Discussion Session Problems 2

01/22/2026

1. Which of the following methods do we use to best fit the data in Logistic Regression?

   A) Least Squares Error

   B) Maximum Likelihood

   C) Both A and B

   Answer: B)
   Logistic regression is based on maximum likelihood estimation.

2. You are given a coin whose probability of landing on Heads is p. We toss the coin 10 times and observe 7 Heads. What is the most likely value of p?

   A) 7/10

   B) 5/10

   C) 3/10

   Answer: A)

3. Which of the following evaluation metrics does **not** make sense if applied to logistic regression output to compare with the target?

   A) Accuracy

   B) Log loss

   C) Mean Squared Error

   Answer: C)
   Mean squared error is appropriate for linear regression, not logistic regression.

4. What can you say about feature normalization in logistic regression?

   A) It is good practice but not strictly required

   B) It is required

   C) It is bad practice and should not be performed

   D) None of the above

   Answer: A)
   Feature normalization helps gradient-based optimization converge faster. It is not strictly required for logistic regression, but it is required when using regularization.

5. Consider m independent and identically distributed (i.i.d.) random variables $x_1, \ldots, x_m$ drawn from a Geometric distribution with parameter p. In other words, $x_i \sim Geo(p)$ for all $1 \leq i \leq m$.

(a) Derive the maximum likelihood estimate (MLE) for p.

The probability mass function of a Geometric random variable with parameter p is

$$f(X_i \mid p) = p(1-p)^{X_i - 1}, \quad X_i \geq 1.$$

Assuming m independent observations $X_1, X_2, \dots, X_m$, the likelihood function is

$$L(p) = \prod_{i=1}^{m} p(1-p)^{X_i - 1}.$$

Taking the logarithm of the likelihood, we obtain the log-likelihood:

$$LL(p) = \sum_{i=1}^{m} \left[\log p + (X_i - 1)\log(1-p)\right].$$

We now take the derivative of $LL(p)$ with respect to p and set it equal to zero:

$$\frac{dLL(p)}{dp} = \sum_{i=1}^{m} \left(\frac{1}{p} - \frac{X_i - 1}{1-p}\right) = 0.$$

Solving for p, we get

$$\frac{m}{p} = \frac{1}{1-p} \sum_{i=1}^{m} (X_i - 1).$$

Rewriting,

$$\frac{1-p}{p} = \frac{1}{m} \sum_{i=1}^{m} (X_i - 1),$$

which implies

$$\frac{1}{p} - 1 = \frac{1}{m} \sum_{i=1}^{m} (X_i - 1).$$

Therefore, the maximum likelihood estimate of p is

$$\boxed{\hat{p}_{\text{MLE}} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} X_i} = \frac{1}{\bar{X}}}$$

(b) Given $x = \{4, 3, 4, 2, 7\}$, compute $\hat{p}_{MLE}$. Namely, the value of p in the Geometric distribution that would maximize the likelihood of these observations.

$$\hat{p} = \frac{1}{(4+3+4+2+7)/5} = \frac{5}{20} = 0.25$$

6. In this problem, we simultaneously estimate the difficulty of problem set questions and the skill level of each student.

Consider a set of 200 students and 10 questions, where each student answers each question. Let S_{ij} be an indicator variable such that

$$S_{ij} = \begin{cases} 1 & \text{if student } i \text{ answers question } j \text{ correctly} \\ 0 & \text{otherwise} \end{cases}$$

We assume that the probability that student i answers question j correctly is

$$p_{ij} = \sigma(a_i - d_j),$$

where:

- $\sigma(\cdot)$ is the sigmoid function,
- a_i represents the ability of student i,
- d_j represents the difficulty of question j.

We use Maximum Likelihood Estimation (MLE) to estimate all parameters.

(a) Write the log-likelihood for a single response S_{ij} in terms of p_{ij}. *Hint: logistic regression also assumes that its output is a probability of a binary event.*

(b) Compute the partial derivative of the log-likelihood for a single response S_{ij} with respect to a_i.

(c) Compute the partial derivative of the log-likelihood for a single response S_{ij} with respect to d_j.

(d) Explain briefly how the parameters can be estimated using derivatives of log-likelihood with respect to those parameters.

(a) Log-likelihood for a single response

$$\text{Liklihood} = p_{ij}^{S_{ij}}(1-p_{ij})^{1-S_{ij}}.$$

$$\text{Log Likelihood} = S_{ij}\log(p_{ij}) + (1-S_{ij})\log(1-p_{ij}).$$

(b) Partial derivative with respect to a_i

For a single response S_{ij}, the log-likelihood is

$$\text{LL}_{ij} = S_{ij}\log(p_{ij}) + (1-S_{ij})\log(1-p_{ij}).$$

We apply the chain rule:

$$\frac{\partial \text{LL}_{ij}}{\partial a_i} = \frac{\partial \text{LL}_{ij}}{\partial p_{ij}} \cdot \frac{\partial p_{ij}}{\partial a_i}.$$

First,

$$\frac{\partial \text{LL}_{ij}}{\partial p_{ij}} = \frac{S_{ij}}{p_{ij}} - \frac{1-S_{ij}}{1-p_{ij}}.$$

Since $p_{ij} = \sigma(a_i - d_j)$ and $\sigma'(x) = \sigma(x)(1-\sigma(x))$,

$$\frac{\partial p_{ij}}{\partial a_i} = p_{ij}(1-p_{ij}).$$

Multiplying,

$$\frac{\partial \text{LL}_{ij}}{\partial a_i} = \left(\frac{S_{ij}}{p_{ij}} - \frac{1-S_{ij}}{1-p_{ij}}\right) p_{ij}(1-p_{ij}) = S_{ij} - p_{ij}.$$

$$\boxed{\frac{\partial \text{LL}_{ij}}{\partial a_i} = S_{ij} - p_{ij}}$$

(c) Partial derivative with respect to d_j

Again applying the chain rule,

$$\frac{\partial \text{LL}_{ij}}{\partial d_j} = \frac{\partial \text{LL}_{ij}}{\partial p_{ij}} \cdot \frac{\partial p_{ij}}{\partial d_j}.$$

From above,

$$\frac{\partial \mathrm{LL}_{ij}}{\partial p_{ij}} = \frac{S_{ij}}{p_{ij}} - \frac{1 - S_{ij}}{1 - p_{ij}}.$$

Since

$$\frac{\partial (a_i - d_j)}{\partial d_j} = -1,$$

we have

$$\frac{\partial p_{ij}}{\partial d_j} = -\, p_{ij}(1 - p_{ij}).$$

Thus,

$$\frac{\partial \mathrm{LL}_{ij}}{\partial d_j} = -\left(\frac{S_{ij}}{p_{ij}} - \frac{1 - S_{ij}}{1 - p_{ij}}\right) p_{ij}(1 - p_{ij}) = -(S_{ij} - p_{ij}).$$

$$\boxed{\frac{\partial \mathrm{LL}_{ij}}{\partial d_j} = -(S_{ij} - p_{ij})}$$

(d) Parameter estimation

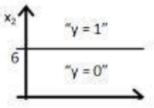
The parameters $\{a_i\}$ and $\{d_j\}$ can be estimated using gradient ascent. At each iteration, we update the parameters in the direction of the gradient using a fixed learning rate until convergence. Just like when we implemented logistic regression, we can program our closed form mathematical solution for gradients to efficiently calculate the gradient for any values of our parameters.

7. Suppose a logistic regression model is

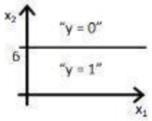
$$f_{w,b}(x) = g(w_1 x_1 + w_2 x_2 + b)$$

with $b = 6$, $w_1 = 0$, $w_2 = -1$, and g is the sigmoid function. Which figure corresponds to the decision boundary?

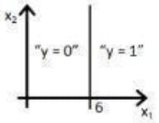
A)



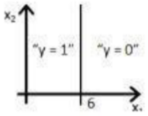
B)



C)



D)



Answer: B)
The decision boundary is defined by:

$$w_1 x_1 + w_2 x_2 + b = 0 \Rightarrow 0 \cdot x_1 + (-1) \cdot x_2 + 6 = 0 \Rightarrow x_2 = 6$$

The boundary is the horizontal line $x_2 = 6$ in options A and B. Option B is the right answer because when you put the value $x_2 > 6$ in the equation $y = g(-x_2 + 6)$, you will get values closer to 0, so the output will be the region $y = 0$.

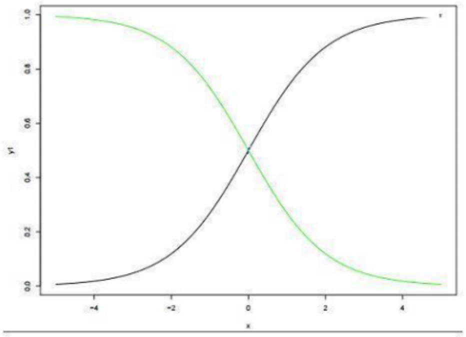
8. If w is constant, what is the slope of the logistic function at $x = 0$ for $g(wx)$?

   A) w
   B) $w/4$
   C) $1/4$
   D) w^2

Answer: B)
Since $g'(z) = g(z)(1 - g(z))$ and $g(0) = 1/2$, we know that $g'(0) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. If we take the derivative with respect to x, we get:

$$\frac{d}{dx} g(wx)\Big|_{x=0} = \frac{w}{4}$$

9. Consider $f_{w,b} = g(w_1 X + b)$, where w_1 is the coefficient and b is the intercept. Below are two different logistic models with different values for w_1 and b. Which of the following statement(s) is true? The model represented by the green line starts on top.



- w_1 for Green is greater than Black
- w_1 for Green is smaller than Black
- w_1 for both models is the same
- We cannot say for certain the relationship between the w_1 for Green and the w_1 for Black.

Answer: B)

10. What can you say about regularized vs. non-regularized logistic regression?

A) It will perform better on the training set
B) We expect it to perform better on the training set
C) It will perform better on the test set
D) We expect it to perform better on the test set

Answer: D)
Regularization reduces overfitting, so we expect improved generalization performance on unseen data.

11. In logistic regression, what happens to the predicted probability $p = \sigma(z)$ as $z \to +\infty$ and $z \to -\infty$?

A) $p \to 0$ as $z \to +\infty$, $p \to 1$ as $z \to -\infty$
B) $p \to 1$ as $z \to +\infty$, $p \to 0$ as $z \to -\infty$
C) $p \to 0.5$ in both cases
D) p oscillates between 0 and 1

Answer: B)
The sigmoid function asymptotically approaches 1 for large positive inputs and 0 for large negative inputs.

12. Suppose a dataset is perfectly linearly separable. What can happen to the magnitude of the weights learned by unregularized logistic regression?

    A) The weights converge to zero

    B) The weights converge to a finite value

    C) The weights can grow arbitrarily large

    D) The algorithm fails immediately

    Answer: C)
    For linearly separable data, the maximum likelihood solution drives the weights toward infinity unless regularization is used.

13. Suppose a logistic regression model outputs probabilities very close to 0 or 1 for most training examples. Which of the following is the most likely explanation?

    A) The model is underfitting

    B) The model weights have very small magnitude

    C) The model is highly confident in its predictions

    D) The learning rate is too small

    Answer: C)
    Probabilities near 0 or 1 indicate that the model is very confident in its predictions.

14. Which of the following best explains why we take the logarithm of the likelihood in logistic regression?

    A) To make the likelihood larger

    B) To convert products over data points into sums

    C) To ensure predictions lie between 0 and 1

    D) To remove the need for regularization

    Answer: B)
    Taking the log turns a product of probabilities into a sum, which simplifies optimization and numerical stability.

15. Consider a binary classification problem with a very imbalanced dataset (e.g., 99% of labels are 0). Which issue is most likely to arise when training logistic regression?

    A) The model cannot be trained using gradient descent

    B) Accuracy may be misleading as an evaluation metric

    C) The sigmoid function becomes non-differentiable

    D) Regularization has no effect

    Answer: B)
    In highly imbalanced datasets, a model that always predicts the majority class can achieve high accuracy while performing poorly.

16. Which of the following statements about the intercept (bias) term in logistic regression is true?

    A) It controls the slope of the decision boundary

    B) It shifts the decision boundary without changing its orientation

    C) It is unnecessary if features are normalized

    D) It must always be regularized

Answer: B)
The intercept shifts the decision boundary while the weights control its orientation.

17. Suppose we increase the strength of L_2 regularization in logistic regression. Which of the following effects do we expect?

    A) Training loss decreases

    B) Weight magnitudes decrease

    C) The decision boundary becomes more complex

    D) The model fits noise more closely

    Answer: B)
    Stronger L_2 regularization penalizes large weights, encouraging smaller parameter values.