Machine Learning

Supervised Learning Fundamentals Quiz – Answer Key

1. **(C) High bias and high training error.** High bias and training error indicate underfitting, not overfitting. Overfitting shows low training error but high test error (high variance).

2. **(B) The size of steps taken toward the minimum.** Learning rate α scales the gradient: $\theta = \theta - \alpha \nabla L$. Too large causes divergence; too small causes slow convergence.

3. **(B) L2 regularization (Ridge).** L2 adds $\lambda \sum w_i^2$ to loss. L1 (Lasso) adds $\lambda \sum |w_i|$, promoting sparsity through zero weights.

4. **(B) True positives divided by all predicted positives.** Precision = TP/(TP+FP). Recall (option A) = TP/(TP+FN). Accuracy (option D) = (TP+TN)/Total.

5. **(C) Provide more reliable performance estimates.** K-fold trains on K-1 folds and validates on 1, rotating through all folds. Averages reduce variance in performance estimates.

6. **True.** An unlimited-depth tree can create a leaf for every training example, achieving 100% training accuracy but memorizing noise and failing to generalize.

7. **False.** Softmax is used for multi-class classification (outputs probability distribution over K classes). Binary classification typically uses sigmoid (outputs single probability).

8. **False.** Feature scaling is essential for distance-based algorithms (KNN, SVM, neural networks) and gradient descent. Tree-based algorithms (Random Forest, XGBoost) are scale-invariant.

9. **Bias-Variance Tradeoff:**

   Definitions:

   - *Bias*: Error from oversimplified assumptions; underfitting; model cannot capture true patterns
   - *Variance*: Error from sensitivity to training data fluctuations; overfitting; model captures noise
   - Total Error = Bias^2 + Variance + Irreducible Error

   Model complexity effects:

   - Simple models (linear regression): High bias, low variance
   - Complex models (deep neural networks): Low bias, high variance
   - Optimal complexity minimizes total error

   Training data size effects:

   - More data reduces variance (model sees more patterns)
   - Bias remains unchanged by data quantity
   - Large datasets allow more complex models

Regularization effects:

- Adds penalty for model complexity
- Increases bias but decreases variance
- L1/L2 regularization, dropout, early stopping
- Example: Ridge regression prevents coefficient explosion

10. Logistic Regression Overview:

Model formulation:

$$z = w^T x + b$$
$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid function:

- Maps any real number to (0, 1)
- Output interpreted as P(y=1—x)
- Decision boundary at $\hat{y} = 0.5$ (when $z = 0$)

Loss function (Binary Cross-Entropy):

$$L = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- Penalizes confident wrong predictions heavily
- Convex function—guarantees global minimum

Gradient descent optimization:

$$\frac{\partial L}{\partial w} = \frac{1}{m} X^T (\hat{y} - y)$$
$$w := w - \alpha \frac{\partial L}{\partial w}$$
$$b := b - \alpha \frac{\partial L}{\partial b}$$

- Iteratively updates parameters to minimize loss
- Learning rate α controls step size
- Converges to optimal parameters