esprit Se former autrement	**EXAMEN** Semestre : 1 ☐ 2 ■ Session : Principale ■ Rattrapage ☐
Unité d'enseignement : Valorisation des données massives Module (s): Big data analytics & Deep learning Classe(s) : 4DS Nombre des questions : 37 Nombre de pages : 8 Date : 01/07/2020 Heure 16h15 Durée : 1h00.	

Partie I: Big Data Analytics

1. Laquelle des commandes suivantes est correcte si vous souhaitez vérifier les contenus d'un fichier texte stocké sur HDFS sans premièrement le copier manuellement dans le stockage local ?

- **A.** C'est impossible de le faire
- **B.** Hdfs dfs -vi /user/Hadoop/myFile.txt
- **C.** Hdfs dfs -nano /user/Hadoop/myFile.txt
- **D.** Hdfs dfs -cat /user/Hadoop/myFile.txt

2. Quel type de contrainte peut-on ajouter à une table Hive ?

- **A.** Clé primaire
- **B.** Clé étrangère
- **C.** Clé unique
- **D.** Aucune des clés précédentes

3. Qu'est ce qui se passe à la suite de la suppression d'une base de données Hive ?

- **A.** Les tables seront aussi supprimées
- **B.** Le répertoire dédié à la base sera supprimé s'il n'y a pas de table
- **C.** Les blocks HDFS seront formatés
- **D.** Aucune des réponses précédentes

4. Etant donné la tâche suivante : vous souhaitez importer une table depuis une base de données MySQL appelée « retail_db » dans Hive. Qu'est-ce qu'il manque dans la commande d'importation Sqoop ci-dessous à la ligne 5 ?

```
1  sqoop import \
2  --connect jdbc:mysql://localhost:3306/retail_db \
3  --username=root \
4  --hive-import \
5  ****
6  --m=1
```

- **A.** Une requête de type formulaire libre en utilisant une clause WHERE
- **B.** Le nom de la table que vous souhaitez importer
- **C.** Des informations de partitionnement
- **D.** L'emplacement du nœud HDFS.

5. Quel est l'effet de la ligne 7 dans la commande d'importation Sqoop suivante ?

```
1  sqoop import-all-tables \
2  --connect jdbc:mysql://localhost:3306/retail_db \
3  --username=root \
4  --compression-codec=snappy \
5  --as-avrodatafile \
6  --warehouse-dir=/user/hive/warehouse \
7  -m 1
```

A. Elle spécifie le nombre de partitions à importer

B. Elle spécifie le nombre de fichiers résultants sur HDFS

C. Elle spécifie le nombre de tâches de mappage à utiliser pour l'importation en parallèle

D. Elle est spécifique uniquement à l'exportation Sqoop et ignorée pour l'importation Sqoop

6. Laquelle des réponses suivantes n'est pas un type valide de sink pour Flume ?

A. HDFS sink

B. HBASE sink

C. HTTP sink

D. JDBC sink

7. Etant donné le fichier de configuration Flume suivant, quel est le but de l'agent Flume résultant ?

```
# Name the source, channel and sink
flume_importer.sources = avro-source
flume_importer.channels = jdbc-channel
flume_importer.sinks = file-sink

# Source configuration
flume_importer.sources.avro-source.type = avro
flume_importer.sources.avro-source.port = 11112
flume_importer.sources.avro-source.bind = localhost

# Describe the sink
flume_importer.sinks.file-sink.type = hdfs
flume_importer.sinks.file-sink.hdfs.path = /user/hadoop/sink
flume_importer.sinks.file-sink.hdfs.fileType = DataStream
flume_importer.sinks.file-sink.hdfs.fileSuffix = .avro
flume_importer.sinks.file-sink.serializer = avro_event
flume_importer.sinks.file-sink.serializer.compressionCodec=snappy

# Describe the type of channel
flume_importer.channels.jdbc-channel.type = jdbc

# Bind the source and sink to the channel
flume_importer.sources.avro-source.channels = jdbc-channel
flume_importer.sinks.file-sink.channel = jdbc-channel
```

A. Il diffuse les données encodées avro depuis une source avro vers HDFS

B. Il diffuse les données encodées avro depuis HDFS vers une source avro

C. Il diffuse les données encodées avro depuis kafka vers HDFS

D. Il diffuse les données encodées snappy depuis une source avro vers HDFS

8. Laquelle des réponses suivantes est considérée comme le principal avantage de spark par rapport à Mapreduce ?

- **A.** Spark supporte scala
- **B.** Spark peut exécuter des calculs en mémoire
- **C.** Spark supporte les RDD
- **D.** MapReduce supporte le calcul distribué

9. Lequel des langages suivants n'est pas supporté par Apache spark

- **A.** Java
- **B.** Scala
- **C.** Python
- **D.** Ruby

10. Lequel des composants suivants ne fait pas partie de l'écosystème de Spark ?

- **A.** Spark core
- **B.** Spark Name-Node
- **C.** Spark GraphX
- **D.** Spark MLlib

11. Laquelle des réponses suivantes est l'une des responsabilités de Spark Driver ?

- **A.** Il est un nœud travailleur (worker node) responsable des processus individuels d'un job de spark.
- **B.** Il stocke les métadonnées sur le job en cours mais il ne contient pas les données réelles.
- **C.** Il est responsable de l'allocation des ressources aux job lancés.
- **D.** Il stocke les résultats des calculs en mémoire, cache ou sur le disque.

12. Laquelle des réponses suivantes n'est pas une caractéristique valide d'un RDD ?

- **A.** L'évaluation paresseuse
- **B.** La tolérance aux pannes
- **C.** L'immuabilité
- **D.** L'évolutivité

13. Etant donné la commande suivante exécutée avec pyspark, quel est le type de données résultant ?

```
sqlContext = SQLContext(sc)
data = sqlContext.read.json("data_cours/sparksql/employee.json")
```

- **A.** Un objet FileInputStream
- **B.** Un dataframe ou chaque élément est une ligne depuis le fichier
- **C.** Un pair RDD ou la clé est le chemin de fichier et la valeur du contenu du fichier
- **D.** Un RDD ou chaque élément est une ligne depuis le fichier.

14. Laquelle des réponses ci-dessous n'est pas une opération valide supportée par un RDD ?

- **A.** Regroup()
- **B.** Count()
- **C.** parallelize()
- **D.** filter()

15. Etant donné le fragment de code suivant, quel est le résultat du programme ?

```
from pyspark import SparkContext
sc =SparkContext.getOrCreate()
RDD = sc.parallelize([0,1,2,3,4,5,6,7,8,9,10])
RDD1 = RDD.map(lambda x : x+2)
```

A. Le code ne sera pas compilé

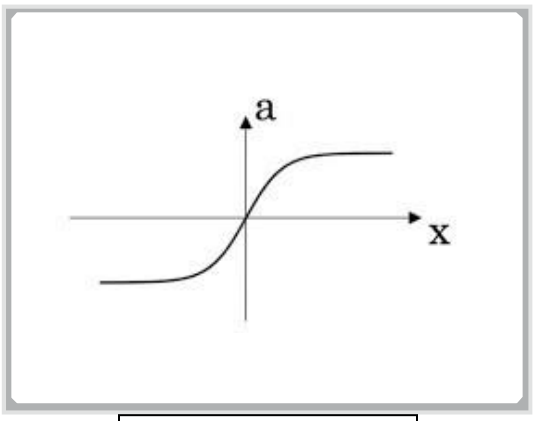
B. [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

C. Le code affichera les numéros dans un ordre non déterminé

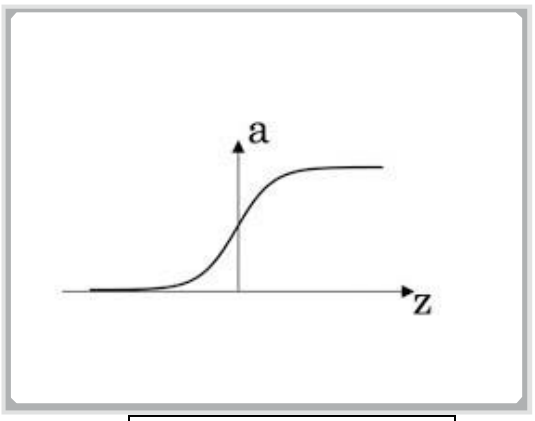
D. Le code sera compilé mais il n'affichera rien

Partie II : Deep Learning

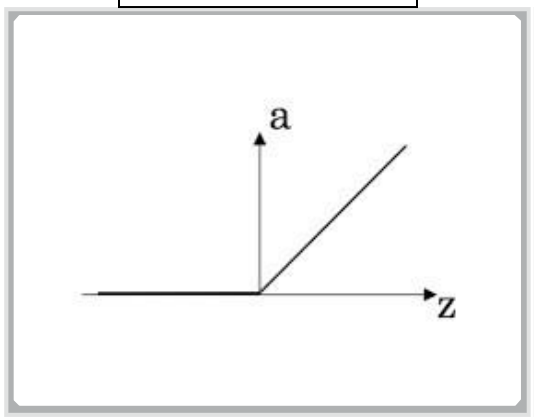
16. Which one of these plots represents a RELU activation function?



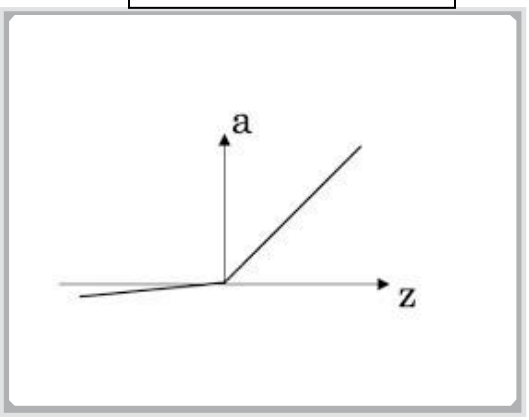
A. Figure n° 1



B. Figure n° 2



C. Figure n° 3



D. Figure n° 4

17. What does a neuron compute?

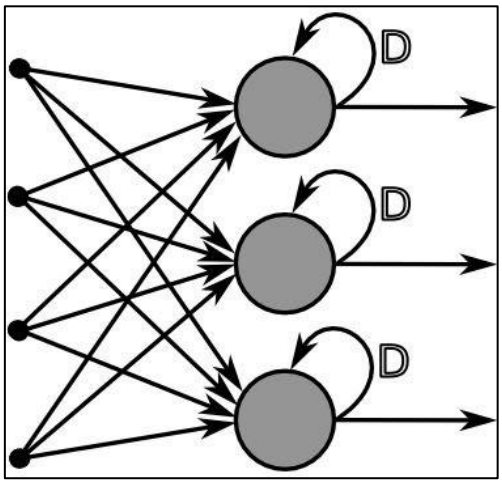
A. A neuron computes an activation function followed by a linear function (z= Wx+b)

B. A neuron computes a function g that scales the input x linearly (Wx+b)

C. A neuron computes the mean of all features before applying the output to an activation function

D. A neuron computes a linear function (z= Wx+b) followed by an activation function

18. You are building a neural network where it gets input from the previous layer as well as from itself. Which of the following architecture has feedback connections?



A. Recurrent Neural network

B. Convolutional Neural Network

C. Restricted Boltzmann Machine

D. None of these

19. Batch Normalization is helpful because?

A. It normalizes (changes) all the input before sending it to the next layer

B. It returns back the normalized mean and standard deviation of weights

C. It is a very efficient backpropagation technique

D. None of these

20. Suppose img is a (32,32,3) array, representing a 32x32 image with 3 color channels red, green and blue. How do you reshape this into a column vector?

A. X = img.reshape((3,32*32))

B. X = img.reshape((1,32*32,*3))

C. X = img.reshape((32*32*3,1))

D. X = img.reshape((32*32,3))

21. In a neural network, which of the following techniques is used to deal with overfitting?

A. Dropout

B. Regularization

C. Batch Normalization

D. All of these

22. You are building a binary classifier for recognizing cucumbers (y=1) vs. watermelons (y=0). Which one of these activation functions would you recommend using for the output layer?

A. Relu

B. Leaky Relu

C. Sigmoid

D. Tanh

23. Among the following, which one is not an " hyperparameter"?

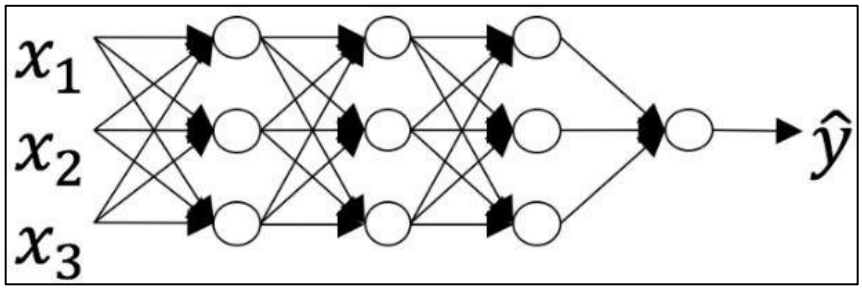
A. Learning rate α

B. Number of iterations

C. Weight matrices W

D. Number of layers L in the neural network

24. Consider the following neural network, how many layers does this network have?



A. The number of layers L is 3. The number of hidden layers is 3.

B. The number of layers L is 4. The number of hidden layers is 3.

C. The number of layers L is 4. The number of hidden layers is 4

D. The number of layers L is 5. The number of hidden layers is 4.

25. In a neural network, knowing the weight and bias of each neuron is the most important step. If you can somehow get the correct value of weight and bias for each neuron, you can approximate any function. What would be the best way to approach this?

A. Assign random values and pray to God they are correct.

B. Search every possible combination of weights and biases till you get the best value.

C. Iteratively check that after assigning a value how far you are from the best values, and slightly change the assigned values to make them better

D. None of these

26. What are the steps for using a gradient descent algorithm?

1. Calculate error between the actual value and the predicted value
2. Reiterate until you find the best weights of network
3. Pass an input through the network and get values from output layer
4. Initialize random weight and bias
5. Go to each neuron which contributes to the error and change its respective values to reduce the error

A. 1, 2, 3, 4, 5

B. 5, 4, 3, 2, 1

C. 3, 2, 1, 5, 4

D. 4, 3, 1, 5, 2

27. In a neural network, knowing the weight and bias of each neuron is the most important step. If you can somehow get the correct value of weight and bias for each neuron, you can approximate any function. What would be the best way to approach this?

A. Assign random values and pray to God they are correct.

B. Search every possible combination of weights and biases till you get the best value.

C. Iteratively check that after assigning a value how far you are from the best values, and slightly change the assigned values to make them better

D. None of these

28. Which of the following gives non-linearity to a neural network?

A. Stochastic Gradient Descent

B. Rectified Linear Unit

C. Convolution function

D. None of the above

29. What is autoencoder?

A. A neural network that codes itself

B. A neural network that decodes itself

C. An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner.

D. An autoencoder is a type of artificial neural network used to learn efficient data codings in a supervised manner.

30. Which of the following is an application of autoencoders?

A. Feature learning

B. Dimensionality reduction

C. Information retrieval

D. All the above

31. Consider the following statements regarding Artificial Neural Networks (ANN) and convolutional Neural Networks (CNN):

1. There are sparse connections between inputs and outputs between two consecutive layers in CNN.
2. CNNs can be used only for image data
3. Parameters are shared between output neurons in CNN layer
4. Both CNNs and ANNs can take image data as input

Which of the above statements are TRUE?

A. 1 and 2

B. 1 and 4

C. 1, 3 and 4

D. 2,3 and 4

32. What will be the size of the output after the following operations?

Input size = [227 x 227 x 3]
Filter size = [11 x 11 x 3]
Stride = 4
2 x2 Max-pooling with stride of 2

A. [54 x 54]
C. [216 x 2016]
B. [55 x 55]
D. [68 x 68]

33. Pooling layers are used to accomplish which of the following?

A. To progressively reduce the spatial size of the representation.
B. To reduce the number of epochs while ensuring best accuracy.
C. To ovoid local minima.
D. To always select maximum value over pooling region.

Answer 34-36 for the CNN given below:

The whole network is composed of Conv layers that perform 3 x 3 convolutions with stride 2 and padding is 'valid' (P=1). POOL layers perform 2 x 2 max pooling with stride 2 (and no padding). Number of filters in the Conv layers and number of neurons in fully connected layers are shown in brackets.

34. The output size after pool1, pool2 are:

A. [32 x 32 x 128], [5 x 5 x 64]
B. [31 x 31 x 128], [3 x 3 x 64]
C. [32 x 32 x 128], [3 x 3 x 64]
D. [31 x 31 x 128], [5 x 5 x 64]

35. Number of parameters till pool2 including bias are:

A. 89186
C. 75648
B. 297042
D. 147584

36. Total Number of parameters from pool2 layer till the output layer including bias are

A. 855818
C. 590848
B. 262400
D. 2570

37. Which of the following is true for most CNN architectures?

A. Size of input (height and width) decreases, while depth increases
B. Fully connected layers in the first few layers
C. Multiple convolutional layers followed by pooling layers
D. A and C