Section 1: Multiple Choice Questions (Select the best answer for each question)

1. In the context of the Waymo self-driving system mentioned in class, why is a 99% model accuracy considered a "disaster" in the industry?
   A) Because 99% means the model's processing Frames Per Second (FPS) is too low.
   B) Because competitors usually achieve a 99.99% academic benchmark accuracy.
   C) Because rule-based methods can theoretically achieve 100% accuracy.
   D) Because a 1% error rate at 30 FPS means a critical decision failure occurs roughly every 3.3 seconds.

2. What is the primary goal of an Agent in Reinforcement Learning (RL)?
   A) To maximize the expected cumulative reward through interaction with an environment.
   B) To discover hidden topological feature structures in unlabelled data.
   C) To predict the correct categorical labels on new test data.
   D) To transform continuous time-series data into discrete categorical features.

3. In a machine learning team, what is the typical difference in focus between a Data Scientist and a Software Engineer?
   A) Scientists focus on reducing model latency, while engineers focus on improving recall.
   B) Scientists manage hardware infrastructure, while engineers design features and models.
   C) Scientists focus on model prototypes and theoretical accuracy, while engineers focus on system reliability, latency, and deployment.
   D) Scientists only evaluate online dynamic data, while engineers evaluate offline static data.

4. According to the "Strawberry Paradox" mentioned in the lectures, what is the root cause of early LLMs failing to count the number of letters in a word correctly?
   A) A lack of high-quality Reinforcement Learning from Human Feedback (RLHF) data.
   B) Tokenization destroys the character-level structural information of words before the model even sees them.
   C) The Transformer architecture lacks sufficient depth (number of layers).
   D) The model's parameter count has not reached the threshold for emergent abilities.

5. To prevent "Model Collapse" caused by training on highly repetitive text, which technique is widely used in the industry for deduplicating petabyte-scale datasets?
   A) Target Encoding
   B) K-Means Clustering
   C) Locality Sensitive Hashing (LSH)
   D) Principal Component Analysis (PCA)

6. When using Target Encoding for categorical features, how must you perform the operation to strictly avoid Data Leakage?
   A) Calculate the mean across the entire combined dataset (Train, Validation, and Test sets).
   B) Calculate the statistics strictly within the current training folds during cross-validation.
   C) Calculate the categorical statistics using only the validation set.
   D) Assign random target statistics to normalize the distribution.

7. When scaling numerical features, what is a significant advantage of Standardization (Z-score) over Normalization (Min-Max scaling)?
   A) It is much less sensitive to extreme outliers in the dataset.
   B) It strictly bounds all values within the [0, 1] range.
   C) It automatically imputes missing values during the scaling process.
   D) It is the only valid scaling method for distance-based algorithms like KNN.

8. What is the core philosophical difference between Generative and Discriminative models?
   A) Generative models directly learn the decision boundary, while discriminative models learn the joint distribution.
   B) Generative models are exclusively for text data, while discriminative models are for image data.
   C) Generative models require no labeled data at all, while discriminative models require massive labeled datasets.
   D) Generative models learn the joint probability distribution (how data is generated), while discriminative models directly learn the decision boundary (how to separate classes).

9. Why is the Naive Bayes algorithm called "naive"?
   A) It assumes all features strictly follow a Gaussian normal distribution.
   B) It assumes that, given the class label, the features are conditionally independent of each other.
   C) It assumes all features are perfectly linearly correlated with the target variable.
   D) It assumes the dataset contains absolutely no noise or missing values.

10. When implementing Naive Bayes in code, probabilities are typically converted into log probabilities and added together instead of multiplied. What is the main reason for this?
    A) To prevent numerical underflow when multiplying many extremely small probabilities in a computer.
    B) To convert non-linear classification boundaries into linear ones.
    C) To automatically filter out highly infrequent outlier words.
    D) To accelerate the convergence speed during gradient descent optimization.

11. In AdTech precise ranking, why is a Logistic Regression layer often added after the Deep Neural Network (DNN) outputs?
    A) Because DNN inference latency always exceeds the strict 10ms limit.
    B) Because DNNs are often uncalibrated, and Logistic Regression optimizing Log-Loss ensures the output represents true click probabilities.
    C) Because deep networks can only output binary results (0 or 1) and cannot output decimal probabilities.
    D) Because Logistic Regression can automatically discover complex non-linear cross features better than DNNs.

12. What phenomenon does "The Death of Public Benchmarks" refer to in the evaluation of large language models?
    A) A universal drop in model accuracy, indicating benchmarks are severely outdated.
    B) Models accidentally training on test set questions found online (Data Contamination), leading to falsely high benchmark scores.

C) Benchmark datasets are too small to reflect true zero-shot generative capabilities.
D) The extreme cost of manual evaluation causing companies to abandon public tests entirely.

13. In a medical diagnosis dataset that is highly imbalanced (e.g., 99% healthy, 1% sick), which evaluation metric is the most misleading to rely on?
    A) Accuracy
    B) Recall
    C) Precision
    D) Area Under the PR Curve (AUCPR)

14. In a binary classification model, if you lower the probability threshold for predicting the Positive class, what typically happens?
    A) Both Precision and Recall increase significantly.
    B) Precision increases, while Recall decreases.
    C) The number of False Positives decreases dramatically.
    D) Recall increases, while Precision decreases.

15. Which regularization method tends to produce sparse solutions (driving some feature coefficients exactly to zero), effectively performing automatic feature selection?
    A) Lasso (L1 Regularization)
    B) Ridge (L2 Regularization)
    C) Dropout
    D) Early Stopping

16. Compared to k-Fold Cross-Validation, what is a major risk of using the simple Hold-Out Method (a single Train/Val/Test split)?
    A) It consumes multiple times more computational resources.
    B) The evaluation results are highly dependent on the random split, making it unstable, especially on small datasets.
    C) It cannot be used for hyper-parameter tuning in neural networks.
    D) It inevitably causes severe overfitting to the training dataset.

17. In the Bias-Variance Tradeoff, what does "High Variance" typically indicate about a model?
    A) The model is overly complex and highly sensitive to small noise in the training data (Overfitting).
    B) The model is too simple to capture the underlying patterns between features and labels (Underfitting).
    C) The model has excellent robustness and perfect generalization capabilities.
    D) The model consistently and systematically deviates from the true values in its predictions.

18. From a geometric perspective, what is the shape of the penalty constraint region for Ridge Regression (L2 Regularization)?
    A) A diamond, with sharp corners that encourage sparse solutions.
    B) A circle, which shrinks parameters smoothly but rarely drives them exactly to zero.

C) An open rectangular region.
D) A single point at the origin.

19. [Extension] When certain features in a dataset are highly correlated (multicollinearity), Lasso might randomly keep one and drop the others. To combine the advantages of both L1 and L2 regularization, which powerful regression method is commonly used in the industry?
    A) Robust Regression
    B) Elastic Net
    C) Isotonic Regression
    D) Polynomial Regression

20. [Extension] When dealing with highly imbalanced datasets like credit card fraud, besides changing evaluation metrics, we can address the issue at the data level. Which classic technique generates synthetic "minority class" samples to balance the training data?
    A) Synthetic Minority Over-sampling Technique (SMOTE)
    B) Term Frequency-Inverse Document Frequency (TF-IDF)
    C) Early Stopping
    D) Chi-Square Test

Section 2: Short Answer Questions

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