Week_5_mlt_questions

May 2025 Quiz 1

1. For a linear regression problem, which among the following are true? Select all that apply. (MSQ)

(a) Gradient descent will converge to a global minimum.

(b) Gradient descent will converge to a local minimum (which is not a global minimum).

(c) Stochastic gradient descent is guaranteed to converge to a global minimum.

(d) Gradient descent will take lesser number of iterations to converge to a local minimum than stochastic gradient descent.

[Ans: a, d]

2. **Comprehension:**
Consider the following dataset with feature vector $\mathbf{x}$ and label y:

$\mathbf{x}$	y
$[4 \quad 1]^T$	2
$[2 \quad 8]^T$	-14
$[1 \quad 4]^T$	-7
$[3 \quad 2]^T$	-1

Based on the above data, answer the given subquestions.

i. Fit the linear regression model $y = wx,\ w = [w_1 \quad w_2]^T$ using the squared error.

(a) $y = x_1 + 2x_2$

(b) $y = x_1 - 2x_2$

(c) $y = x_1 - x_2$

(d) $y = x_1 + x_2$

[Ans: b]

ii. Find the mean squared error for the training dataset.

[Ans: 0]

2024 Mar 24 Quiz 2

3. **Comprehension:**
Consider a dataset with the following data points and the target variable:

Sample No	x	y
1	3	8
2	0	3
3	5	12
4	6	13

The linear regression model is given by $y = w_0 + w_1 x$. Assume that the Leave-One-Out Cross-Validation technique is applied.
Based on the above data, answer the given subquestions.

i. Enter the value of w_1 obtained when the 3rd sample is used as the validation data point.

[Ans: 1.65 to 1.70]

ii. What will be the predicted value for the left-out data point?

(a) 12
(b) 13
(c) 12.3
(d) 11.3
(e) None of these

[Ans: d]

4. Kernel regression with a polynomial kernel of degree three is applied on a data set $\{X, y\}$. Let the weight vector be given by

$$w = \phi(X)[2.3, -1.0, 0.4, -0.7]^T$$

Here $\phi(X)$ is the transformed data matrix whose i^{th} column is $\phi(x_i)$. What will be the prediction for the data point $[0, 0, 0, 0]^T$?

[Ans: 1]

5. Let X be the data matrix of shape (d, n) and y be the corresponding label vector. A linear regression model of the form $\hat{y}_i = w^T x_i$ is fit using the squared error on the same dataset. If the solution w^* to the optimization problem is orthogonal to the subspace spanned of the data point (columns of matrix X), what will be the squared error?

(a) 0
(b) 1
(c) $\|y\|^2$
(d) Insufficient information to answer

[Ans: c]

2024 Apr 28 ET

6. Let X be a data matrix of the shape (d, n) and y be the associated label vector of shape $(n, 1)$. Assume that a linear regression model with loss as the sum of squared error is trained on the data $\{X, y\}$. In which of the following cases, the loss on the training data will necessarily be zero? Assume that the solution of the model is obtained by the normal equation that is $w^* = (XX^T)^{-1}Xy$. (MSQ)

(a) If y lies in the space spanned of row vectors of X.

(b) If y lies in the space spanned of row vectors of X^T.

(c) If all the data points satisfy the equality $x_1 + x_2 + \cdots + x_d = 0$, where x_i is the i^{th} feature and $y = 0$ for all the data points.

(d) If all the data points satisfy the equality $x_1^3 + x_2^3 + \cdots + x_d^3 = 0$, where x_i is the i^{th} feature and $y = 0$ for all the data points.

[Ans: a]

2024 Aug 04 Quiz 2

7. Consider the following training dataset for a linear regression problem with one feature:

x_i	y_i
1	1.2
-1	0.1
2	2.1
-2	-1

Find the value of w^*, the optimal weight.

[Ans: 0.73]

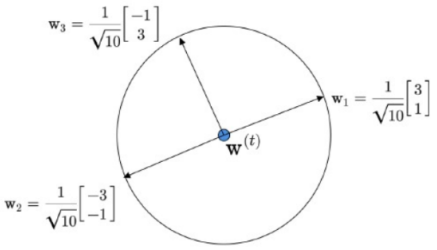
8. Consider the following dataset for a kernel regression problem with a polynomial kernel of degree two along with the coefficient vector $\boldsymbol{\alpha}$:

$$X = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 2 \\ -2 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ -0.625 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}$$

If the prediction for the data-point $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is 0, find the value of α_1.

[Ans: 0.625]

9. **Comprehension:**
Consider a linear regression problem that has two features with a training dataset of 10,000 points. The blue point in the image below is the weight vector, $\mathbf{w}^{(t)}$, at some iteration of gradient descent. $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$ are unit vectors. Specifically, $\mathbf{w}_1$ is a unit vector in the direction of the gradient at this point.



Based on the above data, answer the given subquestions.

i. If the next iteration of gradient descent is expressed as $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\mathbf{w}_k$ with $\eta > 0$, what would be the value of k?

[Ans: 2]

ii. If $\mathbf{w}^{(t)} = \begin{bmatrix}0.4\\0.1\end{bmatrix}$ and $\eta = \frac{1}{\sqrt{10}}$, find $\|\mathbf{w}^{(t+1)}\|$, where we use the usual L_2 norm.

[Ans: 0.1]

2024 Sep 01 ET

10. Consider a linear regression problem. Which of the following is the gradient of the SSE function with respect to $\mathbf{w} \in \mathbb{R}^d$, the weight vector, for a single data-point $\mathbf{x} \in \mathbb{R}^d$? y is the true label and $\hat{y}$ is the predicted label. Note that SSE is the sum of squared errors.

(a) $(\hat{y} - y)\mathbf{x}$

(b) $(\mathbf{w}^T\mathbf{x})\mathbf{w}$

(c) $\hat{y}\mathbf{x}$

(d) $\mathbf{x}y$

[Ans: a]

2024 Dec 01 Quiz 2

11. **Comprehension:**
Kernel regression with a polynomial kernel is applied on the following dataset with two features:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad y = [2, 1, 2]^T$$

Weight vector can be written as $w = \phi(X)\alpha$, where ϕ is the transformation mapping corresponding to the kernel $k(x_i, x_j) = (1 + x_i^T x_j)^2$. The vector α is given by $(K)^{-1}y$, where K is the kernel matrix.
Based on the above data, answer the given subquestions.

i. Compute K.

(a) $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 1 \\ 4 & 1 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \\ 4 & 1 & 1 \end{pmatrix}$

(c) $\begin{pmatrix} 4 & 4 & 1 \\ 1 & 1 & 1 \\ 4 & 1 & 1 \end{pmatrix}$ (d) $\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

[Ans: d]

ii. What will be the prediction for the data point $[1, -1]^T$? Enter the answer correct to two decimal places.

[Ans: 2.31 to 2.35]

12. Consider a linear regression model with loss

$$L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

where x_i's are the d-dimensional training data points and y_i's are their corresponding labels. For the optimal weight vector w^*, which among the following is correct? (MSQ)

(a) If we double all the values of y_i, w^* will also get doubled.

(b) If we double all the values of y_i, w^* will get halved.

(c) If we double all the values of x_i, w^* will get halved.

(d) If we double all the values of x_i, w^* will also get doubled.

[Ans: a, c]

13. Consider the following dataset:

$$D = \{(1, 2), (2, 2), (0, 1)\}$$

Assume that Leave one out cross validation is applied on this dataset and the model used is $y = w_0 + w_1 x$. Suppose $w = [w_0, w_1]^T$ be the weight obtained when $(2, 2)$ is used in the validation set. Then calculate $\|w\|_2^2$.

[Ans: 2]

2024 Dec 22 ET

14. Given a training set $A_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, let A_n^{-i} denote the set obtained by removing the label of the i-th training example. Leave one out cross validation (LOOCV) is performed as follows: for each (x_i, y_i) in the training set, we train the model on the remaining $n-1$ points of A_n^{-i} and test the prediction on (x_i, y_i). The error is defined as

$$error_L(A_n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}^{-i}(x_i))^2,$$

where $\hat{f}^{-i}(x_i)$ denote the predicted value of y_i using the model trained on A_n^{-i}. Now, for the dataset:

$$D = \{(1,2), (2,2), (0,1)\},$$

fit the model $\hat{f} = \beta x_i$ and find $error_L(A_3)$.

[Ans: 2]

15. Perform one round of the gradient descent algorithm for the linear regression on the below data

$$X = \begin{bmatrix} 1 & -2 & 0 \\ 2 & 5 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix}$$

with the initial weight vector $w_0 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$ and learning rate $\eta = 0.05$. Find the updated weight vector w_1 after one iteration of gradient descent.
Note: The objective function is given by $f(w) = \|X^T w - y\|^2$.

(a) $w_1 = [0.2, 3.1]^T$

(b) $w_1 = [0.2, 0.5]^T$

(c) $w_1 = [1, 0.5]^T$

(d) $w_1 = [-0.2, 3.1]^T$

[Ans: d]

2025 Mar 16 Quiz 2

16. Polynomial kernel in kernel regression model is expressed as:

$$K(x, x') = (x^T x' + 1)^d$$

What role does $K(x, x')$ play in kernel regression? (MSQ)

(a) It ensures that the regression model remains linear.

(b) It computes the similarity between data points in a transformed feature space.

(c) It allows the model to capture non-linear relationships by mapping data to a higher-dimensional space.

(d) It helps in reducing the variance of the model.

(e) It minimizes the loss function in regression problems.

[Ans: b, c]

17. Consider a kernel regression problem on the following dataset with two features:

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$$

where:

$$x_1 = [1 \quad 2]^T,\ y_1 = 3, \quad x_2 = [2 \quad 3]^T,\ y_2 = 5, \quad x_3 = [3 \quad 4]^T,\ y_3 = 7$$

The kernel function is given as:

$$K(x, x') = (1 + x^T x')^2$$

and the optimal coefficients α_i are given as:

$$\alpha = [0.5 \quad 1.0 \quad 1.5]^T.$$

Then, what will be the predicted value for the data point $[2 \quad 2]^T$?

[Ans: 483]

18. **Comprehension:**
Consider a linear regression model $y_i = w^T x_i$ with an initial weight $w = [3]$.

x_1	y
1	4
2	7
3	10
4	13

Based on the above data, answer the given subquestions.

i. Using the squared loss function and a learning rate of $\eta = 0.2$, compute the updated weight after one iteration of gradient descent for the given dataset:
Use $\nabla f(w) = 2(XX^T)w - 2Xy$

[Ans: 7]

ii. If we stop the algorithm at the weight calculated in the previous question, what will be the predicted value for the data point, $x_1 = 5$?

[Ans: 35]

2025 Apr 13 ET

19. A kernel regression model with a polynomial kernel of degree 2 is trained on the following dataset with three features:

$$X = \begin{bmatrix} 1 & 4 & 7 & 2 & 6 \\ 2 & 5 & 8 & 1 & 5 \\ 3 & 6 & 9 & 3 & 4 \end{bmatrix}.$$

Let the optimal weight vector w^* be given by:

$$w^* = \phi(X)[1, 1.2, 0, 0.7, -0.5]^T,$$

where ϕ is the transformation mapping corresponding to the kernel $K(x, x') = (x^T x' + 1)^2$. Compute the predicted value y_{pred} for a new data point $x' = [1 \quad 1 \quad 1]^T$. Enter the answer correct to one decimal place.

[Ans: 262.2 to 262.8]

2025 Aug 03 Quiz 2

20. Consider n data points, with each feature $x_i \in \mathbb{R}^+$ and label $y_i \in \mathbb{R}$. Suppose $y_i \mid x_i \sim \mathcal{N}(\log(wx_i), 1)$, then find the ML estimate $\hat{w}_{MLE}$ of w, where $w > 0$.

(a) $\hat{w}_{MLE} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \log(x_i))\right)$

(b) $\hat{w}_{MLE} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i \log(x_i))\right)$

(c) $\hat{w}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \log(x_i))$

(d) $\hat{w}_{MLE} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \log(x_i))\right)$

[Ans: a]

21. Consider a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Let $w^* \in \mathbb{R}^d$ be the solution of the optimization problem $\min_{w \in \mathbb{R}^d} \|X^T w - y\|_2^2$, where $X \in \mathbb{R}^{d \times n}$ and $y \in \mathbb{R}^n$. Construct a new dataset D_{new} by duplicating all points in D and adding 10 new points that lie exactly on the line $y = x^T w^*$. Then, the solution w^*_{new} of the corresponding optimization problem is also w^*.

(a) TRUE

(b) FALSE

[Ans: a]

22. Consider the linear regression model $y = w^T x + \epsilon$. Assuming $\epsilon \sim \mathcal{N}(0, \sigma^2)$, maximizing the conditional log-likelihood of w is equivalent to minimizing the sum of squared errors $\|y - X^T w\|_2^2$.

(a) TRUE

(b) FALSE

[Ans: a]

23. Let $w^* \in \mathbb{R}^d$ be the solution of the optimization problem $\min_{w \in \mathbb{R}^d} \|X^T w - y\|_2^2$, where $X \in \mathbb{R}^{d \times n}$ and $y \in \mathbb{R}^n$. Then, which of the following statements is/are true? (MSQ)

(a) $X^T w^*$ is the projection of the labels onto the subspace spanned by the features.

(b) $X^T w^*$ is the projection of the labels onto the subspace spanned by the data points.

(c) w^* lies in the span of data points.

(d) w^* lies in the span of features.

[Ans: a, c]

2026 Apr 06 Quiz 2

24. Consider a linear regression model applied to a dataset in which each data-point is described by two features. The first feature is always 1. The i^{th} data-point's feature-label pair is represented as $((1, x_i), y_i)$. Let $(w_0^*, w_1^*) \in \mathbb{R}^2$ be the optimal weight vector that minimizes the sum of squared errors. Consider two scenarios:
$S_1 : w_0^* + w_1^* x_i > y_i$, for each data-point $((1, x_i), y_i)$.
$S_2 : w_0^* + w_1^* x_i < y_i$, for each data-point $((1, x_i), y_i)$.
Which of the following is true?

(a) S_1 is possible, but S_2 is impossible.

(b) S_1 is impossible, but S_2 is possible.

(c) One of S_1 or S_2 is possible.

(d) Both S_1 and S_2 are impossible.

[Ans: d]

25. Consider a training dataset $(x_i, y_i)_{i=1}^n$ where all input points $x_i \in \mathbb{R}^d$ are distinct. Which of the following models is **guaranteed** to achieve a training error of zero?

(a) $h(x_i) = \frac{1}{n} \sum_{j=1}^n y_j \quad \forall i \in \{1, \ldots, n\}$

(b) $h(x_i) = 0 \quad \forall i \in \{1, \ldots, n\}$

(c) $h(x_i) = w^T x_i + b \quad \forall i \in \{1, \ldots, n\}$, given $n > d + 1$

(d) $h(x_i) = y_i \quad \forall i \in \{1, \ldots, n\}$

[Ans: d]

26. A researcher is modeling a dataset using linear regression of the form $y = X\theta + \epsilon$. They are considering two different probabilistic assumptions for the independent and identically distributed (i.i.d.) noise term ϵ:

- Assumption A: The noise follows a Gaussian distribution, $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- Assumption B: The noise follows a Laplace distribution, $\epsilon \sim \text{Laplace}(0, b)$.

The training dataset is known to contain several extreme outliers resulting from sensor malfunctions. Which of the following statements correctly describes the behavior of the Maximum Likelihood Estimation (MLE) for the parameter vector θ? (MSQ)

(a) Under Assumption A, the MLE is equivalent to minimizing Mean Squared Error (MSE), which is highly sensitive to outliers because the penalty grows quadratically.

(b) Under Assumption B, the MLE is equivalent to L1 regularization, which is more robust to outliers because the penalty for large residuals grows linearly.

(c) Both Assumption A and Assumption B lead to the same optimal parameter vector $\hat{\theta}$ regardless of the presence of outliers.

(d) All the given options are correct.

[Ans: a, b]

27. Let the linear regression model's solution be written as $w^* = X\alpha$, where $\alpha = [\alpha_1, \ldots, \alpha_n]^T \in \mathbb{R}^n$. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a valid kernel function. What will be the prediction for an unseen data point x_{test}?

(a) $\sum_{i=1}^n k(x_i, x_{\text{test}})$ (b) $\sum_{i=1}^n \alpha_i$

(c) $\sum_{i=1}^n \alpha_i k(x_i, x_{\text{test}})$ (d) $\sum_{i=1}^n \alpha_i^2 k(x_i, x_{\text{test}})$

[Ans: c]

2026 Aug 14 Quiz 2

28. Under the probabilistic view of linear regression, assume $y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$, where the observations are independent. Suppose the true noise variance is $4\sigma^2$ instead of σ^2, while the dataset and the observed labels remain unchanged. What happens to the maximum likelihood estimator $\hat{w}_{ML} = (XX^T)^{-1}Xy$?

(a) It is unchanged.

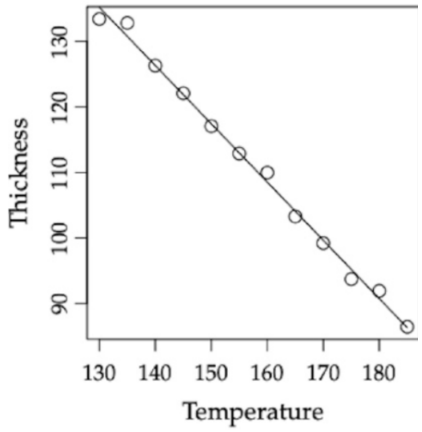
(b) It is scaled by 4.

(c) It is scaled by 1/4.

(d) It cannot be determined without knowing σ^2.

[Ans: a]

29. On a machine that folds plastic film, the operating temperature can be varied between 130°C and 185°C. To study the relationship between temperature and fold thickness, $n = 12$ observations of temperature and the corresponding fold thickness were collected. The scatter plot along with the fitted simple linear regression line is shown below:



Determine, by inspecting the figure, which of the following sets of parameter estimates for the simple linear regression model is most appropriate.
Note: The regression model is Thickness$(y) = \beta_0 + \beta_1(\text{Temperature}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Here, $\hat{\beta}_0$ is the estimated intercept, $\hat{\beta}_1$ is the estimated slope, and $\hat{\sigma}$ is the estimated standard deviation of the error term.

(a) $\hat{\beta}_0 = 0, \quad \hat{\beta}_1 = -0.9, \quad \hat{\sigma} = 360$

(b) $\hat{\beta}_0 = 0, \quad \hat{\beta}_1 = 0.9, \quad \hat{\sigma} = 3.6$

(c) $\hat{\beta}_0 = 252, \quad \hat{\beta}_1 = -0.9, \quad \hat{\sigma} = 3.6$

(d) $\hat{\beta}_0 = -252, \quad \hat{\beta}_1 = -0.9, \quad \hat{\sigma} = 360$

(e) $\hat{\beta}_0 = 252, \quad \hat{\beta}_1 = -0.9, \quad \hat{\sigma} = 360$

[Ans: c]

30. A student P minimizes the sum of squared errors (SSE) loss

$$L(w) = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2,$$

while another student Q minimizes the mean squared error (MSE) loss

$$L'(w) = \frac{1}{2n}\sum_{i=1}^{n}(w^T x_i - y_i)^2.$$

Both students use gradient descent with the same learning rate η and the same initial parameter vector w_0. Which one of the following statements is correct?

(a) They converge to different minimizers, because the losses differ.

(b) They converge to the same minimizer, but Q's steps are n times smaller; Q would need learning rate $n\eta$ to reproduce P's trajectory.

(c) They converge to the same minimizer, and their iterates coincide at every step.

(d) They converge to the same minimizer, but Q's steps are n times larger.

[Ans: b]

31. A dataset has $n = 3$ data points and $d = 2$ features. The two feature vectors are

$$f_1 = [1 \quad 0 \quad 1]^T, \quad f_2 = [0 \quad 1 \quad 1]^T.$$

For which label vector y is the linear model $f(x) = w^T x$ *guaranteed* to achieve exactly zero training error?

(a) $y = [1 \quad 2 \quad 3]^T$

(b) $y = [1 \quad 2 \quad 4]^T$

(c) $y = [2 \quad 1 \quad 4]^T$

(d) $y = [0 \quad 0 \quad 1]^T$

[Ans: a]

32. Suppose the label vector $y \neq 0$ is orthogonal to every feature vector, i.e.

$$f_j^T y = 0 \quad \text{for all } j = 1, \ldots, d.$$

w^* is the optimal weight obtained after minimizing the objective function of linear regression. Which statement is correct?

(a) $X^T w^* = y$ and the training error is 0.

(b) $X^T w^* = 0$ and the squared training error equals $\|y\|_2^2$.

(c) $X^T w^* = 0$ and the squared training error equals 0.

(d) The normal equations have no solution.

[Ans: b]

2026 Sep 13 ET

33. Let w^* be the solution to the ordinary least-squares linear regression problem

$$\min_w \|y - Xw\|^2.$$

Which of the following is always satisfied at w^*? Note that here X is your feature matrix and y is the label vector.

(a) $(Xw^*)^\top(y - Xw^*) = 0$

(b) $(Xw^*)^\top(Xw^*) = 0$

(c) $y^\top(y - Xw^*) = 0$

(d) $y - Xw^* = 0$

[Ans: a]

34. Suppose we model data as $y_i \sim \mathcal{N}(\log(w/x_i), 1)$ with a single unknown parameter $w > 0$ and known variance 1 (each $x_i > 0$). You perform maximum-likelihood estimation of w. The MLE of w must satisfy which of the following equations? Note that x is a feature and y is a label.

(a) $\sum_i y_i = \sum_i \log(w/x_i)$

(b) $\sum_i x_i y_i = \sum_i x_i \log(w/x_i)$

(c) $\sum_i \frac{y_i}{x_i} = \sum_i \log(w/x_i)$

(d) $\sum_i y_i = \sum_i \frac{w}{x_i}$

[Ans: a]