

Machine Learning Exam Questions - Chapter 1

1. Define Machine Learning (ML).

Definition: Machine Learning (ML) is a subset of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. The goal is to develop algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Key Features:

1. **Automatic Learning:** Systems learn patterns from data without explicit programming for each task
2. **Experience-based Improvement:** Performance improves as more data becomes available
3. **Statistical Analysis:** Uses mathematical and statistical methods to make predictions
4. **Adaptability:** Models can update and adapt to new information dynamically

Practical Example: Email spam detection systems learn from thousands of labeled emails (spam/not spam) to automatically classify new incoming emails without being programmed with specific rules for each type of spam.

Significance: ML is fundamental to modern AI applications, enabling computers to solve complex problems that would be impossible to program manually, making it essential for automation and intelligent decision-making across industries.

2. List applications of Machine Learning.

Introduction: Machine Learning has widespread applications across various industries, transforming how businesses operate and solve complex problems through data-driven insights and automated decision-making.

Key Applications:

1. **Image and Speech Recognition:** Computer vision, voice assistants, facial recognition systems
2. **Healthcare:** Medical diagnosis, drug discovery, personalized treatment plans

3. **Finance:** Fraud detection, algorithmic trading, credit scoring, risk assessment
4. **Transportation:** Autonomous vehicles, route optimization, traffic management

Practical Example: Netflix uses ML algorithms to analyze viewing patterns, preferences, and user behavior to provide personalized movie and TV show recommendations, significantly improving user engagement and satisfaction.

Significance: These applications demonstrate ML's versatility in solving real-world problems, improving efficiency, reducing costs, and enabling innovations that were previously impossible with traditional programming approaches.

3. Explain Supervised Machine Learning with example.

Definition: Supervised learning is a type of machine learning where the model is trained using a labeled dataset. Each training example is a pair consisting of an input and a desired output label. The model learns to map inputs to outputs by minimizing the prediction error.

Key Characteristics:

1. **Labeled Data:** Requires a dataset where each example is correctly labeled with the desired output
2. **Goal-oriented:** Aims to predict the label of unseen data accurately
3. **Error Minimization:** Uses error functions to measure how far predictions are from actual labels
4. **Generalization:** Learns patterns that can be applied to new, unseen data

Practical Example: Email spam classification where the system is trained on thousands of emails labeled as "spam" or "not spam." Input: "Congratulations! You've won a prize" → Label: Spam. Input: "Meeting at 10 AM tomorrow" → Label: Not Spam. The algorithm learns from these examples to classify new emails.

Significance: Supervised learning is the most common ML approach, providing reliable and interpretable results for classification and regression tasks, making it essential for applications requiring high accuracy and predictable outcomes.

4. Explain PAC Learning with example.

Definition: Probably Approximately Correct (PAC) learning is a theoretical framework that addresses the problem of learning a function from samples in a way that is both probably

correct and approximately correct. It formalizes conditions under which a learning algorithm can perform well on new, unseen data.

Key Concepts:

1. **Sample Complexity:** Determines the number of samples required to achieve specified accuracy and confidence
2. **Hypothesis Space:** The set of all possible hypotheses the algorithm can choose from
3. **Generalization Bounds:** Provides theoretical guarantees on performance with unseen data
4. **VC Dimension:** Measures the complexity of the hypothesis space affecting sample requirements

Practical Example: An image classification system trained with sufficient data can classify images with $\leq 5\%$ error at 95% confidence. With enough training samples, a PAC learner guarantees that the learned hypothesis will perform within these bounds on new images.

Significance: PAC learning provides theoretical foundations for understanding learning algorithms' behavior, helping determine sample size requirements and offering confidence guarantees crucial for critical applications like medical diagnosis or autonomous systems.

5. Explain Model Selection and Generalization with example.

Definition: Model selection is the process of choosing the best-performing machine learning model from a set of candidates. Generalization refers to a model's ability to perform well on new, unseen data after being trained on a specific dataset.

Key Techniques:

1. **Cross-validation:** K-fold validation to assess model performance across different data splits
2. **Bias-Variance Trade-off:** Balancing model complexity to avoid overfitting and underfitting
3. **Hyperparameter Tuning:** Grid search and Bayesian optimization for optimal parameters
4. **Performance Metrics:** Using accuracy, precision, recall, and F1-score for evaluation

Practical Example: For house price prediction, comparing linear regression, decision trees, and neural networks using cross-validation. Linear regression might generalize better with limited data, while neural networks might perform better with large datasets, requiring careful selection based on data size and complexity.

Significance: Proper model selection and generalization ensure reliable real-world performance, preventing overfitting that leads to poor practical results and enabling deployment of robust ML systems in production environments.

6. Dimensions of Supervised ML algorithms with example.

Definition: Dimensions of supervised ML algorithms refer to various aspects and characteristics that can be used to compare, evaluate, and understand different learning algorithms, helping in algorithm selection and analysis.

Key Dimensions:

1. **Hypothesis Space:** The set of functions the algorithm can learn (linear vs. non-linear)
2. **Model Complexity:** Simplicity vs. ability to capture patterns, measured by VC dimension
3. **Sample Complexity:** Amount of data needed for good generalization
4. **Computational Complexity:** Training time, prediction time, and scalability

Practical Example: Comparing k-NN and SVM: k-NN has infinite VC dimension (can memorize any dataset) but is computationally expensive for prediction, while SVM has bounded complexity with efficient prediction but requires more training time for large datasets.

Significance: Understanding these dimensions helps practitioners choose appropriate algorithms based on data characteristics, computational constraints, and performance requirements, leading to more effective ML system design and deployment.

7. Explain VC dimension with example.

Definition: The Vapnik-Chervonenkis (VC) dimension is a measure of the capacity or complexity of a set of functions (hypothesis space) that a learning algorithm can implement. It is the maximum number of points that can be shattered by a hypothesis class.

Key Concepts:

1. **Shattering:** A set of points is shattered if for every possible labeling, there exists a function that correctly classifies them
2. **Capacity Measure:** Higher VC dimension indicates greater model complexity and flexibility
3. **Generalization Bound:** Relates to sample complexity and overfitting risk
4. **Model Comparison:** Helps compare different algorithms' expressive power

Practical Example: Linear classifier in 2D has VC dimension of 3. It can shatter any 3 non-collinear points (can find a line for any of the $2^3=8$ possible labelings), but cannot shatter 4 points in general (no single line can separate all $2^4=16$ possible labelings of 4 points).

Significance: VC dimension provides theoretical foundation for understanding model complexity, helping determine appropriate model selection, sample size requirements, and generalization capabilities essential for reliable ML system design.

8. Why is VC Important?

Introduction: VC dimension is crucial in machine learning theory as it provides fundamental insights into learning algorithm behavior, generalization capabilities, and the relationship between model complexity and performance.

Key Importance:

1. **Generalization Analysis:** Higher VC dimension models can fit complex data but may overfit without sufficient data
2. **Sample Complexity Bounds:** Determines minimum data requirements for reliable learning
3. **Model Selection Guide:** Helps choose models that balance complexity and performance
4. **PAC Learning Foundation:** Essential for Probably Approximately Correct learning theory

Practical Example: When choosing between linear SVM (VC dimension = $d+1$ for d features) and RBF kernel SVM (potentially infinite VC dimension), VC theory helps determine which is appropriate based on available training data size and desired generalization.

Significance: VC dimension provides theoretical guarantees and practical guidance for ML practitioners, enabling informed decisions about model complexity, data requirements, and expected performance, crucial for developing reliable AI systems.

9. What are the challenges of PAC learning?

Introduction: While PAC learning provides valuable theoretical foundations for machine learning, several challenges limit its direct application to real-world problems, requiring additional considerations and techniques.

Key Challenges:

1. **Real-world Applicability:** Assumptions like finite hypothesis space and known data distribution often don't hold in practice
2. **Computational Complexity:** Finding optimal hypothesis can be computationally expensive for large, complex hypothesis spaces
3. **Model Assumptions:** Assumes true function exists within hypothesis space, which may not be realistic
4. **Noise Handling:** Original PAC framework doesn't account for noisy data common in real applications

Practical Example: In image recognition, the assumption that there exists a perfect classifier within the hypothesis space is unrealistic due to inherent ambiguity, lighting variations, and noise in real images, making pure PAC learning insufficient.

Significance: Understanding these challenges drives development of more robust learning frameworks, noise-tolerant algorithms, and practical approximations that bridge the gap between theoretical guarantees and real-world ML applications.