1- Which are two “main approaches to achieve AI” where by AI we mean the definition in Lecture one "A machine is considered intelligent if it can perform tasks which are considered intelligent when carried out by a human being"?

A. Supervised learning and unsupervised learning
B. Supervised learning and reinforcement learning
C. Supervised learning, unsupervised learning, and reinforcement learning
D. Rule-based and Machine learning

2- Which statement is true? (you may interpret “kind of” as “subset of”)

A. AI is a kind of Machine Learning
B. Machine Learning is a kind of AI
C. Machine Learning is a kind of Deep Learning
D. Deep Learning is a kind of Machine Learning

3- Which statement is false?

A. Machine learning explicitly codes knowledge as formulas or rules
B. Machine learning is inductive
C. Rule-based approach is deductive
D. The goal in Machine Learning is to generalize to unseen data

4- Which statement does characterize Artificial Intelligence correctly based on the Turing test?

A. A machine is intelligent when its output is indistinguishable from that of a human.
B. A machine is considered intelligent if it can perform tasks which are considered intelligent when carried out by a human being.
C. A machine is considered intelligent if it can perform tasks which are considered intelligent when carried out by an intelligent mammal.
D. A machine is considered intelligent if it can perform tasks which are considered intelligent when it can adapt to a new environment.

5- In simulated annealing, how can we increase “exploitation”?

A. Increasing the step-size
B. Decreasing the step-size
C. Increasing the temperature
D. Decreasing the temperature

6- What is the major and common issue that algorithms such as greedy search and hill climbing face when solving a “minimization problem”?

A. Converging to a local maximum
B. Getting stuck in a local optima
C. Converging to a saddle point
D. Diverging to infinity

7- In gradient descent and ascent, which of the following does tell us the *direction* of solution update?

A. Gradient
B. Step-size
C. Batch-size
D. Temperature

8- Which one is NOT a “Discrete” optimization problem?

A. Traveling salesman problem.
B. Finding a course time table with the minimum number of clashes for registered students.
C. Knapsack problem: Given a set of items, each with a weight and a value, determine which items to include in the collection so that the total weight is less than or equal to a given limit and the total value is as large as possible.
D. Designing parts of an airplane.

9- Which type of Machine Learning is inspired by Operant Conditioning in Psychology (behaviors are modified by association with the addition of reward)?

A. Supervised Learning
B. Unsupervised Learning
C. Reinforcement Learning
D. Logistic Regression

01- For regression, we often use the mean squared error (MSE) loss function instead of for example average of errors. Why is this?

A. To eliminate negative values and to check how far the prediction is from the correct value so we don’t want positive and negative errors from individual data points to cancel each other out.
B. To eliminate negative values and to check how far the prediction is from the correct value so we want smaller loss when the gap is larger.
C. Because we want to make sure that the overall loss function is differentiable.
D. Because we want to make sure that the overall loss function becomes a convex function.

11- Consider the problem of predicting whether a loan application will be approved (Yes/No) based on applicant information using the following dataset:

ApplicantID	Gender	Marital Status	Education	Employment Type	Income	Loan Amount	Property Area	Loan Status
1001	Male	Married	Graduate	Salaried	High	Large	Urban	Yes
1002	Female	Single	Non-Grad	Self-Employed	Low	Medium	Rural	No
1003	Male	Married	Graduate	Salaried	High	Medium	Semi-Urban	Yes
1004	Female	Married	Non-Grad	Salaried	Medium	Small	Urban	No

What is the most common algorithm to solve this problem given the types of features?

A. kNN
B. Perceptron
C. Decision trees
D. Deep Neural Network such as ResNet-150

12- Which statement is Wrong about about kNNs?

A. For a given *k*, it calculates the distance to all the training instances and picks the *k* nearest ones, followed by choosing the majority class.
B. Selecting a small *k* may lead to overfitting to training data.
C. Selecting a large *k* may lead to a very general but ineffective model.
D. kNN can be trained very efficiently and it is also easy, fast, and efficient to make predictions for new instances.

13- Suppose you trained a machine learning model for a binary classification task and after evaluation on the test set and presenting the confusion matrix, you obtained these numbers for the evaluated examples:

True positives, tp=100
False positives, fp=50
False negatives, fn=50
True negatives, tn=200

What is the correct accuracy in percent?

A. 66.67%
B. 75%
C. 25%
D. 50%

14- Regarding feature scaling and normalization, which approach is recommended according to Marsland?

A. Use the training data to determine the scalar/normalizer and then use the same scalar on test data
B. Use the test data to determine the scalar/normalizer and then use the same scalar on test data
C. Use both training and test data to determine the scalar/normalizer and then use the same scalar on test data
D. Use both training and test data to determine the scalar/normalizer but do not apply any scalar on test data

15- A single-layer perceptron with a step activation function can only classify linearly separable data. Which of the following statements explains why a single-layer perceptron fails to classify data that is not linearly separable, such as data points forming an XOR pattern?

A. It is computationally infeasible to adjust the weights and bias of a single-layer perceptron to learn non-linear decision boundaries.
B. The perceptron activation function can only produce outputs of 0 or 1, which limits its ability to handle continuous outputs.
C. A single-layer perceptron lacks multiple layers and non-linear activation functions that would allow it to model complex relationships in data.
D. The learning rate in a perceptron is too low to converge on non-linear patterns.

16- Suppose you have a function $h(x,y) = (a + bx + cy)^2$ where x and y are input variables and a, b, and c are some constants. What is the gradient of h at the origin (x=0, y =0)

A. (0, 0)
B. (a, b, c)
C. (2ab, 2ac)
D. (2b, 2c)

17- Consider the following problems:

Problem A: Predict the temperature tomorrow

Problem B: Assign a label from a set of 5 labels to each image

Which Statement is true:

A. Problem A is regression and Problem B is classification.
B. Problem A is classification and Problem B is regression.
C. Problem A is regression and Problem B is regression.
D. Problem A is classification and Problem B is classification.

18- Consider a simple linear regression problem with parameters *w_0, w_1, …., w_m* To measure the performance, one may consider, mean squared error (MSE), root mean squared error (RMSE), or squared error (SE). What is common among MSE, RMSE, and SE?

A. The value of different error functions are the same under the same parameters
B. All error functions are practical and useful when we have training and test sets with different sizes
C. All of them are minimized under the same parameters *w_i*'s
D. None of them are practical and useful when we have training and test sets with different sizes

19- For binary classification and one-hot labels, what is the cross-entropy loss of a single data point *(x_i,y_i)*?

A. The absolute difference between the prediction of model and ground truth
B. The probability of predicting the ground truth
C. Applying ReLU on top of a linear mapping (weighted sum of input and weights)
D. Negative logarithm of the predicted probability of the ground-truth class

20- When we have a classification problem with a finite set of *n* different labels with *n>2,* and we choose one label from the set of labels for each input, what kind of problem do we have?

A. Binary classification
B. Multi-class classification
C. Multi-label classification
D. Logistic regression

21- In multinomial logistic regression for multi-class classification, we first obtain the weighted sums of the input features and weights and biases.Then we apply some activation functions on the weighted sums to obtain predicted probabilities for each class (*y*'s) whose cross-entropy loss is computed with the one-hot ground-truth targets. What activation function is used to obtain prediction probabilities for each class?

A. Softmax function
B. Logistic Sigmoid function
C. ReLU function
D. Step function

22- What is the main reason sigmoid activation is rarely used nowadays to provide non-linearity within hidden layers of basic neural networks such as fully-connected feed-forward neural networks?

A. Because sigmoid activation is not differentiable
B. Because computing the output of sigmoid for a given input is computationally very complex tasks for computers
C. Because it suffers from gradient saturation (vanishing gradient)
D. Because its output can be very large and even unbounded

23- To minimize the empirical risk, we may use stochastic gradients to control computation complexity when the number of training data is very large. But then the variance and randomness in gradient updates may be substantial. How can one control the variance?

A. Using a larger step-size
B. Using a smaller step-size
C. Using a development set and cross-validation
D. Computing gradients over a mini batch

24- Which of the following does NOT explain why convolutional neural networks work and become popular?

A. Having an overparameterized architecture with one fully-connected and very wide layer and a few other fully-connected layers
B. Parameter sharing
C. Localized pattern detection over all the input using inner product as a similarity measure
D. Hierarchical feature/pattern recognition across layers where complex features are extracted in deeper layers

25- If a genetic algorithm is stuck in a local minimum, which of the following modifications is most likely to help escape it?

- A) Increase the mutation rate
- B) Decrease the crossover rate
- C) Increase the selection pressure
- D) Reduce the population size

26- Suppose we use Partially Mapped Crossover (PMX) on the following two parent tours with a randomly chosen cut point between the 3rd and 4th bits and another cut point between the 6th and 7th bits:

Parent 1: P1 = (3, 4, 8, |2, 7, 1,| 6, 5)

Parent 2: P2 = (4, 2, 5, |1, 6, 8,| 3, 7)

Hint: Partially Mapped Crossover (PMX) algorithm:

- *1. Substring selection: Cut each parent into two substrings, and then select one substring for each parent at random.*
- *2. Substring exchange: Exchange the two selected substrings to produce proto-offspring.*
- *3. Mapping list determination: Determine the mapping relationship based on the selected substrings.*
- *4. Offspring legalization: Legalize proto-offspring with the mapping relationship.*

What offspring can PMX generate from these parents? (Choose TWO answers)

- A) (3, 4, 5, 2, 7, 1, 6, 8)
- B) (4, 8, 5, 2, 7, 1, 3, 6)
- C) (4, 2, 5, 1, 6, 8, 3, 7)
- D) (4, 2, 5, 2, 7, 1, 3, 5)
- E) (4, 2, 5, 1, 6, 8, 3, 7)
- F) (3, 4, 2, 1, 6, 8, 7, 5)
- G) (4, 2, 7, 1, 6, 8, 5, 3)
- H) (3, 4, 8, 2, 7, 1, 6, 5)

27- For which of the following problems is a genetic algorithm generally *not* very efficient?

Hint:

- *N-Queens Problem: For an NxN chess board, place N queens such that no queen is able to attack another.*
- *Job-Shop Scheduling : Assigning a series of jobs to resources (like machines) while minimizing total completion time and ensuring no overlap for jobs requiring the same machine.*
- *Graph Coloring Problem: Coloring the graph vertices in such a way that no two adjacent vertices have the same color.*
- *Knapsack Problem: Given a set of items, each with a weight and a value, determine which items to include in the collection so that the total weight is less than or equal to a given limit and the total value is as large as possible.*
-
- A) Finding Large Prime Numbers
- B) N-Queens Problem (for large N)
- C) Job-Shop Scheduling
- D) Graph Coloring Problem
- E) Knapsack Problem

28- In a genetic algorithm, a population has the following fitness values for five individuals: [12, 15, 20, 25, 28]. Calculate the probability of selecting the individual with a fitness of 28 using fitness-proportionate selection (roulette wheel).

- A) 0.14
- B) 0.28
- C) 0.4
- D) 0.56

29- You are using a roulette wheel selection method in a genetic algorithm for a problem with a fitness landscape where most individuals have similar fitness values. What issue might arise with this selection method, and which alternative selection strategy would address this issue?

- A) Premature convergence; switch to tournament selection
- B) Loss of selection pressure; switch to rank-based selection
- C) Increased diversity; switch to elitist selection
- D) Lack of convergence; switch to random selection

30- Which of the following statements about tournament selection in a genetic algorithm are true? (Select TWO)

- A) Increasing tournament size reduces selection pressure.
- B) Larger tournament sizes favor fitter individuals.
- C) Tournament selection has constant selection pressure regardless of tournament size.
- D) Smaller tournament sizes increase diversity.
- E) Tournament selection is guaranteed to select the top individual.
- F) Picking with replacement increases selection pressure.

31- For tournament selection with a sample size k=3 in a population of 6 individuals, calculate the probability that the individual with rank 1 is selected in a single tournament (assuming no replacement.)

- A) 0.2
- B) 0.3
- C) 0.33
- D) 0.5

32- If an evolutionary algorithm applies fitness sharing with a niche size σ_share=2, and three individuals with fitness values [9, 8, 7] are located within a distance less than σ_share, calculate the shared fitness for the individual with fitness 8. Assume sharing function sh(d)=1−d/σ_share.

- A) 2.6
- B) 4
- C) 6
- D) 8.1

33- Suppose you are tuning hyperparameters using *grid search* and *k-fold cross-validation* with *k=5*. Approximately how many models will be trained in total if *three* hyperparameters are each tested with *five* different values?

- A) 75
- B)125
- C) 625
- D)1215

34- For a binary classification problem with a large class imbalance, you decide to evaluate the model using both ROC-AUC and Precision-Recall AUC. The model achieves a high ROC-AUC but a low Precision-Recall AUC. What is the most likely reason for this discrepancy?

- A) The model has high sensitivity but low specificity.
- B) The model has high precision but low recall.
- C) The model performs well on the majority class but poorly on the minority class.
- D) The model is underfitting both classes.

35- For a model's predictions on a binary classification problem at a given threshold, the F1-score is 0.75. The model has 40 true positives (TP) and 10 false positives (FP). Calculate the model's recall.

- A) 0.50
- B) 0.60
- C) 0.67
- D) 0.71

36 -For a binary classifier, the precision-recall curve is particularly useful compared to the ROC curve when:

- A) Classes are balanced, providing equal distribution of positives and negatives.
- B) The model requires a balance between sensitivity and specificity.
- C) The dataset is highly imbalanced with a rare positive class.
- D) The classifier has high accuracy but low AUC.

37- Which of the following techniques can help reduce variance in a model?

- A) Adding more features
- B) Using ensemble methods like bagging
- C) Decreasing the size of the training data
- D) Increasing the model complexity

38- In a situation where the data is linearly separable, a logistic regression model is trained without regularization. This model achieves perfect accuracy on the training set but fails on the test set. Which issue best explains this performance?

- A) High bias due to underfitting
- B) High variance due to overfitting
- C) Model collapse due to insufficient data
- D) Inadequate features for classification
- E) high bias due to overfitting

39- Which of the following statements are true? (Select TWO)

- A) Ridge regularization is likely to lead to a sparse model by shrinking some coefficients exactly to zero.
- B) Lasso regularization is likely to lead to a sparse model by shrinking some coefficients exactly to zero, which is helpful for feature selection.
- C) Early stopping is likely to lead to a sparse model by reducing model complexity.
- D) Adding more data to a model with high bias is likely to reduce the model's overfitting.
- E) Adding more data to a model with high variance can sometimes reduce the model's overfitting.

40- In a regularized linear regression model, both Ridge (L2) and and Lasso (L1) regularization terms are applied. The regularization parameters are set as $\lambda_2=0.5$ and $\lambda_1=0.3$ for L1 regularization. Given the weight vector $w=[2, -1, 3]$, what is the total penalty added to the loss function by the regularization terms? (*Hint: Calculate the L2 penalty and the L1 penalty separately, then add them to get the total penalty*.)

- A)2.8
- B) 4.2
- C) 6
- D) 8.8

41- Which of the following methods is best suited for visualizing and preserving the local neighborhood relationships in data when reducing dimensions?

- A) Principal component analysis (PCA)
- B) Linear Discriminant Analysis (LDA)
- C) t-Distributed Stochastic Neighbor Embedding (t-SNE)
- D) Autoencoders
- E) K-means

42- In hierarchical clustering, suppose you choose Ward linkage to determine the distance between clusters. What characteristic does Ward linkage prioritize when merging clusters?

- A) Minimizing the maximum distance between samples in merged clusters
- B) Minimizing the increase in within-cluster variance
- C) Maximizing the number of clusters at each stage
- D) Maximizing the mean distance between clusters

43- In a k-means clustering application, the final clusters have high within-cluster variance. Which of the following techniques could help improve clustering by reducing within-cluster variance?

- A) Change the distance metric to Manhattan distance
- B) Decrease the number of clusters k
- C) Increase the number of clusters k
- D) Use average linkage

44- Given a dataset with highly correlated features, you decide to apply Principal Component Analysis (PCA). What would be the expected effect of PCA on the dataset?

- A) It will create new features that are linear combinations of the original features with maximized variance and reduced dimensionality.
- B) It will increase the dimensionality of the dataset by adding orthogonal features.
- C) It will reduce the dataset to one-dimensional data regardless of the original dimensionality.
- D) It will remove all correlations between features without changing the dimensionality.

45- Which of the following statements is NOT true about autoencoders?

- A) Autoencoders learn to produce an output similar to the input.
- B) Autoencoders require labeled data for training.
- C) Autoencoders can be used for dimensionality reduction by learning a compressed representation of the data through middle layers.
- D) Autoencoders can be used as a pretraining method to provide a useful weight initialization for other tasks like classification.

46- In Support Vector Machines (SVMs) with a linear kernel, what is the primary purpose of maximizing the margin between classes, and how does it impact the decision boundary?

- A) Maximizing the margin reduces the generalization error by increasing the distance between support vectors.
- B) The decision boundary becomes a non-linear function, making it suitable for complex datasets.
- C) Maximizing the margin reduces the impact of outliers by minimizing the distance of each data point from the boundary.
- D) The support vectors themselves are adjusted to minimize the classification error within each class.

47- What is the primary difference between Strong AI and Weak AI? (Choose TWO)

- A) Strong AI claims AI can truly understand and have minds.
- B) Weak AI is limited to performing specific tasks without understanding.
- C) Strong AI is less computationally intensive.
- D) Weak AI seeks to create machines with consciousness.
- E) Strong AI simulates tasks with human-like efficiency.
- F) Weak AI operates independently of human-designed rules.

48- The "Chinese Room" argument by John Searle challenges which assumption in AI?

- A) That AI can replicate human emotions
- B) That AI can only perform narrow, specific tasks
- C) That symbolic AI is inferior to connectionist approaches
- D) That computation alone is sufficient for understanding

49- How does the complexity of an AI model generally affect its explainability, and why is this relationship important in the context of Explainable AI? (Select TWO)

- A) As model complexity increases, explainability typically decreases, making it harder for humans to understand the decision-making process.
- B) Complex models, such as deep neural networks, are usually easier to explain than simpler models like decision trees.
- C) Explainable AI techniques are more effective on complex models than on simple models.
- D) Simple models, like linear regression, are usually less transparent, making them more difficult to explain compared to more complex models.
- E) Lack of explainability in complex AI models can lead to ethical issues, as it becomes difficult to identify and correct biases in decisions that impact individuals.

50- Choose **two correct statements** among all statements:

- Pure exploitation may find a local optimum by gradually improving solutions until we can’t find a better solution
- Global optimum can be worse than some neighboring solutions.
- Exploration means moving (randomly) around in the search landscape, trying to find new types of solutions that potentially are better than what we have found before.
- Local optimum can be worse than some neighboring solutions.