_t)$ is the gradient direction:

$$p_t = \sum_{i=1}^N (y_i - f_{w_t}(x_i))x_i.$$

Because we are maximising $\log L(w)$, we use gradient ascent (adding the gradient). If we were minimising a loss (e.g. negative log-likelihood), we would subtract the gradient.

12) **Single-step gradient ascent (computational).**

Consider a single training example with feature vector $x = (1, 2)^\top$ and label $y = 1$. Suppose current parameters $w = (0, 1)^\top$.

- a) Compute $f_w(x)$.
- b) Compute the gradient $\nabla_w \log p(y | x; w)$.
- c) Using learning rate $\alpha = 0.1$, compute the updated w' after one gradient ascent step on this single example.

Answer:

- a) Linear score:

$$z = w^\top x = 0 \cdot 1 + 1 \cdot 2 = 2.$$

So

$$f_w(x) = \sigma(2) = \frac{1}{1 + e^{-2}} \approx 0.881.$$

- b) From the single-sample gradient:

$$\nabla_w \log p(y | x; w) = (y - f_w(x))x = (1 - 0.881)(1, 2)^\top \approx 0.119(1, 2)^\top \approx (0.119, 0.238)^\top.$$

- c) Update:

$$w' = w + \alpha \nabla_w \log p(y | x; w) = (0, 1)^\top + 0.1 \cdot (0.119, 0.238)^\top \approx (0.0119, 1.0238)^\top.$$

13) **Decision boundary from logistic regression.**

For a logistic regression classifier with model $f_w(x) = \sigma(w^\top x)$, show that the decision boundary for predicting class 1 vs class 0 at threshold 0.5 is linear. Write its equation.

Answer: We predict class 1 if

$$p(y = 1 | x; w) = f_w(x) = \sigma(w^\top x) \geq 0.5.$$

The sigmoid is monotonic increasing, and $\sigma(z) = 0.5$ when $z = 0$. Thus:

$$\sigma(w^\top x) \geq 0.5 \quad \Leftrightarrow \quad w^\top x \geq 0.$$

So the decision boundary is given by

$$w^\top x = 0,$$

which is a linear hyperplane (a straight line in 2D).

Interpretation & Applications

14) **Interpreting parameters in 1D.**

In a 1D case with $f_w(x) = \sigma(w_0 + w_1 x)$, interpret the parameters w_0 and w_1 in terms of the shape and position of the curve.

Answer:

- w_0 (bias/intercept) horizontally shifts the logistic curve along the x-axis. Changing w_0 moves the point where $p(y = 1 | x; w) = 0.5$.

- w_1 determines the slope (steepness and direction):
 - Larger $|w_1|$ makes the transition between probabilities near 0 and 1 steeper.
 - The sign of w_1 controls whether the curve is increasing ($w_1 > 0$) or decreasing ($w_1 < 0$).

15) **Probability calculation in 1D.**

A learned logistic regression model for a single feature is:

$$p(y = 1 \mid x) = \sigma(-3 + 2x).$$

- Compute $p(y = 1 \mid x = 2)$.
- For which x -value is $p(y = 1 \mid x) = 0.5$?
- For which region of x does the classifier predict class 1 if we use a 0.5 threshold?

Answer:

- Plug in $x = 2$:

$$z = -3 + 2 \cdot 2 = 1, \quad p(y = 1 \mid x = 2) = \sigma(1) = \frac{1}{1 + e^{-1}} \approx 0.731.$$

- Set the logit to zero:

$$-3 + 2x = 0 \Rightarrow 2x = 3 \Rightarrow x = 1.5.$$

At $x = 1.5$, $p = 0.5$.

- Decision rule with threshold 0.5:

$$\sigma(-3 + 2x) \geq 0.5 \Leftrightarrow -3 + 2x \geq 0 \Leftrightarrow x \geq 1.5.$$

So class 1 is predicted for $x \geq 1.5$; class 0 for $x < 1.5$.

16) **Interpreting a 2D decision boundary.**

Suppose your learned parameters are $w_0 = -2$, $w_1 = 1$, $w_2 = 3$ with $x = (1, x_1, x_2)^\top$.

- Write the equation of the decision boundary in terms of x_1, x_2 .
- Rearrange it to express x_2 as a linear function of x_1 .
- If a point has coordinates $(x_1, x_2) = (1, 1)$, which class would be predicted at threshold 0.5?

Answer:

- Decision boundary: $w^\top x = 0$:

$$-2 + 1 \cdot x_1 + 3 \cdot x_2 = 0.$$

- Solve for x_2 :

$$3x_2 = 2 - x_1 \Rightarrow x_2 = \frac{2 - x_1}{3}.$$

- For $(1, 1)$,

$$z = -2 + 1 \cdot 1 + 3 \cdot 1 = 2.$$

Then $p(y = 1 \mid x) = \sigma(2) \approx 0.881 > 0.5$, so we predict class 1.

17) **Comparing logistic and linear regression.**

Briefly compare logistic regression to linear regression in terms of:

- a) Type of output.
- b) Typical loss/optimisation objective.
- c) Suitability for classification.

Answer:

- a) Linear regression outputs a real-valued prediction (unbounded). Logistic regression outputs a probability in $(0, 1)$ via the sigmoid.
- b) Linear regression typically minimises mean squared error. Logistic regression typically maximises log-likelihood (equivalently minimises negative log-likelihood).
- c) Logistic regression is designed for binary classification tasks (and can be extended to multi-class), whereas linear regression is unsuitable for directly modelling probabilities or discrete labels because its outputs are unbounded and not probabilistic.