Introduction to Data Science
DS2001

26 October 2023
15:00 – 19:00

ITE, Halmstad University
Contact: Mohamed-Rafik BOUGUELIA

Instructions:

The exam consists of a total of 80 multiple-choice-questions. Each question has four options (A, B, C, D), and **only one** of the four options is correct. You get 1 point if you select the correct option, and 0 otherwise.

The questions from 1 to 48 are related to Python programming; and the questions from 49 to 80 are related to data science / machine learning.

Your answers should be marked on the provided **answer sheet form**. For each question, mark your selected option with an **X** (a cross) in the corresponding box. Remember: **only one** option should be marked with an **X** (marking more than one option would give you 0 point for that question). So, fill in the answer sheet form carefully.

To pass with grade G, you need at least 40 points in total (i.e., 50%). To pass with grade VG, you need at least 64 points in total (i.e., 80%).

Good Luck!

First part – Questions 01 to 48 related to Python programming

Question 01:

What is the correct way to create a variable called `firstname` in Python?

```
(A)       var firstname = "Emma"
(B)       firstname = "Emma"
(C)       variable firstname: "Emma"
(D)       create firstname as "Emma"
```

Question 02:

What is the result of the expression `(2 > 4 or 5 > 3)` in Python?

(A) True
(B) False
(C) "No"
(D) None

Question 03:

What will be the output of the following code?

```
a = 10
b = 3
print(a % b)
```

(A) 1
(B) 0
(C) 3
(D) 10

Question 04:

What does the `!=` operator represent in Python?

(A) Less than or equal to
(B) Equal to
(C) Not equal to
(D) Greater than or equal to

Question 05:

What does the `super()` function do in Python?

(A) It prints the attributes of the parent class
(B) It is used to call a method from the child class
(C) It is used to call a method from the parent class
(D) It is used to create a new object of the class

Question 06:

Consider the matrix multiplication M @ D shown in the following code:

```
import numpy as np

M = [[1, 2, 3],
     [4, 5, 6]]

D = [[7,  8,  9, 10],
     [11, 12, 13, 14],
     [15, 16, 17, 18]]

M = np.array(M)
D = np.array(D)

result = M @ D
```

(A) The code is ok, and the two matrices can be multiplied.
(B) The matrices cannot be multiplied due to a mismatch in dimensions: the number of rows in M should be equal to the number of columns in D.
(C) The matrices can be multiplied, but the result will be an empty matrix `[]`.
(D) Python will raise a syntax error because the matrices are not compatible for multiplication.

Question 07:

You have a robot that can walk one step forward if you call the function `step_forward()`, or can turn (left or right to explore its surroundings) if you call the function `make_turn()`. The robot continuously moves. However, each time, you want it to randomly make a step forward with a probability of 0.8 (i.e., most of the time), or make a turn with a probability of 0.2 (i.e., sometimes). Which of the following codes allows you to do that?

Code (A):

```
import numpy as np
r = np.random.uniform(0, 1)
while True:
    if r < 0.8:
        step_forward()
    else:
        make_turn()
```

Code (B):

```
import numpy as np
while True:
    r = np.random.uniform(0, 1)
    if r > 0.8:
        step_forward()
    else:
        make_turn()
```

Code (C):

```
import numpy as np
while True:
    r = np.random.uniform(0, 1)
    if r < 0.2:
        make_turn()
    else:
        step_forward()
```

Code (D):

```
nb_forward = 0
nb_turn = 0
nb_total = 0
while True:
    nb_total = nb_total + 1
    if nb_forward / nb_total < 0.8:
        step_forward()
        nb_forward = nb_forward + 1
    elif nb_turn / nb_total < 0.2:
        make_turn()
        nb_turn = nb_turn + 1
```

Question 08:

You have some input data stored in a matrix X (two-dimensional numpy array) where the rows correspond to different patients, and the columns correspond to different numerical features describing those patients. You want to compute the average value for each separate feature. How would you do that in Python?

(A) `X.mean(X)`
(B) `np.mean(X)`
(C) `X.mean(axis=1)`
(D) `np.mean(X, axis=0)`

Question 09:

You have a two-dimensional numpy array called X (i.e., a matrix with several rows and columns).
What does it mean when you do X[:2, 4]

(A) Selects rows 0 and 1, and columns 0, 1, 2 and 3.
(B) Selects rows 0, 1 and 2, and column 4
(C) Selects all rows, and columns 2 and 3
(D) Selects rows 0 and 1, and column 4

Question 10:

How do you correctly call the function defined below?

```
def power(x):
    p = 2
    result = x ** p
    return result
```

(A) def power(5)
(B) x = power(4)
(C) y = power(5, 2)
(D) z = power(3, p=2)

Question 11:

Choose the correct function declaration of `my_function` so that we can execute the following function calls successfully:

```
my_function(15, 25, 10)
my_function(10, 20)
```

(A) `def my_function(a, b, c):  ...`
(B) `def my_function(a, b, c=10):  ...`
(C) `def my_function(a=10, b=20, c):  ...`
(D) `def my_function(a, b):  ...`

Question 12:

Select the correct statement about Python functions:

(A) A Python function can return only a single value.
(B) A Python function can take three arguments at most.
(C) A Python function does not always need to have a return statement.
(D) A Python function should have at least one argument.

Question 13:

In Python you want a data-structure that allows you to store the name of some people with their phone numbers. You want to use this data-structure to design a service that allow you to quickly get a name corresponding to a given phone number. Note that you don't care about searching for the phone number corresponding to a given name. Which of the following data-structures would you choose?

(A) A dictionary where the keys correspond to the names and the values correspond to the phone numbers.
(B) A dictionary where the keys correspond to the phone numbers and the values correspond to the names.
(C) A list where each element is another list containing a name and a phone number.
(D) Two lists, one containing the names, and the other one containing the corresponding phone numbers.

Question 14:

Which of the following statements is correct?

(A) If A and B are two numpy arrays, then C = A + B would concatenate them into one array C
(B) If A and B are two lists of floats of the same length, then C = A + B would compute the elementwise sum (i.e., sum of each element in A with the corresponding element in B)
(C) The lists A and B should have the same length to be able to do A + B
(D) If A and B are two numpy arrays, then sum(A) + sum(B) would be the same as sum(A+B)

Question 15:

What would be the output of the following code? (Please read the code carefully)

```
my_result = 2 + 3 * 5.0
print(myResult)
```

(A) 17
(B) 17.0
(C) this will cause an error
(D) 25.0

Question 16:

What is the purpose of the `__init__` method in a Python class?

(A) It is called when an object of the class is created and initializes the object's attributes
(B) It is a special method used for inheritance
(C) It is a destructor method
(D) It is called when an object of the class is destroyed

Question 17:

Suppose that you have two numpy arrays of the same length called A and B. For example:

```
import numpy as np
A = np.array([1, 5, ..., 3])
B = np.array([2, 0, ..., 6])
```

Which one of the following options is **not** equivalent to the following code:

```
s = 0
for i in range(len(A)):
   diff = A[i] - B[i]
   diff2 = diff ** 2
   s = s + diff2
```

```
(A)        s = sum((B - A) ** 2)
(B)        s = sum([ (A[i] - B[i]) ** 2 for i in range(len(B)) ])
(C)        s = sum(A**2 - B**2)
(D)        s = sum([ (a - b)*(a - b) for (a, b) in zip(A, B) ])
```

Question 18:

Which of the following statements is correct?

(A) We usually use the while loop when we know in advance how many iterations we want to perform.
(B) We should use a for loop as much as possible, since it is faster than a while loop.
(C) The while loop stops, when the condition being tested becomes False.
(D) In this example `for nbr in range(n): ...` we should increment the variable `nbr` within the loop, otherwise we can have an infinite loop.

Question 19:

These two Python codes are equivalent. Select True (option A) or False (option B).

Code 1:

```
x = int( input() )
if x%2 == 0 and x < 100:
   print("Access granted")
else:
   print("Access refused")
```

Code 2:

```
x = int( input() )
if x >= 100 or x%2 != 0:
   print("Access refused")
else:
   print("Access granted")
```

(A) True
(B) False

Question 20:

You have a variable called sentence that contains a string entered by the user. You want to check if this string starts with "Hello". If so, you print "Welcome", otherwise you print "Bye". Is the following code a correct way of doing that? Select True (i.e., yes it is correct) or False (i.e., no it is not correct).

```
sentence = input("Please enter a sentence: ")
if sentence[0:5] = "Hello":
   print("Welcome")
else:
   print("Bye")
```

(A) True
(B) False

Question 21:

The following three instructions are all equivalent (i.e., they produce the same result). Select True (option A) or False (option B).

Instruction 1:

```
passed = True if grade > 0.5 else False
```

Instruction 2:

```
passed = (grade > 0.5)
```

Instruction 3:

```
if grade > 0.5:
    passed = True
else:
    passed = False
```

(A) True
(B) False

Question 22:

What would be an equivalent condition to the following:

```
if not (age <= 18 or country != "Sweden"): ...
```

(A) if age > 18 and country == "Sweden": ...
(B) if age > 18 or country == "Sweden": ...
(C) if age <= 18 or country != "Sweden": ...
(D) if age >= 18 and country != "Sweden": ...

Question 23:

In Python, the two names `myAge` and `myage` refer to the same variable.

(A) True
(B) False

Question 24:

What is the output of the following code?

```
class A:
    def show(self):
        print("Sun")

class B(A):
    def show(self):
        print("Moon")

obj = B()
obj.show()
```

(A) Sun (B) Moon (C) Sun Moon (D) Moon Sun

Question 25:

Which one of the following is **not** a correct way to assign value(s) to variable(s) :

```
(A)        salary, month = 38000, "January"
(B)        size == 49
(C)        age = (3 == 2)
(D)        age = height = weight = 55
```

Question 26:

All the following three codes can be used to correctly compute the dot product between u (matrix) and v (vector). True or False?

Code 1:

```
u = np.array([[1, 3], [2, 0], [2, 1]])
v = np.array([1, 5])
result = u @ v
```

Code 2:

```
u = np.array([[1, 3], [2, 0], [2, 1]])
v = np.array([1, 5])
result = np.dot(u, v)
```

Code 3:

```
u = np.array([[1, 3], [2, 0], [2, 1]])
v = np.array([1, 5])
result = sum([u[i]*v[i] for i in range(len(u))])
```

(A) True
(B) False

Question 27:

Product of two matrices: if you compute the following matrix by matrix multiplication in Python, what would be the shape of A ?

$$\text{A} = \begin{bmatrix} 1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix} @ \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

(A) The shape of A would be (2, 3)
(B) The shape of A would be (3, 3)
(C) The shape of A would be (2, 1)
(D) The shape of A would be (3, 1)

Question 28:

You want to create a dictionary with names as keys and ages as values. The dictionary must contain the following information in any order: Helena 21; Mark 18; John 28; Jess 32.

Which code can you use?

Code A:

```
names = ["Helena", "John", "Mark"]
ages = [21, 28, 26]
dico = dict(zip(names, ages))
dico["Jess"] = 32
dico["Mark"] = 18
```

Code B:

```
names = ["Helena", "John", "Mark"]
ages = [21, 28, 26]
dico = dict(zip(ages, names))
dico[32] = "Jess"
dico[18] = "Mark"
```

Code C:

```
names = ["Helena", "Mark", "John"]
ages = [21, 28, 26]
dico = dict(zip(ages, names))
dico["Jess"] = 32
dico["Mark"] = 18
```

Code D:

```
names = ["Helena", "Mark", "John", "Jess"]
ages = [21, 18, 28, 32]
items = zip(names, ages)
dico = [k for k in items]
```

Question 29:

The output of the following code is YES. Select True (option A) or False (option B).

```
x = True
y = False
z = False
if x or (y and z):
     print("YES")
else:
     print("NO")
```

(A) True
(B) False

Question 30:

What is the output of the following code?

```
a = [1, 2, 3, "Hello", [], True]
print(len(a))
```

(A) 5
(B) 1
(C) 3
(D) 6

Question 31:

What is the output of the following code?

```
countries = ["Sweden", "Spain", "Finland", "Denmark"]
i = countries.index("Finland")
countries[i] = "Italy"
print(countries)
```

(A) ["Sweden", "Spain", "Italy", "Denmark"]
(B) ["Sweden", "Spain", "Finland", "Denmark"]
(C) ["Sweden", "Spain", 'Finland", "Denmark", "Italy"]
(D) {"Sweden": 0, "Spain": 1, "Italy": 2, "Denmark": 3}

Question 32:

After running this code: `a` will be 1, and `b` will be 0. Select True (option A) or False (option B).

```
a = 0
b = 1
for i in range(2):
    a, b = b, a
```

(A) True
(B) False

Question 33:

Select the initialization (of the variables x, y, z) that would allow the following code to display `Good`.

```
# Here initialization of x, y, z
# ...
if (x == "YES") or (y >= z):
    print("Bad")
else:
    print("Good")
```

(A) Initialize the variables as follows:

```
x = "YES"
y = 0
z = 1
```

(B) Initialize the variables as follows:

```
x = "YES"
y = 3
z = 1
```

(C) Initialize the variables as follows:

```
x = "NO"
y = 5
z = 2
```

(D) Initialize the variables as follows:

```
x = "NO"
y = 0
z = 1
```

Question 34:

Which of the following is correct if you want to print the sum of two numbers?

Code (A):

```
a = int(input())
b = 2
print("The sum is", a+b)
```

Code (B):

```
a = input()
print("The sum is", a+2)
```

Code (C):

```
a = input(); b = 2
print("The sum is", a+b)
```

Code (D):

```
a = int(input())
b = input()
print("The sum is", a+b)
```

Question 35:

What is the output of the following code?

```
class Morning:
    def greeting(self):
        print("Good Morning!", end=" ")

class Evening(Morning):
    def greeting(self):
        super().greeting()
        print("Good Evening!", end=" ")

time_of_day = Evening()
time_of_day.greeting()
```

(A) Good Morning! (B) Good Evening!
(C) Good Morning! Good Evening! (D) Good Evening! Good Morning!

Question 36:

What is the error in the following code?

```
def factorial(n):
    if n = 0:
        return 1
    else:
        return n * factorial(n - 1)
```

(A) Logic Error: The function must have two arguments (one for `n` and one for `n-1`)
(B) Syntax Error: Single equals sign should be a double equals sign in the if statement
(C) There is no error in this code
(D) Logic Error: Recursive function call is incorrect

Question 37:

What is the error in the following code?

```
def remove_duplicates(lst):
    unique_lst = []
    for item in lst:
        if item not in unique_lst:
            unique_lst.append(item)
    return lst
```

(A) Incorrect list is returned (should return `unique_lst` instead of `lst`)
(B) Syntax Error: Missing quotation marks around item in the if statement
(C) Duplicates are not removed correctly
(D) the lists will always remain empty

Question 38:

What is the error in the code?

```
def find_majority_element(nums):
    majority_count = 0
    for num in nums:
        count = nums.count(num)
        if count > majority_count:
            majority_count = count
            return num
    return None
```

(A) The function may not find the true majority element as it returns a value as soon as the condition is True.
(B) Incorrect comparison of `majority_count` and `count` variables.
(C) The function will always return the last element in the input list, regardless of its count.
(D) Syntax Error: Missing indentations in the for loop

Question 39:

What does the following `mystery_function` do?

```
def mystery_function(word):
    new_word = ""
    for letter in word:
        if letter.isupper():
            new_word += letter.lower()
        elif letter.islower():
            new_word += letter.upper()
        else:
            new_word += letter
    return new_word
```

(A) Returns the given word reversed
(B) Returns the word with alternating uppercase and lowercase letters: upper, lower, upper, etc.
(C) Returns the given word with all uppercase letters converted to lowercase and vice versa
(D) Returns the given word without any changes

Question 40:

```
def merge_lists(list1, list2):
    merged_list = []
    i, j = 0, 0
    while i < len(list1) and j < len(list2):
        if list1[i] < list2[j]:
            merged_list.append(list1[i])
            i += 1
        else:
            merged_list.append(list2[j])
            j += 1
    merged_list += list1[i:]
    merged_list += list2[j:]
    return merged_list
```

Suppose that `list1` and `list2` are already sorted. What does the `merge_lists` function do?
(A) Merges two sorted lists into a new sorted list
(B) Merges two lists into a new list without sorting (elements can end-up in a random order)
(C) Sorts each list in descending order using bubble-sort.
(D) Always returns an empty list

Question 41:

```
class Shape:
    def __init__(self, color):
        self.color = color

class Circle(Shape):
    def __init__(self, color, radius):
        super().__init__(color)
        self.radius = radius

circle = Circle("Red", 5)
```

What is the relationship between the `Circle` class and the `Shape` class?

(A) `Circle` inherits from `Shape`
(B) `Shape` inherits from `Circle`
(C) `Circle` and `Shape` are unrelated classes
(D) `Circle` and `Shape` are instances (objects), not classes

Question 42:

```
class Vehicle:
    def __init__(self, wheels, color):
        self.wheels = wheels
        self.color = color

class Car(Vehicle):
    def __init__(self, color):
        super().__init__(4, color)

class Motorcycle(Vehicle):
    def __init__(self, color):
        super().__init__(2, color)

my_car = Car("Blue")
my_motorcycle = Motorcycle("Red")
```

What are the common attributes between `my_car` and `my_motorcycle` objects?

(A) Only the `color` attribute
(B) Only the `wheels` attribute
(C) Both the `color` and `wheels` attributes
(D) Type of vehicle (Car/Motorcycle)

Question 43:

```
def manipulate_list(lst):
    lst.append(1)
    lst = [0]
    return lst

original_list = [1, 2, 3]
modified_list = manipulate_list(original_list)
print(original_list)
```

What is the final value of `original_list` after the given operations? Note that this question is about `original_list`, not `modified_list`.

(A) [1, 2, 3]
(B) [0, 1]
(C) [1, 2, 3, 1]
(D) [0]

Question 44:

What is the output of the following code?

```
def mystery_func(n):
    if n == 1:
        return 1
    else:
        return n + mystery_func(n - 1)

print( mystery_func(4) )
```

(A) 1
(B) 4
(C) 10
(D) 8

Question 45:

Which of the following list comprehensions creates a list of squares of even numbers from 1 to 20?

```
(A)       [x**2 for x in range(1, 21) if x % 2 == 0]
(B)       [x**2 for x in range(1, 21) if x % 2 != 0]
(C)       [(x % 2)**2 for x in range(1, 21)]
(D)       [x**2 for x in range(1, 21)]
```

Question 46:

What does the `continue` statement do in a loop?

(A) Exits the loop
(B) Skips the rest of the code inside the loop for this iteration and continues with the next iteration
(C) Restarts the loop from scratch
(D) None of the above

Question 47:

```
value = 10

def update_value(value):
    value = value + 5
    return value

updated_value = update_value(20)
print(f"value is {value}, updated_value is {updated_value}")
```

What will be the output of the code?

(A) value is 10, updated_value is 25
(B) value is 10, updated_value is 10
(C) value is 25, updated_value is 25
(D) value is 20, updated_value is 25

Question 48:

```
def func(x):
    x = x + 1

x = 2
func(x)
print(x)
```

(A) 3
(B) 2
(C) 0
(D) None

Second part – Questions 49 to 80 related to DS/ML

Question 49:

What is the main difference between supervised learning and unsupervised learning in machine learning?

(A) Supervised learning requires labeled data, while unsupervised learning does not
(B) Supervised learning algorithms are faster than unsupervised learning algorithms
(C) Unsupervised learning requires labeled data, while supervised learning does not
(D) There is no difference between supervised and unsupervised learning

Question 50:

What is the primary goal of clustering in machine learning?

(A) Maximizing the accuracy of the model
(B) Dividing a set of data points into different groups
(C) Minimizing the number of features in the dataset
(D) Determining the strength of the correlation between features

Question 51:

Select the correct answer about `GridSearchCV` in scikit-learn.

(A) It helps in finding the most complex model, thus reducing overfitting
(B) It fine-tunes hyperparameters, thus reducing overfitting
(C) It increases the training dataset size, thus reducing overfitting
(D) It decreases the number of features, thus reducing underfitting

Question 52:

In this question, you need to select the **wrong** option.

If we treat anomaly detection as a supervised learning problem (with two classes: "normal" vs "abnormal"), we may face the following problems (remember to select the **wrong** option) :

(A) One class will have much more data than the other class, which creates a class-imbalance problem, and may mislead the classifier.
(B) Anomalies may be of various kinds, so data-points from the "abnormal" class may not be close to each others and may be scattered anywhere in the feature space.
(C) It is usually easy to collect examples of normal data, but it is harder to collect various examples of abnormal data.
(D) When we have labels, it becomes very hard to estimate the density around each data-point.

Question 53:

With unsupervised anomaly detection (or outlier detection), we need to fit a model to a training dataset that contains only normal data (i.e., without any anomalies).

(A) True
(B) False

Question 54:

You are interested in classifying patients into three specific classes (healthy, covid19, pneumonia). However, the patients data that you have access to is not labeled (i.e. you don't know the class-label corresponding to each patient). Can you use clustering to achieve the same goal?

(A) Yes, if we use unsupervised clustering to find 3 clusters, they will correspond to the 3 classes we are interested in.
(B) No, the clusters that we find may not necessarily correspond to the three class-labels we are interested in.

Question 55:

You have a regression model that depends on a set of parameters theta. What does it mean "*to train*" such a model?

(A) It means using a training dataset to find the values of the parameters theta that make the predicted output as close as possible to the true output.
(B) It means using a training dataset to find the relation between the different features.
(C) It means finding which training data-points allow us to achieve the smallest cost (or error).
(D) It means finding the values of the parameters theta that maximize the cost (or loss, or error) function.

Question 56:

Select the correct answer. To avoid overfitting in Decision Tree, you can:

(A) Further increase the depth of the tree.
(B) Limit the number of nodes in the tree.
(C) Use the Gini measure, instead of the Entropy to compute the impurity.
(D) Decrease the number of data-points then retrain the decision tree.

Question 57:

You have a training dataset consisting of the input X (matrix of n instances/rows and d features/columns) and the output y (array of n labels). Before training a machine learning model, you want to normalize your training data since you noticed that the features have very different scales. You do that with the following code:

```
from sklearn.preprocessing import MinMaxScaler
import numpy as np

X = ... # input from the training set
y = ... # labels from the training set

sc = MinMaxScaler()
sc.fit(X)
X_scaled = sc.transform(X)

# Then you train a classifier using your normalized training dataset
clf = RandomForestClassifier()
clf.fit(X_scaled, y)
```

Later, you receive new test data consisting of input X_test (matrix of n samples/rows and d features/columns). You want to predict the labels corresponding to the data in X_test. What should you do before making such prediction?

(A) We don't need to do anything. Scaling was only necessary for training. Now, we just need to predict using `clf.predict(X_test)`
(B) Fit the scaler using `sc.fit(X_test)` then scale the data using `sc.transform(X_test)`
(C) Scale the data using `sc.transform(X_test)`
(D) We first need to transform X_test using PCA (Principal Component Analysis).

Question 58:

Suppose that you have text documents where each document is labeled as "politics", "sports", or "entertainment" (i.e., 3 classes). To train a classifier, you need to represent these text documents as feature-vectors. What could be a reasonable way to do this?

(A) Create an array of the same size as the vocabulary (i.e., all possible words). The jth item in this array will be either 0 or 1, indicating whether or not the jth vocabulary word is present in the document.
(B) Represent each document by one feature which is the number of words that the document contains.
(C) It is not possible to classify text documents with usual classification methods. We need advanced NLP (Natural Language Processing) methods for this.
(D) Create an array containing the length (i.e., number of characters) of each word in the document.

Question 59:

You trained a classifier to diagnose some disease in patients. It seems to achieve a low error on the training patients (which is good). However, when you test it on a new set of patients, you find that the error is quite high (which is bad). Which of the following are promising actions that you can do?

(A) Use fewer patients to train the classifier.
(B) Try adding much more features to characterize patients better.
(C) Combine the test set with the training set, re-train the classifier, then test it again on the same test set.
(D) Include more patients in the training set and possibly reduce the number of features.

Question 60:

The reason why we normalize the data (i.e., scale the features) is to eliminate overfitting and underfitting. Is that True or False?

(A) True
(B) False

Question 61:

Which of the following is **less likely** to lead to overfitting?

(A) Having a large number of features compared to the number of training data-points.
(B) Using a small portion of training data.
(C) Using a simple linear model.
(D) Choosing hyper-parameter values that make the model as complex as possible.

Question 62:

Regardless of whether there is overfitting or underfitting or not, the training error is usually lower than the test error.

(A) True
(B) False

Question 63:

We trained a classification model. Then we evaluated it using a training set (to compute the training error err_train) and on a test set (to compute the test error err_test).

Which of the following statements is **wrong**:

(A) if err_train is very low, but err_test is very high, then it's probably overfitting.
(B) if err_train is high and err_test is high, then it's probably underfitting.
(C) if err_train is very high and err_test is very high, then it's probably overfitting.
(D) if err_train is low and err_test is low, then it's probably a good model.

Question 64:

You want to solve a simple problem (e.g., where you know that the target variable depends linearly on the features) with a very complex machine learning model. What can this lead to?

(A) Underfitting
(B) Overfitting
(C) High training error
(D) Low test error

Question 65:

In random forest, making the decision trees as similar as possible to each other (ideally the same) is better than making them diverse. True or False?

(A) True
(B) False

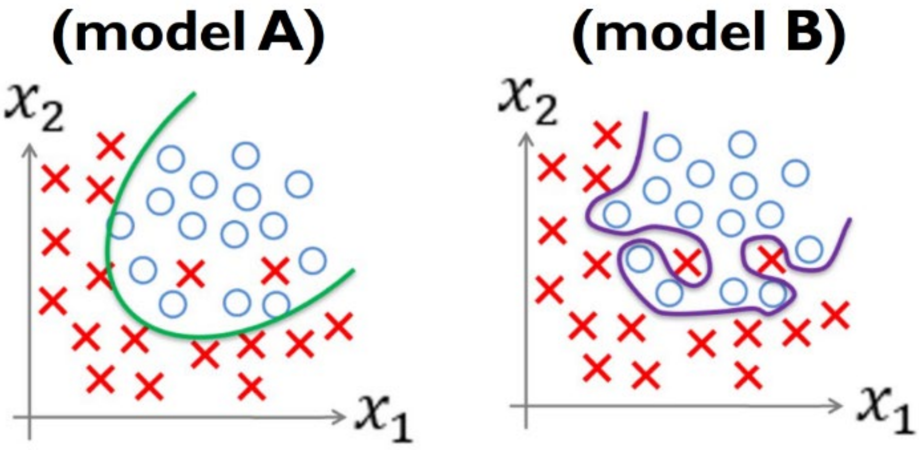
Question 66:

Which of the following is **not** a recommended action to avoid overfitting?

(A) Adding much more features to the dataset
(B) Fine-tuning the hyperparameters of the model
(C) Collecting much more data-points
(D) Using an ensemble of models (e.g., several decision trees instead of one)

Question 67:

Choose the correct answer about the models below:



(A) Model B achieves 0 errors on the training dataset, so it is better than model A
(B) Model A suffers from severe underfitting.
(C) Model A is likely more accurate than model B, on new data.
(D) Both models A and B are good models.

Question 68:

Consider the following input dataset:

```
X = np.array([
    [1,  20],
    [6,  5],
    [3,  50],
    [11, 25]
])
```

What would be the values of the first feature (i.e., column 0 of X) if the features were normalized using the min-max scaler?

(A) 0, 0.5, 0.2, 1
(B) 0.01, 0.06, 0.03, 0.11
(C) 0.1, 0.6, 0.3, 1.1
(D) We can't normalize this dataset since the labels y are missing in this question.

Question 69:

You have a set of five houses characterized by their sizes (in sqm) and their number of rooms. The first house has size 30 and 1 room. The second house has size 65 and 2 rooms. The third house has size 110 and 3 rooms. The fourth house has size 25 and 1 room. The fifth house has size 80 and 2 rooms.

Predict the number of rooms in a new house of 90 sqm using the k-nearest-neighbours algorithm with k=1.

(A) The predicted number of rooms is 3
(B) The predicted number of rooms is (1+2+3+1+2) / 5
(C) The predicted number of rooms is 2
(D) We can never use regression to predict the number of rooms. We should rather predict a target variable such as the price.

Question 70:

You have a very large collection of unlabeled news articles from the Web. You want to automatically group together the news articles that talk about similar topics. What kind of machine learning methods would you use?

(A) Regression
(B) Classification
(C) Clustering
(D) Anomaly Detection

Question 71:

What's the main difference between regression and classification?

(A) In regression the features take continuous values, whereas in classification the features are mainly categorical features (such gender, education level, etc.)
(B) We can use regression and classification interchangeably depending on the dataset size (number of data-points) and the number of features
(C) In regression, the output is a linear combination of the input features. In classification, the output class is predicted based on a nonlinear combination of the input features.
(D) In regression, the target variable is continuous values, whereas in classification the target variable consists of classes (or categories).

Question 72:

When is feature normalization (scaling) important?

(A) When the number of features in the training dataset is different from the number of features in the test dataset.
(B) When the features have very different range of values. For example, the values of feature 1 are in the range [0, 1], but the values of feature 2 are in [50, 100].
(C) Every time we train a complex machine learning model.
(D) When we have multiple machine learning models.

Question 73:

Complete the following sentence about clustering:

The choice of K (the number of clusters) in K-means ...

(A) is random, the algorithm will automatically choose it.
(B) should always be either K=2 or K=3.
(C) should always be as large as possible.
(D) is selected by the user beforehand depending on the dataset and the prior knowledge about the application domain.

Question 74:

You are a data analyst at Google, and you want to predict how many visitors (people) will look at two Wikipedia pages by tomorrow (e.g., "Galaxy - Wikipedia" and "Star - Wikipedia"). You have data about how many visitors visited these two pages in the past. Which of the following statements is correct?

(A) I can use two regression models (one for each page). The target variable will be the number of visitor for each page.
(B) I can use a classification model to predict the number of visitors.
(C) I can use a regression model. The target variable will be the number of words in each page.
(D) I can use two classification models (one for each page). the target variable will be the number of words in each page.

Question 75:

Select the correct statement about clustering.

(A) Clustering can always be used for classification (when we don't have labels).
(B) If we run K-Means several times on the same dataset using the same value of K, then the clustering result is guaranteed to always be the same.
(C) Clustering is grouping the data points such that instances (data-points) from different clusters are similar (close to each others)
(D) A goal of clustering is to identify groups (clusters) such that instances (data-points) within the same cluster have similar properties.

Question 76:

Select the **wrong** statement about overfitting.

(A) Overfitting may happen when the number of data-points in the training dataset is small.
(B) Overfitting can be addressed using Random Forest instead of Decision Tree.
(C) Overfitting can be addressed by increasing the number of data-points in the test dataset and decreasing the number of data-points in the training set.
(D) Overfitting may happen when the model is too complex.

Question 77:

How can we measure the impurity in Decision Tree?

(A) Using Density
(B) Using Entropy
(C) Using the Euclidean Distance
(D) Using the Mean Squared Error

Question 78:

In classification, using kNN (k-Nearest-Neighbors) with k=1 produces a simpler decision boundary than k=10. True or False?

(A) True
(B) False

Question 79:

Which of the following statements is **wrong**.

(A) After training a Linear Regression model, the estimated parameters can be used to make accurate predictions on any other dataset, even if it's from a completely different application domain.
(B) Binary classification can be used for a classification problem with 4 classes. This is done by using the one-vs-rest strategy and training 4 binary classification models.
(C) k-Nearest-Neighbors can be used to solve nonlinear regression problems.
(D) Logistic Regression is a classification method.

Question 80:

What is the predicted output of a binary logistic regression model?

(A) Any number (from -infinity to +infinity)
(B) A number between 0 and 1 (representing a probability)
(C) A number having the same range as the range of the features.
(D) A number which is between 1 and the number of data-points.