CS7641 Machine Learning
Practice Final Exam (Fall 2023)

Created by Kyle Nakamura with the help of GPT-4

Question 1

Which of the following statements are true about supervised learning?

A) It only includes classification problems.
B) It involves assigning labels to instances.
C) It can include regression problems.
D) It doesn't use instances as input vectors.

Question 2

What are characteristics of decision trees in machine learning?

A) They cannot handle Boolean functions.
B) They are optimized to prevent overfitting.
C) They use edges to predict outcomes.
D) They are enhanced by algorithms like ID3.

Question 3

In the context of neural networks, which of the following statements are correct?

A) They cannot model complex, discontinuous functions.
B) Gradient descent is used for linear cases.
C) Sigmoid functions facilitate backpropagation.
D) They adjust weights through algorithms like the Perceptron Rule.

Question 4

Which of the following are true regarding instance-based learning (IBL)?

A) It is insensitive to noise in data.
B) k-NN is an example of IBL.
C) It always requires low space during the learning phase.
D) Curse of Dimensionality impacts its performance.

Question 5

In computational learning theory, what relates to the size of the hypothesis space?

A) The complexity of the learning algorithm.
B) The sample complexity bounds.
C) The learner's speed.
D) The teacher's role in the learning process.

Question 6

What are characteristics of boosting in machine learning?

A) It can't integrate simple rules.
B) It focuses on difficult data points.
C) It always reduces the expected error.
D) It uses a weighted mean for error calculation.

Question 7

Which statements are true about Support Vector Machines (SVMs)?

A) They are ineffective in reducing overfitting.
B) They find optimal hyperplanes for classification.
C) They cannot handle high-dimensional data.
D) Kernel tricks are used for data separation.

Question 8

What are aspects of computational learning theory in machine learning?

A) It emphasizes the quantity of training data.
B) It does not consider the problem definition important.
C) It involves interactive learning between learners and teachers.
D) It focuses solely on sample complexity.

Question 9

Regarding hypothesis space in machine learning, which of the following are true?

A) Larger hypothesis spaces always lead to more accurate models.
B) The hypothesis space size affects the model's error rate.
C) Agnostic learning assumes a target space.
D) Haussler's Theorem relates hypothesis space size and sample size.

Question 10

In Bayesian learning, which statements are correct?

A) It always integrates priors via MAP hypothesis.
B) Maximum likelihood approaches ignore prior probabilities.
C) It is irrelevant in artificial intelligence.
D) Bayes' Rule is used for probable hypothesis identification.

Question 11

What are true statements about gradient descent in machine learning?

A) It is exclusively used in neural networks.
B) It involves iterative optimization using calculus.
C) It cannot be used for complex functions.
D) Randomized optimization can be an alternative to gradient descent.

Question 12

In machine learning, what are characteristics of Simulated Annealing?

A) It uses temperature changes metaphorically.
B) It is less effective than Hill Climbing for global optimization.
C) It risks suboptimal solutions with too fast cooling.
D) It always finds the global optimum.

Question 13

Which statements are true about Genetic Algorithms (GAs) in machine learning?

A) They don't use crossover and mutation for optimization.
B) Fitness-based strategies are essential in GAs.
C) They are ineffective for problem spaces with complex structures.
D) They assume solution segment independence for effective crossover.

Question 14

In the context of clustering in machine learning, which of the following are true?

A) K-means algorithm is a type of soft clustering.
B) EM algorithm alternates between likelihood assignments and center updates.
C) Single Linkage Clustering is inefficient for large datasets.
D) K-means always provides the best clustering irrespective of initial center placement.

Question 15

Regarding dimensionality reduction techniques, which statements are accurate?

A) PCA and ICA are both used for supervised dimensionality reduction.
B) PCA uses eigenvalues to maintain variance while reducing dimensions.
C) ICA separates independent variables orthogonally.
D) LDA requires labels for dimensionality reduction.

Question 16

Which are true statements about Reinforcement Learning (RL) in the context of machine learning?

A) RL exclusively focuses on immediate rewards.
B) Finite horizons in RL prompt changes due to time limits.
C) The Bellman Equation is not relevant in solving MDPs.
D) Q-learning updates Q estimates without known rewards or transitions.

Question 17

In game theory applied to machine learning, which of the following statements are correct?

A) Minimax strategy is irrelevant in game trees.
B) Nash equilibrium occurs in all finite games.
C) Stochastic games generalize MDPs and are always solvable in polynomial time.
D) In zero-sum games, coop strategies lead to the best outcomes.

Question 18

What are characteristics of the Minimax Q learning update in machine learning?

A) It converges to multiple Q* values.
B) It is used in zero-sum stochastic games.
C) It is ineffective in two-player competitive scenarios.
D) In general-sum games, Nash Equilibrium is easily computable.

Question 19

Regarding optimization algorithms in machine learning, which statements are true?

A) Randomized optimization is always slower than full evaluations.
B) Mimic uses probability distributions in evolutionary algorithms.
C) Hill climbing algorithms effectively handle multi-modal problem spaces.
D) Simulated Annealing outperforms methods like Hill Climbing in certain scenarios.

Question 20

In the context of supervised and unsupervised learning, which of the following are true?

A) Supervised learning only uses labeled data.
B) Unsupervised learning, like clustering, finds patterns in unlabeled data.
C) Unsupervised learning cannot handle metric spaces.
D) Supervised learning cannot generalize from labeled data.

Question 21

What is true about the concept of overfitting in machine learning?

A) It occurs when a model is too simple for the data.
B) It is desired in decision tree models.
C) It involves a model capturing noise in the data.
D) It is prevented by reducing the number of features in a model.

Question 22

In the context of neural networks, what does the Perceptron Rule accomplish?

A) It assists in non-linear problem solving.
B) It adjusts weights based on error margins.
C) It is used for optimizing network architecture.
D) It prevents overfitting in deep neural networks.

Question 23

What are the characteristics of polynomial regression in machine learning?

A) It always uses matrix multiplication.
B) It minimizes squared error.
C) It is suitable only for categorical data.
D) It cannot handle continuous outcomes.

Question 24

Regarding the ID3 algorithm for decision trees, which statements are accurate?

A) It focuses on minimizing information gain.
B) It selects attributes that maximize entropy.
C) It is used for splitting decision nodes.
D) It is less efficient than random attribute selection.

Question 25

What are true statements about cross-validation in machine learning?

A) It is only used for testing neural networks.
B) It helps in estimating the accuracy of a model.
C) It is irrelevant for determining optimal model complexity.
D) It involves dividing the dataset into training and testing sets.

Question 26

In the context of k-Nearest Neighbors (k-NN), which of the following are true?

A) The choice of 'k' is irrelevant for the algorithm's performance.
B) It is more efficient in the learning phase than in the query phase.
C) Weighted averages can be used for determining influence.
D) It is unaffected by the Curse of Dimensionality.

Question 27

What is true about ensemble learning in machine learning?

A) It is less effective than using single models.
B) Bagging is a method that averages predictions from data subsets.
C) Boosting is ineffective in focusing on difficult examples.
D) It only uses boosting for improving outcomes.

Question 28

Which statements are correct regarding the handling of errors in machine learning?

A) Machine learning models aim for a 0% error rate.
B) Boosting is unrelated to learning from errors.
C) Errors can arise from the distribution of examples during training/testing.
D) A weak learner is defined as an algorithm with an error rate over 50%.

Question 29

In Bayesian learning, what role does maximum likelihood estimation play?

A) It always incorporates prior probabilities.
B) It estimates parameters that maximize the likelihood of the data.
C) It is used to reduce the complexity of Bayesian networks.
D) It only applies to supervised learning scenarios.

Question 30

Regarding gradient descent, which of the following are true?

A) It is used only in linear regression models.
B) It can be used in conjunction with Newton's method.
C) It cannot escape local optima in complex functions.
D) It is unsuitable for optimizing neural network parameters.

Question 31

What are the characteristics of the Simulated Annealing algorithm in machine learning?

A) It is a type of genetic algorithm.
B) It uses a metaphorical temperature parameter.
C) It always achieves global optimization.
D) It is less efficient than deterministic algorithms.

Question 32

In the context of evolutionary algorithms, what is true about the Boltzmann distribution?

A) It is irrelevant in the selection process.
B) It is used for fitness-based selection strategies.
C) It assumes that all solutions are equally likely.
D) It is less effective than random selection.

Question 33

What are true statements about K-means clustering in machine learning?

A) It is a type of hierarchical clustering.
B) It iteratively recalculates cluster centers.
C) The algorithm ensures global optimum clustering.
D) The initial placement of centers has no impact on the results.

Question 34

In dimensionality reduction, what is a characteristic of Principal Component Analysis (PCA)?

A) It is used for supervised learning.
B) It maintains variance orthogonally while reducing dimensions.
C) It assumes data features are dependent on each other.
D) It is less effective than Linear Discriminant Analysis (LDA).

Question 35

Which statements are true about Reinforcement Learning (RL) and Markov Decision Processes (MDPs)?

A) RL does not use the concept of regret in its framework.
B) Optimal policies in MDPs seek to maximize discounted rewards.
C) Value iteration is irrelevant for solving MDPs.
D) Model-based RL does not involve planners or simulations.

Question 36

In game theory, what is true about the concept of Nash equilibrium?

A) It only occurs in deterministic games.
B) It is a state where no player benefits from changing strategy.
C) It is less relevant in stochastic games.
D) It ensures the highest possible reward for all players.

Question 37

Regarding the Minimax Q learning update, which of the following are true?

A) It is primarily used for cooperative games.
B) It is effective in solving general-sum stochastic games.
C) It applies only to single-player scenarios.
D) It converges to a unique Q* in zero-sum games.

Question 38

What are characteristics of optimization algorithms in machine learning?

A) They all guarantee to find the global optimum.
B) Hill climbing is effective in all types of problem spaces.
C) Randomized optimization can be more efficient than full evaluations.
D) Simulated Annealing represents problem space structure.

Question 39

What is true about the relationship between supervised and unsupervised learning?

A) Supervised learning is more effective than unsupervised learning in all scenarios.
B) Supervised learning generalizes from labeled data.
C) Unsupervised learning always requires metric spaces.
D) Clustering is an example of unsupervised learning.

Question 40

In machine learning, what are the characteristics of Random Restart Hill Climbing?

A) It guarantees finding the global optimum.
B) It involves restarting the algorithm from different points.
C) It is less efficient than standard Hill Climbing.
D) It maintains a record of all visited points to avoid repetition.

Question 41

What is a primary characteristic of the "Curse of Dimensionality" in machine learning?

A) It refers to the increased complexity of models in two-dimensional space.
B) It describes how algorithm performance improves linearly with added dimensions.
C) It indicates that high-dimensional spaces can lead to data sparsity.
D) It suggests that dimensionality reduction always improves model accuracy.

Question 42

In the context of neural networks, how does the concept of weight initialization impact learning?

A) Incorrect weight initialization guarantees faster convergence.
B) Proper weight initialization can prevent vanishing or exploding gradients.
C) Weight initialization has no impact on the learning rate.
D) Weights should always be initialized to zero for optimal performance.

Question 43

Regarding the ID3 algorithm in decision trees, what does the concept of information gain imply?

A) It suggests choosing the attribute that increases the entropy the most.
B) Higher information gain indicates less useful attributes for splitting.
C) Information gain is used to select the attribute that best separates the samples.
D) Information gain is irrelevant when dealing with continuous attributes.

Question 44

What is a significant challenge in applying Bayesian learning to complex problems?

A) Bayesian learning cannot handle categorical data.
B) It often involves computations that are NP-complete.
C) Bayesian methods always require large datasets for accuracy.
D) It is incompatible with non-probabilistic models.

Question 45

In machine learning, what is an inherent challenge of using the k-Nearest Neighbors algorithm in high-dimensional spaces?

A) The algorithm becomes computationally cheaper.
B) It tends to overfit due to the increased number of neighbors.
C) Distance metrics become less meaningful in high-dimensional spaces.
D) The choice of 'k' becomes irrelevant in higher dimensions.

Question 46

What is the primary objective of ensemble methods like boosting in machine learning?

A) To simplify the computation by using a single powerful model.
B) To reduce bias and variance by combining multiple weak learners.
C) To focus exclusively on reducing the model's variance.
D) To ensure that the model only learns from the most difficult examples.

Question 47

In the context of Support Vector Machines, what is the significance of using kernel tricks?

A) They are used to linearly separate data that is linearly inseparable in original space.
B) Kernel tricks simplify the computation by reducing the dimensionality of the data.
C) They are primarily used to increase the computational speed of SVMs.
D) Kernel tricks are irrelevant when dealing with non-linear data.

Question 48

What challenge does the concept of "epsilon exhaustion" present in computational learning theory?

A) It refers to the difficulty of finding a hypothesis within a small epsilon of the best possible error.
B) It suggests that algorithms become more efficient as epsilon approaches zero.
C) It indicates that large hypothesis spaces can be easily managed with small epsilon values.
D) Epsilon exhaustion is not a recognized concept in computational learning theory.

Question 49

In the context of dimensionality reduction, what is a primary limitation of using Principal Component Analysis (PCA)?

A) PCA cannot be used on datasets with missing values.
B) It assumes linear relationships between features.
C) PCA is only effective in reducing dimensions to two or three.
D) It increases the computational complexity of the learning algorithm.

Question 50

In Reinforcement Learning, how does the concept of "regret" influence policy optimization?

A) It measures the difference between the chosen action and the best possible action.
B) Regret is used to penalize policies that deviate from deterministic strategies.
C) It is irrelevant in the context of finite horizon problems.
D) Regret minimization is not a concept used in Reinforcement Learning.

Answer Key

1. B, C	11. B, D	21. C, D	31. B, D	41. C
2. B, D	12. A, C	22. B	32. B	42. B
3. C, D	13. B, D	23. A, B	33. B	43. C
4. B, D	14. B, C	24. C	34. B	44. B
5. B	15. B, D	25. B, D	35. B	45. C
6. B, C, D	16. B, D	26. B, C	36. B	46. B
7. B, D	17. B	27. B	37. D	47. A
8. A, C	18. B	28. C, D	38. C, D	48. A
9. B, D	19. B, D	29. B	39. B, D	49. B
10. B, D	20. A, B	30. B	40. B, D	50. A

Detailed Answers with Explanations

Question 1: Supervised Learning

Answer: B, C - Supervised learning involves both classification (assigning labels) and regression (assigning continuous values).

Question 2: Decision Trees

Answer: B, D - Decision trees are optimized to prevent overfitting (e.g., via pruning) and algorithms like ID3 improve their efficiency by selecting attributes that maximize information gain.

Question 3: Neural Networks

Answer: C, D - Sigmoid functions enable smooth transitions states, facilitating backpropagation in neural networks. They adjust weights using algorithms like the Perceptron Rule and gradient descent.

Question 4: Instance-Based Learning (IBL)

Answer: B, D - k-Nearest Neighbors (k-NN) is an example of IBL. IBL methods, including k-NN, are impacted by the Curse of Dimensionality, affecting their performance.

Question 5: Computational Learning Theory

Answer: B - Sample complexity bounds relate to the size of the hypothesis space, determining how much data is necessary for learning.

Question 6: Boosting

Answer: B, C, D - Boosting focuses on difficult data points, reduces expected error, and uses a weighted mean for error calculation, combining weak learners to form a stronger model.

Question 7: Support Vector Machines (SVMs)

Answer: B, D - SVMs find optimal hyperplanes for classification and use kernel tricks for data separation in non-linear cases.

Question 8: Computational Learning Theory

Answer: A, C - It emphasizes the importance of the quantity of training data and involves interactive learning between learners and teachers.

Question 9: Hypothesis Space

Answer: B, D - The hypothesis space size affects the model's error rate, and Haussler's Theorem relates hypothesis space size and sample size for error reduction.

Question 10: Bayesian Learning

Answer: B, D - Maximum likelihood approaches ignore prior probabilities, focusing on the likelihood of the data given the model. Bayes' Rule is central in Bayesian learning for probable hypothesis identification.

Question 11: Gradient Descent

Answer: B, D - Gradient descent involves iterative optimization (often used with calculus methods like Newton's method). Randomized optimization can be an alternative in some scenarios.

Question 12: Simulated Annealing

Answer: A, C - Simulated Annealing uses a metaphorical temperature parameter and risks suboptimal solutions with too fast cooling, aiming for a balance between exploration and exploitation.

Question 13: Genetic Algorithms (GAs)

Answer: B, D - GAs use fitness-based strategies and assume solution segment independence for effective crossover, which is crucial for their success.

Question 14: Clustering

Answer: B, C - EM algorithm alternates between likelihood assignments and center updates (a form of soft clustering), while Single Linkage Clustering can be inefficient for large datasets due to its computational complexity.

Question 15: Dimensionality Reduction

Answer: B, D - PCA uses eigenvalues to reduce dimensions while maintaining variance orthogonally. LDA (Linear Discriminant Analysis) requires labels for supervised dimensionality reduction.

Question 16: Reinforcement Learning (RL) and MDPs

Answer: B, D - In RL, finite horizons influence policy changes due to time limits. Q-learning updates Q estimates based on rewards and transitions, crucial in RL scenarios without a model of the environment.

Question 17: Game Theory

Answer: B - Nash equilibrium occurs in all finite games and is a state where no player can benefit by changing their strategy unilaterally.

Question 18: Minimax Q Learning

Answer: B - Minimax Q learning update is particularly relevant in zero-sum stochastic games, where it converges to a unique Q* value.

Question 19: Optimization Algorithms
Answer: B, D - Randomized optimization can be more efficient than full evaluations in certain scenarios. Simulated Annealing often outperforms methods like Hill Climbing by avoiding local optima.

Question 20: Supervised and Unsupervised Learning
Answer: A, B - Supervised learning uses labeled data, while unsupervised learning, like clustering, finds patterns in unlabeled data.

Question 21: Overfitting
Answer: C, D - Overfitting occurs when a model captures noise in the data, and it can be prevented by techniques like reducing the number of features.

Question 22: Perceptron Rule in Neural Networks
Answer: B - The Perceptron Rule adjusts weights based on error margins, an essential aspect of learning in neural networks.

Question 23: Polynomial Regression
Answer: A, B - Polynomial regression often uses matrix multiplication and aims to minimize squared error, fitting data to a polynomial curve.

Question 24: ID3 Algorithm
Answer: C - The ID3 algorithm for decision trees is used for splitting decision nodes based on maximizing information gain (reducing entropy).

Question 25: Cross-Validation
Answer: B, D - Cross-validation is a technique for estimating model accuracy and optimal complexity by dividing the dataset into training and testing sets.

Question 26: k-Nearest Neighbors (k-NN)
Answer: B, C - k-NN is more efficient in the query phase than in the learning phase. Weighted averages can be used for determining influence in the algorithm.

Question 27: Ensemble Learning
Answer: B - Bagging is a method in ensemble learning that averages predictions from different data subsets to improve model outcomes.

Question 28: Handling Errors
Answer: C, D - Errors in machine learning can arise from the distribution of examples during training/testing. A weak learner is defined as an algorithm with an error rate just better than random chance (less than 50%).

Question 29: Maximum Likelihood in Bayesian Learning
Answer: B - Maximum likelihood estimation in Bayesian learning estimates parameters that maximize the likelihood of the data, focusing on the probability of observing the data given certain parameters.

Question 30: Gradient Descent
Answer: B - Gradient descent can be used in conjunction with methods like Newton's method for optimization in various contexts, including neural networks.

Question 31: Simulated Annealing
Answer: B, D - Simulated Annealing uses a metaphorical temperature parameter and is often more efficient than deterministic algorithms in certain optimization scenarios.

Question 32: Boltzmann Distribution in Evolutionary Algorithms
Answer: B - The Boltzmann distribution is used for fitness-based selection strategies in evolutionary algorithms, aiding in choosing solutions for crossover and mutation.

Question 33: K-means Clustering
Answer: B - K-means clustering iteratively recalculates cluster centers to minimize the distance of points to these centers, aiming for better clustering with each iteration.

Question 34: Principal Component Analysis (PCA)
Answer: B - PCA reduces dimensions while maintaining variance orthogonally, effectively capturing the most significant variance in fewer dimensions.

Question 35: Reinforcement Learning (RL) and MDPs
Answer: B - In RL and MDPs, optimal policies seek to maximize discounted rewards, a key concept in these frameworks for handling time-sensitive decision-making.

Question 36: Nash Equilibrium in Game Theory
Answer: B - Nash equilibrium is a state in game theory where no player benefits from changing their strategy, given the strategies of the other players.

Question 37: Minimax Q Learning in Stochastic Games
Answer: D - In zero-sum stochastic games, Minimax Q learning converges to a unique Q* value, optimizing the strategies for two-player competitive scenarios.

Question 38: Optimization Algorithms
Answer: C, D - Randomized optimization can be more efficient than full evaluations, and Simulated Annealing represents problem space structure, helping to avoid local optima.

Question 39: Supervised vs. Unsupervised Learning

Answer: B, D - Supervised learning generalizes from labeled data, and unsupervised learning, like clustering, finds patterns in unlabeled data.

Question 40: Random Restart Hill Climbing

Answer: B, D - Random Restart Hill Climbing involves restarting the algorithm from different points, improving the chances of finding a global optimum compared to standard Hill Climbing.

Question 41: Curse of Dimensionality

Answer: C - The "Curse of Dimensionality" refers to the phenomenon where the high-dimensional space leads to sparsity of data, making it difficult to find patterns and increasing the risk of overfitting. This sparsity means that more data is required to form meaningful insights. Options A, B, and D are not aligned with this concept.

Question 42: Weight Initialization in Neural Networks

Answer: B - Proper weight initialization in neural networks is crucial for preventing issues like vanishing or exploding gradients, which can hinder the learning process. Incorrect weight initialization (like all zeros) can lead to poor convergence. Options A, C, and D are either incorrect or oversimplified statements about weight initialization.

Question 43: Information Gain in ID3 Algorithm

Answer: C - In the ID3 algorithm for decision trees, information gain is used to select the attribute that best separates (classifies) the samples at each node in the tree. It doesn't suggest choosing attributes that increase entropy or are less useful, nor is it irrelevant for continuous attributes, although handling them might require additional steps.

Question 44: Challenge of Bayesian Learning

Answer: B - One significant challenge in applying Bayesian learning, especially to complex problems, is the computational difficulty, often involving calculations that are NP-complete. This makes it challenging to compute exact probabilities for large datasets or complex models. The other options either misrepresent Bayesian learning or are too extreme.

Question 45: k-Nearest Neighbors in High-Dimensional Spaces

Answer: C - A key challenge with the k-Nearest Neighbors (k-NN) algorithm in high-dimensional spaces is that distance metrics (like Euclidean distance) can become less meaningful, making it harder to determine the actual 'nearness' of points. High dimensions do not inherently lead to overfitting or make the choice of 'k' irrelevant, and they generally increase computational complexity.

Question 46: Objective of Ensemble Methods like Boosting

Answer: B - The primary objective of ensemble methods, such as boosting, is to reduce both bias and variance in the model by combining multiple weak learners. This approach leverages the strengths of various models to achieve better overall performance. The other options misrepresent the purpose and function of ensemble methods.

Question 47: Significance of Kernel Tricks in SVMs

Answer: A - In Support Vector Machines (SVMs), kernel tricks are used to enable linear separation of data that is not linearly separable in the original feature space. By mapping data to a higher-dimensional space, kernels like the radial basis function (RBF) can make it possible to find a linear separator. The other options either misrepresent the purpose of kernel tricks or are inaccurate.

Question 48: Epsilon Exhaustion in Computational Learning Theory

Answer: A - "Epsilon exhaustion" refers to the difficulty encountered when trying to find a hypothesis within a small epsilon of the best possible error, especially in large hypothesis spaces. It is a recognized concept in computational learning theory and highlights a challenge in achieving low-error solutions. The other options either misinterpret or deny the existence of this concept.

Question 49: Limitation of PCA

Answer: B - A primary limitation of Principal Component Analysis (PCA) is that it assumes linear relationships between features. PCA projects data onto directions that maximize variance, which may not be effective if the underlying data relationships are non-linear. The other options either overstate the limitations or are incorrect.

Question 50: Concept of Regret in Reinforcement Learning

Answer: A - In Reinforcement Learning, "regret" measures the difference between the reward of the chosen action and the reward that would have been obtained by the best possible action. This concept helps in evaluating and optimizing policies by considering what could have been gained by acting differently. The other options either misrepresent the role of regret or deny its use in RL.

Additional Information

How did I make this exam?

I used GPT-3 and GPT-4 to recursively summarize the transcripts of the lectures.
The transcripts were acquired by converting the video subtitles into plain text, stripping out the subtitle formatting, and combining the hundreds of text files into one.
The transcripts were then condensed to 1% of their original size, starting with over 1.5 million characters and ending with just 15,000.

This process made it possible to copy all the lectures into a single ChatGPT prompt.
That single prompt is included below:

```
The following text is a summary of the transcripts from the lectures in CS7641
ML. Use the information in the summary to get an idea of what was taught in this
course.

Then, create a practice exam with 50 questions, following these rules:
1. The questions must be multiple-choice, with multiple correct choices per
question.
2. Each question should have at least one correct choice but no more than three
correct choices.
3. Each question must have four choices.
4. The choices should be labeled A, B, C, D.
5. Make the exam difficult but not impossible.
6. Try to disambiguate the choices and avoid writing questions that rely on a
technicality or a "trick" answer.

Do not show me the answers. Instead, show me an answer key at the end of your
response so that I can test myself.

Here is the summary of the transcribed lectures:

---

Supervised learning (SL) includes classification and regression, assigning labels
and values respectively. ML uses instances as input vectors and aims to learn
target concept mappings from input-output pairs. The hypothesis class is limited
by data and inductive learning derives functions from training sets, with testing
sets ensuring model generalization. Decision trees, with decision nodes and
edges, use attributes to predict outcomes, and are optimized for accuracy and
overfitting prevention. They handle Boolean functions, but complex ones like XOR
increase computational load. ALGs like ID3 improve decision tree efficiency by
selecting attributes that maximize information gain and minimize entropy.
Continuous outcomes in trees involve specific splitting criteria and fitting
ALGs. ML models use squared error minimization and matrix multiplication for
polynomial regression. Handling data noise and using cross-validation are crucial
in ML for estimating model accuracy and optimal complexity. Skepticism around
using RGB for categorical data like hair color persists. NN concepts include mean
```

squared error importance and their similarity to brain operations. NNs use artificial neurons and adjust inputs and weights through algos like the Perceptron Rule and gradient descent, the latter for non-linear cases. Separate adjustments are made for linear separability, gradient direction, and local optima avoidance. Sigmoid functions enable smooth transition states for backpropagation in NNs. NNs have evolved to include layers for complex, discontinuous function modeling, with cross-validation helping select the optimal architecture. Emphasis on architecture, weight initialization, and simplicity, according to Occam's razor, is critical. The shift towards instance-based learning (IBL) is noted, where ML challenges & opt. strategies are discussed. k-NN alg's noise sensitivity, overfitting, & high space-time req. during learning/query phases noted. Compares to linear regression w/ higher learning cost but faster query time. Emphasizes selecting suitable distance metrics & feature significance for k-NN, notes Curse of Dimensionality's impact & suggests weighted distance func. to counter. Optimal 'k' selection & weighted avgs mentioned for adaptability in determining influence, alongside locally weighted regression for better prediction accuracy by fitting models to local data. Decision trees, NNs, & linear regression can be improved with complex funcs. over avgs. Computational learning theory's learner-teacher dynamics, true error, epsilon exhaustion, & sample complexity bounds discussed, linking hypothesis space size to error & failure rates. Agnostic learning & infinite hypothesis space challenges acknowledged, w/ future discussions on the latter. Ensemble learning, esp. boosting, is highlighted for its effectiveness in integrating simple rules & aggregating diverse evidence. It leverages randomness in data subsets, reduces expected error & increases prediction precision in housing data. Bagging & boosting are methods that improve ensemble outcomes by averaging data subset predictions & focusing on difficult data points, respectively. Boosting also increases focus on difficult examples by using a weighted mean for error calculation. ML error rate is explained as the mismatch between actual & predicted outcomes due to example distribution during training/testing. Boosting is connected to learning from errors & distribution importance, defining a weak learner as an alg w/ an error rate under 50%. Boosting aims to create a strong classifier from weak ones by adjusting learners to outperform chance. (Note: Abbreviations were used as per instruction. However, concepts/terms like 'weighted average' and 'decision trees' do currently not have widely recognized abbreviations and were thus kept unabbreviated.)
The alg adjusts ex weights iteratively based on perf, beginning with unif dist. Alpha values are calc'd from errs, influencing which weak learners are selected. Final hyp is a wt'd avg of these learners. Boosting targets hard ex, reduces misclass and errs, similar to neural nets and wt'd nearest neighbor alg. SVMs find optimal hyperplanes via quad prog, focusing on support vectors, using dot products and high-dim proj w/ kernel tricks for data separation and err balance. Boosting and SVMs reduce overfitting and improve ML perf by combining weak hyps. Boosting may cause overfitting, esp. when strong learners are used or due to pink noise. Computational learning theory emphasizes the need for proper problem def, alg analysis, and qty of training data. In ML, interactive learning between learners and teachers, as illustrated by the "20 questions" game, aims to efficiently reduce hyp space. Challenges include large potential hyp numbers and exponential time for guessing. Sample complexity is prioritized over computation, despite the limitations of linear "20 questions" method. A mistake-bound model is suggested for minimizing errs. Self-sufficient ML relies on errors without

teacher guidance for refining alg and removing irrelevant variables. Optimal learning requires ex to differ by only one variable.
ML highlights the significance of hypothesis space's size and complexity, focusing on PAC learning w/ error ε and confidence δ to ensure high-confidence predictions. Computational and sample complexity are crucial, and version space must include the true concept, emphasizing the uniform selection in scarce data scenarios. Haussler's Theorem relates ε, hypothesis space size, and sample size for error reduction. ML compares to CS complexities, valuing data acquisition and learnable strategies. Hypothesis complexity is tested in ML through sample complexity, error tolerance, and failure rates, with agnostic learning choosing the best hypothesis sans target space assumptions. High VC dimensions indicate large learning capacity but require more data, with learnability feasible if VC dimensions are finite. Bayesian learning applies Bayes' Rule for probable hypothesis identification, essential in AI, integrating priors via MAP hypothesis or ignoring them in maximum likelihood approaches.
V:S ratio in ML lwrs w/ noise & incomp conc knowl. Max likelihood est nmlz data w/ norm noise. Simplif techniques in Bayesian lrng & grad desc incl logs for sum prod & const discard. CS7641 showcases Bayesian eval ind of hyp class, while linear reg > consts & data means. Logarithms preserve max sols & inform theory aids optimal code. ML pref simple models, dec trees w/ fewer nodes to avd overfitting; yet, Bayesian minimizes error consistent w/ Occam's Razor. Bayesian lrng sup on avg w/ comprehensive hyp assess & Bayes Nets recommended for complex distribs. Bayes Nets, using DAGs & CPTs, simplify vars for inference. Sampling is crucial for approx inference in ML; marginalization & Bayes' rule pred in uncertain conditions. Naive Bayes for spam detec efficient but risks overfitting, curbed by smoothing; faces inf datas & cond indep issues. Bayesian inference & nets tackle NP-complete probs in ML, useful for classification w/ missing attrs. Sup lrng transitions to UL; discusses ML's randomized optim for ind like chem eng. (Note: Some terms like "Bayesian Learning" were not abbreviated as their abbreviations were not provided in the original text.)
ML fine-tunes neural net and decision tree params to reduce error. A quiz, then an optimization prob using calculus, specifically iterative gradient descent via Newton's method, is introduced. For complex funcs, randomized opt or Hill Climbing alg might be used to evade local optima. "Guess My Word" opt employs a fitness func, alg w/ neighbor func, and Random Restart Hill Climbing, achieving global optima in avg 5.39 steps. V is set at 29.78, with enumeration suggested for small input ranges. To circumvent local optima, algs like quick random restarts and keeping track of visited points are used, with randomized opt often being more efficient than full evals. Simulated Annealing, similar to Metropolis-Hastings, uses metaphorical temp changes, adjusts exploration and exploitation balance, and prefers global optima as temp drops, with too fast cooling risking suboptimal sols. The text discusses GAs, inspired by natural evolution, using Boltzmann distribution for selection, fitness-based strategies, crossover, and mutation for opt. GAs assume solution segment independence for effective crossover and compare one-point and uniform crossover methods. It considers limitations of minimal memory algs like hill climbing and simulated annealing, but notes the latter can represent problem space structure. It advocates mixing alg concepts and prob distribution modeling, as in Mimic, and highlights using prob distributions in evolutionary algs. Dependability trees and minimizing KL divergence are used for feature interdependencies in ML, with Prim's alg used to maximize mutual information in dependency trees. MIMIC is found efficient for prob sampling through dependency trees and stresses precise prob distribution

identification for solving Boolean opt probs, offering a balance between complexity and capturing relationships to find optimal prob distributions, especially as fitness values near uniformity. The text addresses clustering and optimization algs in ML, noting the challenges of estimating prob distributions from uniform samples and advising generating more samples for better theta approximation (>2). For complex tasks, Mimic uses data structure for efficient solutions, despite slower iteration times and overfitting risks, outperforming methods like Simulated Annealing. In supervised learning, algs generalize from labeled data, while UL, like clustering, finds dense representations from unlabeled data, using various measures for clusters without metric spaces. Single Linkage Clustering is mentioned.
SLC clusters 'n' pts into 'K' cls w/ cubic time in 'n', which may become linear w/ data structures like Fibonacci heaps. CS7641 covers SLC for 'K=2' & K-means, latter improves elongated cls handling by iteratively recalculating centers. K-means minimizes distance-based scores until convergence depends on initial center placement & uses tie-breaking. EM does soft cls, alternating between likelihood assignments & center updates based on probabilities, likened to K-means w/ binary probabilities. Cls alg properties of richness, scale-invariance, & consistency are domain-specific, can't be simultaneously fulfilled (Kleinberg). 'Wrap' & 'filter' are fsel techniques w/ trade-offs. Transformation prioritized over fsel for preds & dimensionality mgmt., w/ PCA & ICA for data transformation. PCA uses eigenvalues to reduce dims & maintain variance orthogonally. ICA separates ind. variables w/o orthogonality, useful in the Cocktail Party Problem. RCA noted for speed but with info loss. LDA req. labels for supervised dimensionality reduction. Grad projects prove ICA's effectiveness in data structure revelation. Lecture discusses ICA vs. PCA in UL, w/ ICA more favorable in complex situations. UL, DP, & RL emphasized for homework & projects. Shannon's info theory foundational to ML, introducing entropy & variable-length encoding (e.g., Morse code), mutual information (I), and KL divergence in SL. MDPs in RL discuss maximizing expected long-term rewards via policy optimization.
RL, rooted in psych. and CS, focuses on maximizing rewards. Finite/infinite horizons affect policies, with finite prompting changes due to time limits. Regret and gamma address the time value of rewards, setting Rmax/(1-gamma) reward bounds. Optimal policy (π^*) seeks to maximize discounted rewards, distinct from non-deterministic policies focused on avg rewards. Bellman Equation, crucial for solving MDPs, finds π^* through iterative alg. utilizing value iteration and action-transition analysis. Policy iteration in RL refines policies using linear methods for expected utility max. MDPs leverage discounting for finite valuing of infinite rewards, overcoming the immortality problem. RL, seen as an API in MDPs, learns to maximize rewards based on transitions. Model-based RL uses planners, models, simulations, and algs. like value/policy iteration. Value funcs and Bellman equations are key for π^* but computationally demanding. Q-learning bypasses MDP solutions by updating Q estimates ($\hat{Q}$) iteratively without known rewards or transitions. Strategies like $\alpha t = 1/t$ ensure convergence and all state-action pairs need infinite visits for accurate sampling. Epsilon Greedy Exploration and methods like random restarts, simulated annealing address exploration-exploitation trade-off and local minima. Finally, game theory extends RL to multi-agent, zero-sum games with perfect info, integrating model learning, planning, and exploration-exploitation strategies. RL ignores function approximation and broader ML issues in this context.
Player maximizes or minimizes rewards utilizing MDPs and decision trees. RL likened to game theory strategies. Minimax strategy key for game trees and AI

alg's, esp. in alpha-beta pruning. Challenges in translating game trees to matrices noted, with matrices aiding in game value determination. Minipoker highlights hidden info and complex games, where mixed strategies outperform pure strategies due to unpredictability and stable expected values. Calculating probabilities essential in optimal strategies within a "bow tie" payoff space. Von Neumann's theorem noted for non-deterministic games; minimax/maximin strategies not always aligning in non-zero-sum games. Game value constant across different strategies, emphasizing rationality. In zero-sum games like Prisoner's Dilemma, outcomes of defection or cooperation impact jail time. Nash equilibrium is a state where no player benefits from a strategy change, given others' strategies, occurring in all finite games. Repeated gameplay, threats, sunk costs, trust, and game theory principles impact decision-making. Game theory explores strategies and incentives, with continued probability (gamma) influencing coop. or defection tendencies. Nash equilibria emerge from mutual coop. or defection with no deviation incentive. Repeated games can foster coop. through potential retaliation, with any payoff better than a minmax profile sustainable indefinitely. The Prisoner's Dilemma examined, with coop. leading to -1 avg. payoff each within possible outcomes. Minmax profiles analyzed, showing mixed strategies as superior to pure in securing scores against adversaries. Feasible preferable acceptable region identified, with Nash equilibrium maintained through adequate discount factors and coordinated strategies, deviations met with punishment. Strategies like grim trigger and tit for tat differ in retaliation, promoting ongoing coop. Nash equilibrium also critical in stochastic games, computed using linear programming, suggesting that sub-game perfect Nash equilibria with adaptive Pavlov-like strategies are computable. Poly time for 2-player games is achieved by analyzing Nash Equilibrium in stochastic games, generalizing MDPs and reducible to models like zero-sum. Discount factors affect outcomes and strat in these games vs. repeated games. In ML, zero-sum game Q values determine outcomes for 2-player competitive scenarios; 3-player games are considered general-sum. Mini-max Q learning update converges to a unique Q* but solving zero-sum stochastic games is hard. For general-sum games, Nash Equilibrium faces non-convergence and NP-hard computational challenges. Coop games use "coco values" for collaborative strategies and connect to iterated prisoner's dilemma and RL. Topics include credible threats, Folk Theorem's equilibria, ED in equilibrium computation, and cognitive hierarchy theories. The import of ongoing research and coop is highlighted.