Machine Learning Questions and Answers

Question 1 (14 marks)

What is the need for machine learning in modern systems? How is it different from traditional programming?

Machine learning (ML) is crucial in modern systems because it enables computers to learn from data and adapt to new situations without requiring explicit programming for every task. It is particularly valuable when: - Rules or patterns are too complex to code manually. - Data is vast and constantly changing. - Human expertise cannot fully address the problem.

In **traditional programming**, developers write specific instructions for a computer to execute. For example, to sort numbers, a programmer might implement a sorting algorithm like quicksort. The logic is static and predefined.

In contrast, **machine learning** allows the system to learn patterns from data. For instance, instead of coding rules to recognize handwritten digits, an ML model is trained on a dataset of labeled digits and learns to identify them. This makes ML ideal for tasks like image recognition or predictive analytics.

Key Difference:
- Traditional programming: Logic is explicitly coded.
- Machine learning: Logic is inferred from data.

Question 2 (14 marks)

Explain Cross Validation. Give an example.

Cross-validation is a technique to assess how well a machine learning model generalizes to unseen data. It involves splitting the dataset into multiple subsets, training the model on some, and testing it on others.

A popular method is **k-fold cross-validation**: - Divide the dataset into *k* equal parts (folds). - Train the model on *k-1* folds and test it on the remaining fold. - Repeat *k* times, using a different fold as the test set each time. - Compute the average performance across all iterations.

Example:
In 5-fold cross-validation, a dataset is split into 5 parts. The model trains on 4 parts and tests on the 5th, repeating this 5 times. The average accuracy provides a robust performance estimate.

This method maximizes data usage and helps detect overfitting.

Question 3 (14 marks)

Explain the classification of Machine Learning algorithms. Give real-world examples of:
- **Supervised Learning**
- **Unsupervised Learning**
- **Reinforcement Learning**
Also explain the difference between these types in terms of data availability and learning behavior.

Machine learning algorithms are classified into three main types:

1. **Supervised Learning**:
   - Uses labeled data (input-output pairs) to train the model.
   - Goal: Predict outputs for new inputs.
   - **Examples**:
     - *Classification*: Predicting if an email is spam.
     - *Regression*: Predicting house prices based on size and location.
2. **Unsupervised Learning**:
   - Works with unlabeled data to find patterns or structures.
   - Goal: Discover hidden relationships.
   - **Examples**:
     - *Clustering*: Grouping customers by purchase behavior.
     - *Dimensionality Reduction*: Simplifying data while retaining key information.
3. **Reinforcement Learning**:
   - Learns by interacting with an environment, using rewards or penalties as feedback.
   - Goal: Maximize cumulative reward.
   - **Examples**:
     - *Game Playing*: Training an AI to play chess.
     - *Robotics*: Teaching a robot to navigate obstacles.

Differences:
- **Data Availability**:
- Supervised: Requires labeled data.

- Unsupervised: Uses unlabeled data.
- Reinforcement: Needs an interactive environment.
- **Learning Behavior**:
- Supervised: Learns from examples.
- Unsupervised: Finds patterns without guidance.
- Reinforcement: Learns through trial and error.

Question 4 (14 marks)

Discuss the Machine Learning Life Cycle in detail. Describe each phase of the ML project from data collection to model deployment. Explain how feedback loops and monitoring play a role in post-deployment.

The **Machine Learning Life Cycle** includes:

1. **Problem Definition**: Define the problem and objectives (e.g., predict sales).
2. **Data Collection**: Gather data from sources like databases or APIs.
3. **Data Preprocessing**: Clean data (e.g., handle missing values, normalize).
4. **Feature Engineering**: Select or create features to improve model performance.
5. **Model Selection**: Choose an algorithm (e.g., random forest for classification).
6. **Model Training**: Train the model on the training dataset.
7. **Model Evaluation**: Test performance using metrics like accuracy on a test set.
8. **Hyperparameter Tuning**: Optimize model settings (e.g., via grid search).
9. **Model Deployment**: Integrate the model into a production system.
10. **Monitoring and Maintenance**: Track performance and update as needed.

Post-Deployment:
- **Feedback Loops**: Collect user or system feedback to detect issues (e.g., inaccurate predictions).
- **Monitoring**: Continuously assess performance to address data drift or degradation, retraining the model when necessary.

Question 5 (14 marks)

What is the bias-variance tradeoff in machine learning? Use graphs and examples to explain:
- **High bias vs high variance**
- **Underfitting vs overfitting**
How can this trade-off be optimized during model training?

The **bias-variance tradeoff** balances a model's complexity and its ability to generalize: - **Bias**: Error from overly simple assumptions. High bias causes underfitting.
- **Variance**: Error from sensitivity to training data. High variance causes overfitting.

Examples:
- **High Bias (Underfitting)**: A linear model on nonlinear data performs poorly on both training and test sets.
- **High Variance (Overfitting)**: A complex decision tree fits training data perfectly but fails on test data.

Graph:
- Training error decreases with complexity.
- Test error decreases, then increases (overfitting).
- Optimal point: Minimum test error.

Optimization:
- Use cross-validation to assess generalization.
- Apply regularization (e.g., L2) to reduce variance.
- Adjust model complexity.
- Increase training data to lower variance.

Question 6 (14 marks)

Compare and contrast Machine Learning with Statistical Modeling. Mention the key philosophical, theoretical, and application-based differences. Support your answer with at least two examples from real-world problems.

Philosophical:
- **Statistical Modeling**: Emphasizes understanding relationships and inference.
- **Machine Learning**: Prioritizes prediction and generalization.

Theoretical:
- **Statistical Modeling**: Assumes data distribution and estimates parameters.
- **Machine Learning**: Learns patterns without strict assumptions.

Application-Based:
- **Statistical Modeling**: Used in economics, biology for hypothesis testing.
- **Machine Learning**: Applied in vision, NLP for predictive tasks.

Examples:
1. **Customer Churn**:
- *Statistical*: Logistic regression to interpret factors.
- *ML*: Random forest for accurate prediction.
2. **Drug Effect**:
- *Statistical*: Tests for significance in trials.
- *ML*: Predicts patient outcomes.

Question 7 (14 marks)

Explain the steps involved in Machine Learning Model Development. Include:
- **Problem definition**
- **Data collection and preprocessing**
- **Feature selection**
- **Model selection, training, and evaluation**
- **Hyperparameter tuning and validation**

1. **Problem Definition**: Specify the task (e.g., regression) and metrics.
2. **Data Collection and Preprocessing**: Gather and clean data (e.g., remove duplicates).
3. **Feature Selection**: Identify key features (e.g., using correlation).
4. **Model Selection**: Choose an algorithm (e.g., SVM).
5. **Model Training**: Fit the model to training data.
6. **Model Evaluation**: Measure performance (e.g., precision) on a test set.
7. **Hyperparameter Tuning**: Optimize settings (e.g., random search).
8. **Validation**: Confirm generalization on a holdout set.

Question 8 (14 marks)

Describe model evaluation techniques used in supervised learning. Explain the following with examples:
- **Accuracy**
- **Precision**
- **Recall**

- **F1-Score**
- **Confusion Matrix**
Discuss when to use each of these metrics appropriately.

- **Accuracy**: Correct predictions / total predictions.
  - *Example*: 85% correct in a balanced dataset.
  - *Use*: Balanced datasets.
- **Precision**: True positives / predicted positives.
  - *Example*: Spam detection accuracy.
  - *Use*: High cost of false positives.
- **Recall**: True positives / actual positives.
  - *Example*: Disease detection coverage.
  - *Use*: High cost of false negatives.
- **F1-Score**: Harmonic mean of precision and recall.
  - *Example*: Search engine performance.
  - *Use*: Balance precision and recall.
- **Confusion Matrix**: Table of true/false positives/negatives.
  - *Example*: Binary classification breakdown.
  - *Use*: Detailed analysis.

Question 9 (14 marks)

What are Parametric and Non-Parametric Models? Define both with mathematical examples. Explain their differences with respect to:
- **Flexibility**
- **Assumptions**
- **Interpretability**
- **Computation**
Mention suitable use-cases for each.

Parametric Models:
- Assume a fixed form and estimate parameters.
- *Example*: Linear regression: (y = _0 + _1 x).

Non-Parametric Models:
- Adapt to data without a fixed form.
- *Example*: KNN predicts based on nearest neighbors.

Differences:
- **Flexibility**: Non-parametric models are more flexible.
- **Assumptions**: Parametric models assume a distribution.
- **Interpretability**: Parametric models are easier to interpret.
- **Computation**: Parametric models are faster.

Use-Cases:
- **Parametric**: Linear data (e.g., economics).
- **Non-Parametric**: Complex data (e.g., image recognition).
