Answer Key & Rubric

Multiple Choice Answers:

1. D (A 1% error rate is catastrophic in high-frequency real-time systems like autonomous driving)
2. A (RL agents learn by maximizing cumulative rewards)
3. C (DS focuses on models/metrics; SWE focuses on latency/reliability/scale)
4. B (Tokenizers group characters into chunks, blinding the model to letter-level details)
5. C (Locality Sensitive Hashing is the standard for fuzzy deduplication at scale)
6. B (Target encoding must be calculated only on the training folds to avoid future data leakage)
7. A (Standardization handles outliers much better than Min-Max scaling)
8. D (Generative learns P(x,y); Discriminative learns P(y|x) or boundaries)
9. B (It naively assumes all features are conditionally independent given the class)
10. A (Log probabilities prevent float underflow and turn multiplication into addition)
11. B (Ad pricing requires accurate probability; Log-Loss in LR calibrates the DNN scores into real probabilities)
12. B (Data contamination: models memorizing public test sets)
13. A (A model predicting "all healthy" gets 99% accuracy but is completely useless)
14. D (Lowering the threshold catches more true positives [higher recall] but brings in more false alarms [lower precision])
15. A (L1 penalty creates diamond constraint corners, forcing weights to zero)
16. B (A single split is highly subject to randomness; k-fold smooths this out)
17. A (High variance = overfitting to noise)
18. B (L2 forms a circular constraint, providing smooth shrinkage)
19. B (Elastic Net uniquely combines L1 and L2 penalties)
20. A (SMOTE synthesizes new examples for the minority class)

Short Answer Questions

Question 1: What is the "impossible triangle" in AdTech system design? Explain, using the lecture examples, why the industry uses a "Two-Stage Cascade" (Retrieval and Ranking) hybrid architecture to satisfy these constraints.

Grading Rubric (Total: 10 points)

- (3 pts) Correctly identify and explain the three competing constraints of the "impossible triangle."
- (4 pts) Describe the two-stage architecture clearly, including what each stage prioritizes and what type of model each stage might use.
- (3 pts) Explain *why* this hybrid design is necessary — i.e., why a single-stage approach cannot satisfy all three constraints simultaneously.

Hints: Think about what happens if you try to run a heavy, accurate model on millions of candidates in under 10ms. What tradeoffs does each stage make?

Question 2: Compare Normalization (Min-Max) and Standardization (Z-score) in data processing. Give an example of a machine learning algorithm where feature scaling is absolutely necessary, and explain why.

Grading Rubric (Total: 10 points)

- (3 pts) Clearly describe how each scaling method works and what output range/distribution it produces.
- (3 pts) Discuss at least one key difference between them (e.g., sensitivity to outliers, output bounds, assumptions).
- (2 pts) Provide a valid algorithm example where unscaled features would cause problems.
- (2 pts) Explain *why* that algorithm breaks or degrades without scaling — what goes wrong mechanically?

Hints: Consider algorithms that rely on computing distances or similarities between data points. What happens when one feature has a range of 0–1 and another has a range of 0–1,000,000?

Question 3: Briefly contrast the core philosophies of Generative and Discriminative models. Provide one classic algorithm example for each from the lectures.

Grading Rubric (Total: 10 points)

- (4 pts) Clearly articulate what each type of model tries to learn (what probability or function it models). Using mathematical notation (e.g., P(x,y) vs. P(y|x)) is encouraged but not required.
- (3 pts) Provide a valid and appropriate classic example for each type.
- (3 pts) Use an analogy, contrast, or explanation that demonstrates you understand the *why* behind the difference, not just the definition.

Hints: Think about it this way — one approach tries to understand how the data was created; the other doesn't care how data was created and just focuses on telling classes apart. There are multiple valid examples for each type.

Question 4: What is the Bias-Variance Tradeoff in machine learning? Why is it practically impossible to reduce both bias and variance to zero in real-world tasks?

Grading Rubric (Total: 10 points)

- (3 pts) Define Bias and Variance separately and correctly, ideally connecting each to underfitting or overfitting.
- (4 pts) Explain the tradeoff mechanism — why does reducing one tend to increase the other?
- (3 pts) Explain why achieving zero total error is impossible in practice. Your answer should reference the concept of irreducible error or noise inherent in real-world data.

Hints: Consider what happens as you make a model progressively more complex. At what point does "learning the signal" turn into "memorizing the noise"? And why can't you perfectly separate the two?