# Week_6_mlt_questions

## 2024 Mar 24 Quiz 2

1. **Comprehension:**
Consider a regression problem where you are tasked with predicting the sale prices of houses based on their square footage. You decide to experiment with two different models:

$$\text{Model } P: \quad \hat{y}_i = w_0 + w_1 x$$
$$\text{Model } Q: \quad \hat{y}_i = w_0 + w_1 x + w_2 x^2$$

The training dataset consists of information on 200 houses, and you use the models to make predictions on a test dataset of 50 houses. The Mean Squared Error (MSE) is chosen as the evaluation metric. Based on the above data, answer the given subquestions.

i. Considering the specific context of predicting house prices based on square footage, which model is more likely to provide accurate predictions on the training dataset?

(a) Model P
(b) Model Q
(c) Both models are equally likely to provide accurate predictions
(d) It depends on the distribution of house prices in the dataset

[Ans: b]

ii. Identify the factors that could influence the model's performance on the training dataset in this housing price prediction scenario. Select all correct statements: (MSQ)

(a) Model P may struggle to capture non-linear relationships present in houseprice data.
(b) Model Q might be sensitive to outliers in the square footage variable.
(c) The choice between Model P and Model Q depends on the budget constraints of potential homebuyers.
(d) Model Q will always perform well on the test dataset.

[Ans: a, b]

2. Given a design matrix $X \in \mathbb{R}^{d \times n}$ and a target vector $Y \in \mathbb{R}^{n \times 1}$, where $d$ represents the number of features, $n$ represents the number of data points, and the data is defined as:

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
$$Y = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

Calculate the coefficients $\beta$ for Ridge regression with $\lambda = 1$.

(a) $\beta = [0.5, 0.5]$
(b) $\beta = [1, 0.5]$
(c) $\beta = [0.54, 0.88]$
(d) $\beta = [0.67, 0.33]$
(e) None of these

[Ans: c]

## 2024 Apr 28 ET

3. Consider the following two models fitted on a one-dimensional dataset:

$$\text{Model 1:} \quad \hat{y} = w_0 + w_1 x$$
$$\text{Model 2:} \quad \hat{y} = w_1 x^2 + w_2 x + w_3$$

If both models are trained on the same one-dimensional dataset and evaluated on the same test dataset, which model is more likely to have higher bias and lower variance?

(a) Model 1

(b) Model 2

(c) Both models are equally sensitive to outliers

(d) Insufficient data

[Ans: a]

## 2024 Aug 04 Quiz 2

4. **Comprehension:**
A dataset for ridge regression has 1000 data-points. $k$-fold cross validation is performed to estimate the value of $\lambda$ in ridge regression with $k = 4$. For a given value of $\lambda$:
Based on the above data, answer the given subquestions.

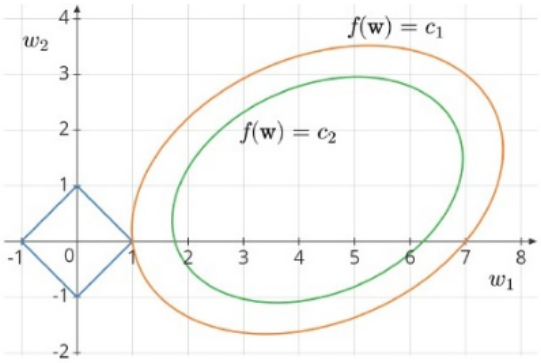
i. How many models have to be trained?

[Ans: 4]

ii. While training each of these models, how many data-points does the training set have?

[Ans: 750]

5. **Comprehension:**
Consider the following image of the parameter space in a regression problem with regularization. $f$ is the sum of squared errors (SSE) and $\lambda$ is the regularization parameter. The weight vector is $\mathbf{w} = [w_1 \quad w_2]^T$. The contours of $f$ and the regularization term in the loss are displayed below:




Based on the above data, answer the given subquestions.

i. What kind of a regression problem is this?

(a) Ridge regression
(b) LASSO regression

[Ans: b]

ii. What is the optimal value of $w_1$?

[Ans: 1]

iii. What is the optimal value of $w_2$?

[Ans: 0]

iv. Which of the following statements is true?

(a) Feature-1 can be discarded.
(b) Feature-2 can be discarded.
(c) Any one of the two features can be discarded.
(d) Both features are equally important and neither of them can be discarded.

[Ans: b]

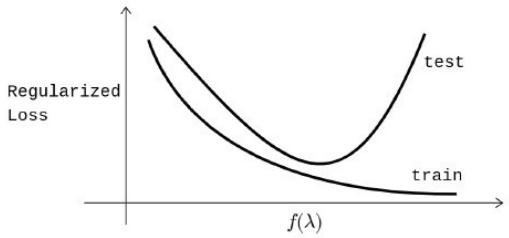
v. Which of the following statements are true? Exactly two options are correct. (MSQ)

(a) $c_1 > c_2$
(b) $c_1 < c_2$
(c) Total loss at optimal point is $c_1 + \lambda$
(d) Total loss at optimal point is $c_2 + \lambda$

[Ans: a, c]

## 2024 Sep 01 ET

6. Consider a regression problem that has a training and test dataset. Ridge regression is applied on the problem for various values of $\lambda$. The training and test loss are plotted against some function of $\lambda$, which we call $f(\lambda)$. Note that the continuous curves are obtained by connecting the points using a smooth curve. What is the most appropriate choice of $f(\lambda)$? Recall that the regularized loss is the sum of the SSE and the regularization term. Note that SSE is the sum of squared errors.




(a) $\frac{1}{\lambda}$

(b) $\lambda$

(c) $\lambda^2$

(d) $\log(\lambda)$

[Ans: a]

## 2024 Dec 01 Quiz 2

7. **Comprehension:**
Consider the following linear regression model

$$y_i \mid x_i = w^T x_i + \epsilon_i,$$

where the noise $\epsilon \sim \text{Normal}(0, 1)$.
Based on the above data, answer the given subquestions.

i. For some $\lambda > 0$, where $\lambda \in \mathbb{R}$, MSE of $\hat{w}_{ML}$ is greater than MSE of $\hat{w}_{Ridge}$.

(a) TRUE

(b) FALSE

[Ans: a]

ii. Consider the Bayesian formulation of the linear regression problem, where the prior for $w$ is assumed to be

$$w \sim \text{Laplace}(0, 2/\lambda).$$

Then, which among the following is true?
**Hint:** If $X \sim \text{Laplace}(\mu, b)$, then $f_X(x) = \frac{1}{2b} e^{-|x-\mu|/b}$.

(a) $\hat{w}_{MAP} = \arg\min_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$, where $\|w\|_2^2 = \sum_{i=1}^d w_i^2$

(b) $\hat{w}_{MAP} = \arg\max_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_2^2$, where $\|w\|_2^2 = \sum_{i=1}^d w_i^2$

(c) $\hat{w}_{MAP} = \arg\min_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_1$, where $\|w\|_1 = \sum_{i=1}^d |w_i|$

(d) $\hat{w}_{MAP} = \arg\max_w \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \|w\|_1$, where $\|w\|_1 = \sum_{i=1}^d |w_i|$

[Ans: c]

## 2024 Dec 22 ET

8. Consider a logistic regression model trained to detect spam emails. Emails containing harmful phishing links are labeled as spam (1), while regular emails are labeled as not spam (0).
A good spam detector should correctly identify almost all emails with phishing links. Even a single phishing email incorrectly classified as not spam could expose users to significant risks. However, the detector may classify some regular emails as spam. This trade-off is acceptable to ensure safety.
To ensure user safety, if the threshold is set to a low value to prioritize the detection of phishing emails, type 1. If the threshold is set to a high value to minimize false positives (misclassifying regular emails as spam), type 0.

[Ans: 1]

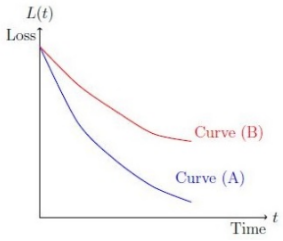
9. Is the following statement true or false? For a fixed size of the training and the test set, increasing the complexity of the model always leads to an increment in the test error.

(a) TRUE

(b) FALSE

[Ans: b]

10. Consider a supervised ML problem where a model is trained using a training dataset, and its performance is evaluated on both the training and validation datasets. During training, the model's loss is plotted over time for both the datasets. Assume that the loss $L(t)$ is decreasing as the model learns.
The following graph shows two curves, Curve (A) and Curve (B), representing the loss for the training and validation datasets, respectively:




Identify the curves corresponding to the training and validation datasets. Select all correct statements. (MSQ)

(a) Curve (A) represents the training loss.

(b) Curve (B) represents the training loss.

(c) Curve (A) represents the validation loss.

(d) Curve (B) represents the validation loss.

[Ans: a, d]

11. Which of the following options are correct? Select all that apply. (MSQ)

(a) Underfitted models generally have less bias and high variance.

(b) A decision tree with the maximum possible length may lead to overfitting.

(c) Weak learners are those whose performance is slightly better than the random classifier.

(d) All the estimators in bagging can be trained parallelly.

[Ans: b, c, d]

## 2025 Mar 16 Quiz 2

12. Below is the constrained version of the ridge regression optimization problem:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\text{subject to } \|w\|^2 \leq \theta.$$

Following are the weight vectors to be considered, along with the mean squared error (MSE) produced by each:

$$w_1 = [1 \quad 1 \quad 1 \quad 1]^T, \quad MSE = 2$$
$$w_2 = [0 \quad 2 \quad 1 \quad 3]^T, \quad MSE = 7$$
$$w_3 = [1 \quad 2 \quad 0 \quad 1]^T, \quad MSE = 1$$
$$w_4 = [2 \quad 1 \quad 1 \quad 2]^T, \quad MSE = 8$$

If $\theta = 10$, which of the following weight vectors will be selected as the final weight vector by ridge regression?

(a) $w_1$

(b) $w_2$

(c) $w_3$

(d) $w_4$

[Ans: c]

13. Let $\hat{w}_R$ and $\hat{w}$ be the ridge coefficient estimate and the least squares estimate, respectively, of the linear regression problem. Which of the following holds true? (MSQ)

(a) For $\lambda = 0$, $\frac{\|\hat{w}_R\|^2}{\|\hat{w}\|^2} = 1$

(b) For $\lambda = 0$, $\frac{\|\hat{w}_R\|^2}{\|\hat{w}\|^2} = 0$

(c) For $\lambda = \infty$, $\frac{\|\hat{w}_R\|^2}{\|\hat{w}\|^2} = 0$

(d) For $\lambda = \infty$, $\frac{\|\hat{w}_R\|^2}{\|\hat{w}\|^2} = 1$

[Ans: a, c]

## 2025 Apr 13 ET

14. What is the role of the regularization term in ridge regression?

(a) It ensures that the dataset is normally distributed.

(b) It penalizes large coefficients to reduce overfitting.

(c) It forces more coefficients to be exactly zero, performing feature selection.

(d) It increases the variance of the model.

[Ans: b]

15. Consider that the three weight vectors $w_1, w_2$, and $w_3$ are learned for a four-dimensional dataset using a linear regression model or regularized linear regression model (Not in any particular order).

$$w_1 = [-0.41 \quad 0.26 \quad 0.54 \quad 0.17]^T$$
$$w_2 = [-0.81 \quad 0.89 \quad 0.93 \quad 0.52]^T$$
$$w_3 = [0 \quad 0 \quad 0.08 \quad 0.2]^T$$

Select the most appropriate match for these weight vectors.

(a) $w_1 \rightarrow$ Linear regression, $w_2 \rightarrow$ Ridge regression, $w_3 \rightarrow$ Lasso regression

(b) $w_1 \rightarrow$ Ridge regression, $w_2 \rightarrow$ Linear regression, $w_3 \rightarrow$ Lasso regression

(c) $w_1 \rightarrow$ Lasso regression, $w_2 \rightarrow$ Ridge regression, $w_3 \rightarrow$ Linear regression

(d) $w_1 \rightarrow$ Lasso regression, $w_2 \rightarrow$ Linear regression, $w_3 \rightarrow$ Ridge regression

[Ans: b]

16. A set of data points is generated by the following process:

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \epsilon, \quad \text{where } \epsilon \text{ is a Gaussian noise.}$$

We use two models to fit the data:
Model 1: $\hat{y} = a_0 + a_1 x$
Model 2: $\hat{y} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5$.
For a fixed number of training examples, select the correct options from the following: (MSQ)

(a) Model 1 will have a higher bias than Model 2.

(b) Model 2 will have a lower variance than Model 1.

(c) Model 1 is more likely to underfit.

(d) Model 1 is more likely to overfit.

[Ans: a, c]

## 2025 Aug 03 Quiz 2

17. In the Bayesian modeling for linear regression, we assume $y \mid x \sim \text{Normal}(w^T x, \sigma^2 I)$ and $w \sim \text{Normal}(0, \gamma^2 I)$. Under this setup, the maximum aposteriori (MAP) estimator of the weights $\hat{w}_{\text{MAP}}$ is given by:

$$\hat{w}_{\text{MAP}} = \arg\min_{w} \left\{ \|y - X^T w\|^2 + \lambda \|w\|^2 \right\}.$$

Using the given information, find the expression for the regularization parameter $\lambda$.

(a) $\lambda = \sigma^2 \cdot \gamma^2$ (b) $\lambda = \frac{\gamma^2}{\sigma^2}$

(c) $\lambda = \frac{\sigma^2}{\gamma^2}$ (d) $\lambda = \frac{1}{\sigma^2 + \gamma^2}$

[Ans: c]

18. You are evaluating the performance of a linear regression model using different hyperparameter settings (learning rate and number of iterations). Below are the results:

| Setting | Learning Rate | Iterations | Training Error | Validation Error |
|---|---|---|---|---|
| 1 | 0.05 | 1000 | 150 | 160 |
| 2 | 0.10 | 800 | 145 | 150 |
| 3 | 0.15 | 500 | 140 | 190 |
| 4 | 0.25 | 350 | 170 | 180 |

Which hyperparameter setting would you choose based on the training and validation errors? Why?

(a) Setting 3, because it has the lowest training error, indicating the best fit to the training data.

(b) Setting 2, because it has the lowest validation error and a relatively low training error, indicating a good generalization.

(c) Setting 1, because it has the highest number of iterations and therefore likely to be more stable.

(d) Setting 4, because it uses the highest learning rate and converges the fastest.

[Ans: b]

19. Consider the linear regression model $y = w^T x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
**Note:** Here $\hat{w}_{\text{MLE}}$ is the Maximum Likelihood Estimator (MLE) of $w$, $\hat{w}_{\text{MAP}}$ is the Maximum A Posteriori (MAP) estimator of $w$, and $\mathbb{E}(\|\hat{w} - w\|^2)$ is the Mean Squared Error (MSE) between an estimator $\hat{w}$ and the true parameter $w$.
Which of the following statements is/are true?

(a) The estimator $\hat{w}_{\text{MLE}}$ always has a lower mean squared error (MSE) than $\hat{w}_{\text{MAP}}$ because it is unbiased.

(b) The estimator $\hat{w}_{\text{MAP}}$ always provides better estimates than $\hat{w}_{\text{MLE}}$ because it uses prior information.

(c) The estimator $\hat{w}_{\text{MAP}}$ can achieve lower MSE than $\hat{w}_{\text{MLE}}$ for some $\lambda > 0$, where $\lambda \in \mathbb{R}$.

(d) The MSE of $\hat{w}_{\text{MLE}}$ and $\hat{w}_{\text{MAP}}$ is always equal when the sample size is small.

[Ans: c]

## 2025 Aug 31 ET

20. **Comprehension:**
Consider a linear regression problem with design matrix $X \in \mathbb{R}^{d\times n}$ and target vector $Y \in \mathbb{R}^{n\times 1}$, where $n$ is the number of data points and $d$ is the number of features. Let $w \in \mathbb{R}^{d\times 1}$ denote the vector of regression coefficients:

$$X = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 \\ 3 \end{bmatrix}.$$

Based on the above data, answer the given subquestions.

i. Calculate $\hat{w}_{ML}$ and $\hat{w}_{Ridge}$ for the regularization parameter $\lambda = 10$.

(a) $\hat{w}_{ML} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad \hat{w}_{Ridge} = \begin{bmatrix} 0.20 \\ 1.49 \end{bmatrix}$

(b) $\hat{w}_{ML} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \quad \hat{w}_{Ridge} = \begin{bmatrix} 0.37 \\ 1.34 \end{bmatrix}$

(c) $\hat{w}_{ML} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}, \quad \hat{w}_{Ridge} = \begin{bmatrix} 0.09 \\ 0.31 \end{bmatrix}$

(d) $\hat{w}_{ML} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \quad \hat{w}_{Ridge} = \begin{bmatrix} 0.37 \\ 1.34 \end{bmatrix}$

[Ans: c]

ii. Repeat the ridge regression coefficients for $\lambda = 0.1$. How do the estimate for $w$ change as $\lambda$ decreases?

(a) The coefficients shrinks towards zero.

(b) The coefficients increases in magnitude.

(c) The coefficients remain unchanged.

[Ans: b]

## 2025 Dec 18 ET

21. In Bayesian linear regression, suppose the prior distribution over the weight vector is

$$w \sim \mathcal{N}(0, \alpha^{-1} I).$$

Which of the following optimization problems is equivalent to finding the Maximum A Posteriori (MAP) estimate of $w$?

(a) $\|y - Xw\|$

(b) $\|y - Xw\|^2 + \alpha\|w\|$

(c) $\|y - Xw\|^2 + \alpha\|w\|^2$

(d) $\|y - Xw\|^2 - \alpha\|w\|^2$

[Ans: c]

## 2026 Apr 06 Quiz 2

22. In Ridge Regression, the optimal weight vector is given by $\hat{w} = (XX^T + \lambda I)^{-1}Xy$. The regularization parameter $\lambda$ is determined using 5-fold cross-validation. The following table summarizes the average Training Error and Validation Error (Mean Squared Error) obtained for various values of $\lambda$:

| $\lambda$ Value | Avg. Training Error (MSE) | Avg. Validation Error (MSE) |
|---|---|---|
| 0.001 | 1.25 | 8.40 |
| 0.1 | 2.10 | 4.20 |
| 1.0 | 3.45 | 3.15 |
| 10.0 | 5.80 | 3.20 |
| 100.0 | 12.40 | 11.90 |

Based on the cross-validation results provided, what is the most appropriate choice for $\lambda$?

(a) 0.001

(b) 0.1

(c) 1.0

(d) 10.0

(e) 100

[Ans: c]

## 2026 Aug 14 Quiz 2

23. Let the training data be $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider the ridge regression estimator defined, for a fixed $\lambda > 0$, as

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^d} \left[ \frac{1}{2} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \frac{\lambda}{2} \|w\|^2 \right].$$

To obtain the maximum a posteriori (MAP) estimator, the following probabilistic model is assumed:

- Prior: $w \sim \mathcal{N}(0, \sigma^2 I)$ where $w \in \mathbb{R}^d$, $0 \in \mathbb{R}^d$ is the zero vector, $I \in \mathbb{R}^{d \times d}$ is the identity matrix, and $\sigma^2 > 0$ is the prior variance.
- Likelihood: Conditioned on $w$, the labels are mutually independent, with

$$y_i \mid x_i, w \sim \mathcal{N}(w^T x_i, 1), \quad i = 1, \ldots, n.$$

Let $\widehat{w}_{MAP}$ denote the maximum a posteriori estimate of $w$.
For which value of $\lambda$ do the two optimization problems become identical? (Equivalently, for which value of $\lambda$ does $\widehat{w}_{MAP} = \widehat{w}_R$ hold for every dataset?)

(a) $\lambda = \sigma^2$

(b) $\lambda = \frac{1}{2\sigma^2}$

(c) $\lambda = 1$

(d) $\lambda = \frac{1}{\sigma^2}$

[Ans: d]

## 2026 Sep 13 ET

24. Consider the effect of each operation below on a model's bias and variance. Which of the following statements are correct? (MSQ)

(a) Regularising the weights in a linear/logistic regression model increases bias and decreases variance.

(b) Increasing the maximum depth of a decision tree increases bias and decreases variance.

(c) Increasing the number of hidden units in a neural network decreases bias and increases variance.

(d) Removing all the non-support-vectors from the training set of an SVM leaves both bias and variance unchanged.

[Ans: a, c, d]

25. Consider one-dimensional ridge regression with no bias term. Given

$$\sum_i x_i^2 = 10 \quad \text{and} \quad \sum_i x_i y_i = 4,$$

with $\lambda = 10$, compute the optimal $w_{\text{ridge}}$. Note that $x_i$ is $i^{\text{th}}$ data point and $y_i$ is $i^{\text{th}}$ label. (Round to 2 decimal places.)

[Ans: 0.19 to 0.21]