

CS7641 ML Practice Midterm Exam

Generated by GPT-4 based on an outline of the Supervised Learning lectures.

1. What is the primary difference between classification and regression?
 - a) The type of algorithm used
 - b) The type of data used
 - c) The type of output produced
 - d) The type of error metric used
2. In a decision tree, the feature used for splitting at each node is chosen based on:
 - a) Random selection
 - b) Feature importance
 - c) Information gain
 - d) Correlation with the target variable
3. Which of the following problems is more suitable for a regression decision tree?
 - a) Predicting housing prices
 - b) Identifying spam emails
 - c) Determining patient disease categories
 - d) Classifying animals based on features
4. A perceptron is capable of solving:
 - a) Linearly separable problems
 - b) Non-linearly separable problems
 - c) Optimization problems
 - d) None of the above
5. In a neural network, what is the primary purpose of the activation function?
 - a) Reduce overfitting
 - b) Introduce non-linearity
 - c) Normalize the input features
 - d) Speed up training
6. Which of the following is true about gradient descent?
 - a) It finds the global minimum of a function
 - b) It's a greedy algorithm
 - c) It requires the computation of derivatives
 - d) It's used only in neural networks

7. What is the key principle behind k-Nearest Neighbors (k-NN) algorithm?
- a) Clustering
 - b) Dimensionality reduction
 - c) Instance-based learning
 - d) Distance-based learning
8. The curse of dimensionality primarily affects:
- a) Tree-based models
 - b) Instance-based models
 - c) Neural networks
 - d) Support vector machines
9. In k-NN, as k increases:
- a) Variance increases
 - b) Bias increases
 - c) Model complexity increases
 - d) Overfitting likelihood decreases
10. Boosting algorithms primarily aim to:
- a) Reduce bias
 - b) Reduce variance
 - c) Increase model complexity
 - d) Decrease training time
11. A weak learner is defined as a model:
- a) That performs slightly better than random guessing
 - b) That overfits to the training data
 - c) With high bias and low variance
 - d) That performs perfectly on the training data
12. In AdaBoost, the final model is a weighted sum of:
- a) All possible models
 - b) The models trained in each iteration
 - c) The models with the lowest error
 - d) The models with the highest weights
13. In SVM, the margin is defined as the:
- a) Distance between the support vectors
 - b) Distance between the closest points of the two classes
 - c) Distance between the hyperplane and the closest point from either class
 - d) Distance between the two hyperplanes separating the classes

14. Kernel trick in SVM is used for:
- a) Reducing training time
 - b) Solving linearly separable problems
 - c) Solving non-linearly separable problems
 - d) Reducing overfitting
15. The main objective of SVM optimization is to:
- a) Maximize margin while minimizing classification error
 - b) Minimize margin while maximizing classification error
 - c) Maximize margin while maximizing classification error
 - d) Minimize margin while minimizing classification error
16. PAC learning stands for:
- a) Probably Approximately Correct learning
 - b) Partially Accurate Computation learning
 - c) Perfectly Accurate Computation learning
 - d) Probably Always Correct learning
17. In PAC learning, a concept is PAC learnable if:
- a) The hypothesis output is always correct
 - b) The hypothesis output is probably approximately correct with high probability
 - c) The learner can find a hypothesis that is probably correct with high probability
 - d) The learner can find a hypothesis that is always correct
18. Version space in computational learning theory refers to:
- a) The set of all possible hypotheses
 - b) The set of hypotheses consistent with the training examples
 - c) The set of hypotheses that minimize the error
 - d) The set of hypotheses that maximize the margin
19. Bayesian learning is based on the application of:
- a) Bayes' theorem
 - b) Gradient descent
 - c) Boosting
 - d) Neural networks
20. In Bayesian classification, the most probable hypothesis given the data is computed using:
- a) Maximum likelihood estimation
 - b) Bayes' theorem
 - c) Gradient descent
 - d) Decision trees

21. Which of the following is a common metric for evaluating splits in a classification decision tree?

- a) Mean Squared Error (MSE)
- b) Gini Impurity
- c) Root Mean Square Error (RMSE)
- d) Pearson Correlation Coefficient

22. What does the ID3 algorithm primarily use to construct a decision tree?

- a) Gini Impurity
- b) Information Gain
- c) Mean Absolute Error
- d) Kullback-Leibler Divergence

23. Decision Trees are known to be prone to:

- a) Underfitting
- b) Overfitting
- c) Ridge Regression
- d) ElasticNet Regression

24. Which of the following activation functions is most commonly used in the hidden layers of a neural network?

- a) Linear Activation Function
- b) Sigmoid Activation Function
- c) Rectified Linear Unit (ReLU) Activation Function
- d) Hyperbolic Tangent Activation Function

25. The backpropagation algorithm is used in training neural networks to:

- a) Reduce variance
- b) Minimize the loss function
- c) Optimize the activation function
- d) Reduce bias

26. The weights in a neural network are updated based on:

- a) The activation function
- b) The loss function
- c) The learning rate
- d) All of the above

27. Which of the following distance metrics is commonly used in the k-NN algorithm?

- a) Manhattan distance
- b) Cosine similarity
- c) Euclidean distance
- d) Both a and c

28. The k in k-NN stands for:

- a) Kernel
- b) K-means
- c) The number of neighbors to consider
- d) The number of clusters

29. In k-NN, a smaller value of k will result in:

- a) A smoother decision boundary
- b) A more complex model
- c) Reduced overfitting
- d) Increased bias

30. Bagging aims to:

- a) Reduce bias
- b) Reduce variance
- c) Increase model complexity
- d) None of the above

31. In Random Forest, what is the main reason for using a random subset of features for splitting at each node?

- a) Reduce overfitting
- b) Increase bias
- c) Speed up training
- d) Both a and c

32. Which of the following ensemble methods trains learners sequentially?

- a) Bagging
- b) Boosting
- c) Stacking
- d) Random Forest

33. In SVM, a soft margin allows for:

- a) Faster training
- b) Some misclassifications
- c) Linear separability
- d) Kernel trick

34. The C parameter in SVM controls:

- a) The width of the margin
- b) The complexity of the kernel function
- c) The penalty for misclassification
- d) The learning rate of the optimization algorithm

35. The dual problem in SVM allows for:
- a) The use of kernel trick
 - b) Faster optimization
 - c) Soft margin classification
 - d) Feature scaling
36. The concept of Occam's Razor in machine learning is closely related to:
- a) Bias-variance trade-off
 - b) Overfitting and underfitting
 - c) Preference bias in learning algorithms
 - d) All of the above
37. In the context of machine learning, what does the No Free Lunch Theorem imply?
- a) There is no single best algorithm for all tasks
 - b) All algorithms perform equally well when averaged over all possible problems
 - c) Complex models always perform better
 - d) Both a and b
38. A hypothesis h is said to generalize well from a training set S if:
- a) The error of h over S is zero
 - b) The error of h over S and unseen examples is similar
 - c) The error of h over unseen examples is zero
 - d) None of the above
39. Maximum Likelihood Estimation (MLE) in the context of Bayesian learning is used to:
- a) Estimate the parameters of the posterior distribution
 - b) Estimate the parameters of the likelihood function
 - c) Estimate the parameters of the prior distribution
 - d) None of the above
40. Which of the following is an assumption of Naive Bayes classifier?
- a) Features are linearly separable
 - b) Features are conditionally independent given the class label
 - c) All features are equally important
 - d) Training data follows a normal distribution
41. In the context of decision trees, what does the term "pruning" refer to?
- a) Reducing the depth of the tree to prevent overfitting
 - b) Removing features that have low importance
 - c) Reducing the number of samples in the dataset to speed up training
 - d) Removing misclassified samples from the dataset

42. In a multi-layer perceptron with a single hidden layer, how is the error back-propagated from the output layer to the hidden layer?

- a) Using forward propagation
- b) By computing the gradient of the loss function with respect to the weights
- c) By adjusting the activation function in the hidden layer
- d) By using a different optimization algorithm

43. Why is it advantageous to use mini-batch gradient descent over batch gradient descent?

- a) It computes the exact gradient of the loss function
- b) It allows for faster convergence to the minimum of the loss function
- c) It requires less memory
- d) Both b and c

44. In k-NN, why might it be beneficial to use a weighted voting scheme when determining the class label?

- a) To give more importance to closer neighbors
- b) To give more importance to further neighbors
- c) To reduce the computational complexity
- d) To ensure that all neighbors have equal influence on the decision

45. In boosting, what happens to the distribution of the training data for the learner at each subsequent iteration?

- a) It remains unchanged
- b) It is skewed towards misclassified samples from the previous iteration
- c) It is skewed towards correctly classified samples from the previous iteration
- d) It is randomly re-sampled

46. In SVM, what is the effect of having a very large value of the C parameter?

- a) It allows more misclassifications
- b) It makes the margin softer
- c) It makes the margin harder
- d) It has no effect on the margin

47. In computational learning theory, what is the significance of the VC dimension?

- a) It measures the capacity of a learning algorithm
- b) It measures the speed of a learning algorithm
- c) It measures the accuracy of a learning algorithm
- d) It measures the robustness of a learning algorithm

48. Which of the following best describes the Minimum Description Length (MDL) principle in the context of machine learning?

- a) Selecting the model that minimizes the description length of the data
- b) Selecting the model that minimizes the description length of the model itself
- c) Selecting the model that minimizes the combined description length of the model and the data
- d) Selecting the model that maximizes the description length of the data

49. In Bayesian learning, what does the Maximum A Posteriori (MAP) hypothesis refer to?

- a) The hypothesis that maximizes the likelihood of the data
- b) The hypothesis that maximizes the prior probability
- c) The hypothesis that maximizes the posterior probability given the data
- d) The hypothesis that minimizes the posterior probability given the data

50. When using Naive Bayes for text classification, why is the "bag of words" model commonly used?

- a) It preserves the order of words in the text
- b) It simplifies the computation of probabilities
- c) It captures semantic relationships between words
- d) It accounts for the frequency of each word in the text

Answer Key

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

43: D

44: A

45: B

46: C

47: A

48: C

49: C

50: B