Quiz-01 — Solutions

Weeks 1–3

Part A: Short Questions (brief answers)

1. **(Week 1) What is the main difference between supervised and unsupervised learning?**
   Answer: Supervised learning uses labeled data (features with target labels) to learn a mapping from inputs to outputs. Unsupervised learning uses unlabeled data and finds structure or patterns (e.g., clustering, dimensionality reduction).
2. **(Week 2) Why do we use the sigmoid function in logistic regression?**
   Answer: The sigmoid maps any real-valued score to the interval (0,1), allowing interpretation as a probability of class membership and enabling thresholding for classification.
3. **(Week 3) What is the key difference between Ridge and Lasso regression?**
   Answer: Ridge adds an ℓ_2 penalty (squared weights) which shrinks coefficients continuously; Lasso adds an ℓ_1 penalty (absolute values) which can drive some coefficients exactly to zero, performing feature selection.

Part B: Fill in the Blanks

1. In Gradient Descent, the learning rate controls the **step size** of the parameter updates.
2. In Lasso Regression, the regularization term is based on the **L1** norm of the weight vector.
3. An underfitting model usually has **high** training error and **high** validation error.

Part C: Multiple Choice Questions

1. Which of the following best describes Ridge Regression?
   (a) Uses L1 norm penalty to shrink coefficients.
   (b) Uses L2 norm penalty to shrink coefficients.
   (c) Combines both L1 and L2 penalties.
   (d) Does not use any regularization.
   Answer: (b)
2. Suppose a model has very low training error but very high validation error. This indicates:
   (a) Underfitting (b) Overfitting (c) Just right complexity (d) Data is perfectly clean
   Answer: (b)
3. In Elastic Net, setting the mix ratio $r = 1$ makes it equivalent to:
   (a) Ridge Regression (b) Lasso Regression (c) Linear Regression without regularization (d) Logistic Regression
   Answer: (b)

Part D: Computational Problem — Full Worked Solution

Problem statement

Consider the linear model with three parameters:

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Training examples (m = 5):

i	x_1	x_2	y
1	1	2	10
2	2	1	9
3	0	1	4
4	3	2	15
5	2	3	16

Initial parameters: $\theta_0 = \theta_1 = \theta_2 = 0$. Learning rate: $\alpha = 0.01$.

We use the MSE-based gradients:

$$\frac{\partial J}{\partial \theta_j} \;=\; \frac{1}{m} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)} \right) x_j^{(i)},$$

with $x_0^{(i)} = 1$.

Step 1: Predictions with initial parameters

All initial thetas are 0, so for every i:

$$h(x^{(i)}) = 0.$$

Residuals are $h - y = -y$.

Step 2: Compute required sums

$$\sum_{i=1}^{5} y^{(i)} = 10 + 9 + 4 + 15 + 16 = 54,$$

$$\begin{aligned} \sum_{i=1}^{5} x_1^{(i)} y^{(i)} &= 1 \cdot 10 + 2 \cdot 9 + 0 \cdot 4 + 3 \cdot 15 + 2 \cdot 16 \\ &= 10 + 18 + 0 + 45 + 32 = 105, \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{5} x_2^{(i)} y^{(i)} &= 2 \cdot 10 + 1 \cdot 9 + 1 \cdot 4 + 2 \cdot 15 + 3 \cdot 16 \\ &= 20 + 9 + 4 + 30 + 48 = 111. \end{aligned}$$

Step 3: Gradients (m = 5)

Using $h - y = -y$:

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{5}\sum_{i=1}^{5}(-y^{(i)}) = -\frac{54}{5} = -10.8.$$

$$\frac{\partial J}{\partial \theta_1} = \frac{1}{5}\sum_{i=1}^{5}(-y^{(i)})x_1^{(i)} = -\frac{105}{5} = -21.0.$$

$$\frac{\partial J}{\partial \theta_2} = \frac{1}{5}\sum_{i=1}^{5}(-y^{(i)})x_2^{(i)} = -\frac{111}{5} = -22.2.$$

Gradients: $\nabla J = (-10.8,\ -21.0,\ -22.2).$

Step 4: One gradient-descent update

Update rule:

$$\theta_j \leftarrow \theta_j - \alpha\frac{\partial J}{\partial \theta_j}.$$

With $\alpha = 0.01$ and initial thetas 0:

$$\theta_0^{\text{new}} = 0 - 0.01 \times (-10.8) = 0.108,$$
$$\theta_1^{\text{new}} = 0 - 0.01 \times (-21.0) = 0.210,$$
$$\theta_2^{\text{new}} = 0 - 0.01 \times (-22.2) = 0.222.$$

$$\boxed{\theta^{\text{new}} = (0.108,\ 0.210,\ 0.222)}$$

Instructor note: Minor differences in rounding are acceptable. Accept answers matching gradients and parameter updates within reasonable tolerance (e.g., ± 0.01).