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.

Which one is NOT a “Discrete” optimization problem? D

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.

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.

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

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.

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

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

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

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

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? C

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