

Trial exam 3050 2025 with answers

Part 1 – Multiple Choice Questions

You can assume there is 1 correct answer unless otherwise specified.

There are 35 questions. The multiple choice part counts for 70% of the final score – each question counts equally (2 points).

Search and Optimization

- 1) Choose **two** correct statements among all statements:
 - a. Pure exploitation may find a local optimum by gradually improving solution until we can't find a better solution
 - b. Global optimum can be worse than some neighboring solutions.
 - c. 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.
 - d. Local optimum can be worse than some neighboring solutions.
- 2) 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
- 3) 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.
- 4) 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

Evolutionary Algorithms

- 5) 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
- 6) Which two are true for permutation representations?
- a. They are used for problems where each variable can only appear once
 - b. Bit-flip mutation is applicable
 - c. A mutation operator that swaps at least two values is applicable
 - d. They are used for problems where each variable can appear multiple times
- 7) 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
- 8) 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 elitist selection
 - b. Loss of selection pressure; switch to rank-based selection
 - c. Increased diversity; switch to selection with fitness sharing
 - d. Lack of convergence; switch to random selection
- 9) 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.

Supervised learning

- 10) 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.

11) Which statement is wrong 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 scales well to large problem sizes

Evaluation of classifiers

12) 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%

13) 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

14) Which scenario is likely to cause a model's macro F1-score to be significantly lower than its accuracy?

- a. The dataset is perfectly balanced between classes.
- b. The classifier performs well on both classes.
- c. The dataset is highly imbalanced, and the model mostly predicts the majority class correctly.
- d. The model's predictions are equally distributed across all classes.

Linear regression

15) 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.
- 16) Consider a simple linear regression problem with parameters $w_0, w_1, \dots, w_m$. 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
- 17) In a regularized linear regression model, Ridge (L2) regularization is applied with a regularization factor $\alpha=0.5$. If the sum of squared coefficients (weights) without regularization is 10, what is the total loss added by the regularization term?
- a. 2.5
 - b. 5.0
 - c. 10.0
 - d. 0.5

Logistic regression

- 18) 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
- 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) 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

21) In a bias-variance trade-off, a highly complex model with low bias is more likely to:

- a. Generalize well to new data
- b. Underfit the training data
- c. Have high variance and overfit the training data
- d. Perform consistently across different datasets

Artificial neural networks (MLPs)

22) A perceptron with a step activation function can only classify linearly separable data.

Which of the following statements explains why such a 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 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 classical perceptron lacks 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.

23) You are given a feed-forward neural network (FFNN), also known as multi-layered perceptron (MLP). It employs ReLU activation function, and the weight values of the first layer are given below. No bias term is used. What is the output of this first layer (after applying the activation function) on the data instance $x=[1, 2, -2, 1]$?

-1 1 -2 2
1 -1 1 1
1 1 -1 1
2 -2 1 -1

- a. [1, 5, 3, 1]
- b. [1, 0, 3, 1]
- c. [1, -5, 3, 1]
- d. [1, 0, -3, 1]

24) You are given a feed-forward neural network (FFNN), also known as multi-layered perceptron (MLP). It performs multinomial classification of the input data instances into three classes. The network employs softmax after the output layer. No bias term is used. The weight values of the last layer are given below. What class this network will predict if the input to the last layer is $h=[4, 0, 2, 0]$?

2 1 1
2 2 1

1 2 2
1 1 2

- a. Unknown class
- b. Class 1
- c. Class 2
- d. Class 3

Unsupervised learning

- 25) In k-means clustering, what happens to the cluster centers at each iteration?
- a. They are randomly repositioned
 - b. B) They are fixed at the initial positions
 - c. They are updated by moving towards neighboring clusters
 - d. They are updated to the mean of the assigned samples
- 26) 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
- 27) 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.
- 28) 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.

Reinforcement Learning

- 29) In Reinforcement Learning
- a. The reward tells us which action we should have taken
 - b. The reward is never delayed
 - c. The reward is always a positive value
 - d. It is possible to receive the entire reward at the end

- 30) The Q-learning algorithm
- a. Assumes future actions will be optimal when calculating Q-values
 - b. Calculates values by averaging all possible actions
 - c. Is a type of on-policy learning
 - d. Results in less risky solutions than SARSA

- 31) What is the objective of a reinforcement learning agent?
- a. To minimize errors
 - b. To maximize accuracy
 - c. To maximize rewards
 - d. To minimize computational resources

Philosophy, history and ethics of AI

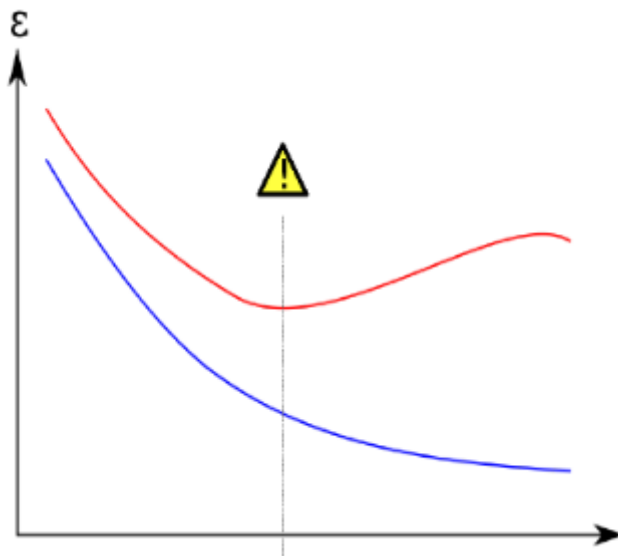
- 32) What is the primary difference between Strong AI and Weak AI? (Choose TWO)
- a. Strong AI claims AI can truly understand and have minds.
 - b. Weak AI is limited to performing specific tasks without understanding.
 - c. Strong AI is less computationally intensive.
 - d. Weak AI seeks to create machines with consciousness.
 - e. Strong AI simulates tasks with human-like efficiency.
 - f. Weak AI operates independently of human-designed rules.
- 33) The "Chinese Room" argument by John Searle challenges which assumption in AI?
- a. A) That AI can replicate human emotions
 - b. 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
- 34) 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.
- 35) The philosophical concept of "connectionism" in AI is most closely associated with which of the following approaches?
- a. Rule-based expert systems
 - b. Symbolic logic and reasoning
 - c. Neural networks
 - d. Autonomous robotics
 - e. Genetic algorithms

Part 2 – Text questions

Evolutionary Algorithms (6p)

In Evolutionary Algorithms, crossover mechanisms and representations are closely connected. Intermediate crossover (also known as arithmetic crossover) is the name of a type of crossover that can be used for a specific type of representation in Evolutionary Algorithms. Describe briefly how intermediate crossover works

Training and evaluating supervised models (6p)



In the plot above, the horizontal axis stands for the number of training steps, and the vertical axis stands for loss values. The blue line shows how the loss is changing on the training set, and the red line shows how the loss is changing on the validation (development) set.

Explain what happens at the point labelled with the yellow exclamation sign. If you were training this model, what would you do at this point? Why?

Artificial neural networks (MLPs) (6p)

Artificial neural networks often use non-linear activation functions. Explain what their role is in supervised machine learning, and what advantages they bring. Come up with one example of a problem which cannot be solved by stacking any number of hidden layers with linear activation functions, but only by adding a non-linearity.

Deep learning (6p)

Why are deep neural networks often described as “representation learners”? What are these “representations” and how they are related to “embeddings”?

Reinforcement Learning (6p)

Describe two methods for action selection in RL, and how these two methods affect the balance between exploration and exploitation in an RL-algorithm. You can assume we have learned estimates of the Q-values already, and are now interested in how to use these Q-values to select the next action.