FINAL QUIZ

Problem 1. (T/F) In cross-validation, having more folds does not guarantee a lower test error. (T)

Problem 2. (T/F) If f is a convex function, then the gradient descent algorithm converges to the global minimum. (F)

Problem 3. (T/F) In hierarchical clustering, when the number of clusters decrease, observations in the same cluster stays in the same cluster. (T)

Problem 4. (T/F) A model with smaller loss on training set tends to be a better model. (F)

Problem 5. (T/F) Increasing the learning rate causes the model to converge quicker, but it increases the likelihood of getting stuck at a local minimum than a lower learning rate during gradient descent. (F)

Problem 6. (T/F) Test data should never be used during the training process. (T)

Problem 7. (T/F) Transformers use context vectors to map words to their respective numerical representations in an effort to capture their semantic meaning. (F)

Problem 8. Which one of the following is an example of unsupervised learning? (b)

a) Linear regression
b) K-means clustering
c) Logistic regression
d) Decision tree classification

Problem 9. Which one of the following is the correct interpretation of $R^2 = 0.8$? (a)

a) 80% of the variance in the response variable is explained by the model.
b) 80% of the predicted values are correct.
c) The correlation between the predicted and actual values is 0.8.
d) The model has an 80% chance of making a correct prediction.

Problem 10. In the context of linear regression, what is the main advantage of applying regularization (Lasso or Ridge)? (c)

a) It increases the model's complexity to better fit the training data.
b) It reduces the size of the dataset required for training.
c) It helps prevent overfitting by penalizing large coefficients.
d) It guarantees zero training error.

Problem 11. Which of the following statements about Random Forest is correct? (d)

a) Random Forest uses a single decision tree to avoid overfitting.
b) All trees in a Random Forest are trained on the same subset of features.
c) Random Forest is more prone to overfitting than a single decision tree.
d) Random Forest reduces variance by averaging predictions from multiple trees.

Problem 12. Which one of the following is *not* a hyperparameter? (d)

a) Regularization strength in Ridge regression
b) Number of neighbors in K-means
c) Maximum depth in Decision Tree
d) Weights in neural network

Problem 13. Which of the following can be used as valid stopping conditions in gradient descent? (Select all that apply) (a,b,d)

a) The number of iterations exceeds a predefined limit
b) The magnitude of gradient becomes smaller than a chosen threshold
c) The sign of the gradient changes three times in a row
d) The change in model parameters is very small between consecutive updates

Problem 14. Compute the MSE and the MAE for given true values and predicted values from a regression model.

Sample	True y	Predicted y
1	3.0	3.5
2	-0.5	0
3	2.0	1.7
4	7.0	6.9
5	1.5	1.2

Answer:

$$\begin{aligned}\text{MSE} &= \frac{1}{5}\left((3.5-3)^2 + (-0.5-0)^2 + (1.7-2)^2 + (6.9-7)^2 + (1.2-1.5)^2\right)\\ &= \frac{0.69}{5}.\end{aligned}$$

$$\begin{aligned}\text{MAE} &= \frac{1}{5}(|3.5-3| + |-0.5-0| + |1.7-2| + |6.9-7| + |1.2-1.5|\\ &= \frac{1.7}{5}.\end{aligned}$$

Problem 15. Compute the accuracy, recall (TP/(TP+FN)), and precision (TP/(TP+FP)) for given true values and predicted values from a classification model.

	Predicted: Diseased	Predicted: Not Diseased
True: Diseased	10	5
True: Not Diseased	10	75

Answer:

$$\begin{aligned} \text{accuracy} &= \frac{10+75}{10+5+10+75} = 0.85, \\ \text{recall} &= \frac{10}{10+5} = 0.67, \\ \text{precision} &= \frac{10}{10+10} = 0.5. \end{aligned}$$

Problem 16. Consider the following multi-layer perceptron (MLP) architecture with an input dimension of 4 and an output dimension of 1. Each hidden layer uses the ReLU activation function, and all linear layers include bias terms.

- input dim = 4
- hidden dims = [8, 16, 8]
- output dim = 1

The MLP consists of four linear layers in the following order: Linear → ReLU → Linear → ReLU → Linear → ReLU → Linear. What is the total number of parameters in this network?

Answer: 40+144+136+9=329

Problem 17. In previous problem, what is the shape of the weight matrix in the second Linear layer (8 → 16)?

Answer: 16×8

Problem 18. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x, y, z) = y^3 - z^2$. Compute ∇f and identify the set $C = \{(x, y, z) : \nabla f(x, y, z) = (0, 0, 0)^\top\}$.

Answer: $\nabla f(x, y, z) = \begin{pmatrix} 0 \\ 3y^2 \\ -2z \end{pmatrix}$. $C = \{(x, 0, 0) : x \in \mathbb{R}\}$.

Problem 19. The following are the four main steps main steps in training a neural network. Arrange them in the correct order:

(1) Backpropagation
(2) Parameter update
(3) Loss computation
(4) Forward propagation

Answer: (4)-(3)-(1)-(2)

Problem 20. Let

$$x = 2,\ w_1 = 1,\ b_1 = -1,\ w_2 = 3,\ b_2 = 0,\ y = 5.$$

Consider following equations:

$$\begin{aligned} z &= w_1 x + b_1 \\ a &= \text{ReLU}(z) \\ \hat{y} &= w_2 a + b_2 \\ L &= \frac{1}{2}(\hat{y} - y)^2 \end{aligned}$$

Compute $\frac{\partial L}{\partial b_1}$.

Answer: -6

Problem 21. What is the Double Descent phenomenon?

Answer: A phenomenon where test error (loss) first decreases, then increases, and then decreases again as model complexity increases.

Problem 22. What problem does Gaussian noise pose in the training process?

Answer: the addition of Gaussian noise helps decrease the risk of overfitting by forcing it to learn more generalizable features of the data

Problem 23. What is the purpose of pooling in the feature learning process?

Answer: Pooling helps extract the dominant features of an image to reduce computational complexity.