جامعة الإسكندرية
ALEXANDRIA
UNIVERSITY
كلية الحاسبات وعلوم البيانات

DST Revision Before midterm

1. **What does the function “search” do?**

   A. matches a pattern at the start of the string.
   [✓ B. matches a pattern at any position in the string.]
   C. replace all matched.
   D. delete all matched.

2. **Which of those functions isn’t related in the Requests module?**

   A. requests.post(...)
   B. requests.put(...)
   C. requests.delete(..)
   [✓ D. requests.copy(…)]

3. **What is the typical file extension for a Comma Separated Values (CSV) file?**

   a. .txt
   [✓ b. .csv]
   c. .py
   d. .xls

4. **Which of the following is a valid data type in JSON?**

   a. Date
   b. Currency
   [✓ c. List]
   d. Font

5. **What type of XML tag is represented as <tagname />?**

   a. StartTag
   b. End Tag
   [✓ c. Empty-Element Tag]
   d. Attribute

6. **Which method from the json library is used to load JSON from a file in Python?**

   [✓ a. Json.load(file)]
   b. json.loads(file)
   c. json.read(file)
   d. json.reads(file)

7. In MongoDB, what does BASE stand for in terms of consistency?

A. Basic Availability, Simple-state, Eventually Consistent

B. Basically Available, Soft-state, Eventually Consistent

C. Basic, Availability, Simple-state, Elastic

D. Basically Atomic, Stable-state, Eventual Consistency

8. What is BSON in MongoDB?

A. Basic Structured Object Notation

B. Binary Structured Object Node

C. Binary JSON

D. BSON Object Naming

9. Which type of NoSQL database is MongoDB classified as?

A. Column Store

B. Key-Value Store

C. Document Store

D. Graph Database

10.What is the purpose of the ObjectId in MongoDB?

A. To identify databases.

B. To ensure document uniqueness.

C. To create timestamps.

11. Which of the following is true about MongoDB's handling of the _id field?

A. It cannot be indexed.

B. It is automatically created as an integer.

C. Developers cannot provide their own values.

D. It is automatically indexed and can be an Object Id or another unique immutable value.

12.What is a major advantage of MongoDB's document model?

A. Reduced scalability

B. Complex querying

C. Flexibility in schema design

D. Strict adherence to ACID principles to manage distributed transactions

13. **What does the term "Horizontally Scalable" mean in the context of MongoDB?**

A. Scaling vertically across multiple servers

B. Scaling a single server to handle more requests.

C. Distributing data across multiple servers to handle growth.

D. Limiting the number of servers for better performance.

MONGO SECTION

1. **What is MongoDB?**

A. Relational database

B. Document-oriented database

C. NoSQL database

D. Both B and C

2. **In MongoDB, what is a document equivalent to in a SQL database?**

A. Table

B. Record

C. Field

D. Column

3. **Which method is used to insert a single document into a MongoDB collection using PyMongo?**

A. add_one()

B. insert_single()

C. insert_one()

D. add_document()

4. **What is the purpose of the PyMongo package in Python with respect to MongoDB?**

A. Web development

B. Data visualization

C. MongoDB driver for Python

D. Machine learning

5. **In MongoDB, what does CRUD standfor?**

   A. Create, Retrieve, Update, Delete

   B. Connect, Read, Update, Delete

   C. Collect, Retrieve, Use, Delete

   D. Create, Read, Upload, Delete

6. **How do you update a document in MongoDB using PyMongo?**

   A. update_single()

   B. modify_one()

   C. update_one()

   D. change_document()

7. **In PyMongo, what does the $set operator do in the context of updating a document?**

   A. Sets the document to null.

   B. Adds a new field to the document.

   C. Updates a specific field in the document

   D. Sorts the document in ascending order.

8. **Which method is used to delete a single document from a MongoDB collection in PyMongo?**

   A. delete_one()

   B. remove_single()

   C. erase_one()

   D. discard_one()

9. **What is the purpose of the sort () method in MongoDB when using PyMongo?**

   A. Group documents in a collection

   B. Filter documents based on a condition.

   C. Order the result in ascending or descending order.

   D. Limit the number of documents returned.

Data Preprocess

1. **Question: Why do we use data normalization in machine learning?**
   A. To make data more complicated.
   B. To reduce the impact of unusual values.
   C. To add noise to the data.
   D. To create subsets of data.
2. **Question: What is the primary purpose of the Euclidean distance metric in machine learning?**
   A. To find the median value.
   B. To measure the straight-line distance.
   C. To calculate the mean value.
   D. To identify outliers in data.
3. **Question: Why is Hamming distance commonly used for comparing binary vectors?**
   A. It is the fastest metric.
   B. It considers statistical measures.
   C. It is a generalization of various metrics.
   D. It is simple and straightforward.
4. **Question: When might noise be introduced in data processing?**
   A. During data normalization
   B. When dealing with inconsistent data.
   C. In the presence of outliers
   D. During vector space representation
5. **Which of the following normalization techniques is based on the range of values and ensures data is scaled between 0 and 1?**
   A. Z-score normalization
   B. Decimal Scaling normalization
   C. Mean normalization
   D. Min-Max normalization
6. **Question: Which of the following are common techniques to replace missing data in a dataset?**
   A. Mean
   B. Median
   C. Mode
   D. Random Value

7. **What is the S-Jaccard similarity between sets x={a,c,d} and y={a,b,e}?**

   a. 0.2

   b. 0.25

   c. 0.4

   d. 0.5

8. **What is the S-edit distance between the strings "Samar" and "Tamer"?**

   a. 1

   b. 2

   c. 3

   d. 4

9. **What is the Euclidean distance between the points (1, 2, 3) and (4, 5, 6)**

   a. 3.0

   b. 4.0

   c. 5.0

   d. 5.19

10. **What is the cosine similarity between the vectors (1, 2) and (2, 3)?**

   a. 0.5

   b. 0.007

   c. 0.707

   d. 1.0

Feature Extraction in Images and Time Series

1. **Question: What does the term "standardization" refer to in the context of image preprocessing?**
   a. Increasing image diversity
   b. Scaling and preprocessing images to have similar characteristics.
   c. Reducing image resolution
   d. Converting images to grayscale

2. **Question: Which image augmentation technique involves reversing rows or columns of pixels either vertically or horizontally?**
   a. Rotation
   b. Shifting
   c. Flipping
   d. Scaling

3. **Question: What is the purpose of changing image brightness during data augmentation?**
   a. To reduce image diversity
   b. To introduce noise in the dataset
   c. To increase contrast
   d. To improve model performance

4. **Question: In the formula for grayscale conversion (gray_image = 0.3 * R + 0.59 * G + 0.11 * B), what do R, G, and B represent?**
   a. Red, Green, and blue channels of the image
   b. Random values for pixel intensities
   c. Rotation, Grayscale, and Brightness parameters
   d. RGB color space

5. **What is the primary purpose of aggregating transactional data into time series data?**
   a. To increase computational complexity
   b. To improve data security
   c. To facilitate analysis over time
   d. To reduce data diversity

6. Which of the following is an example of time series data?

a) Image pixel values
b) Customer names
c) Website visits per hour
d) Employee salaries

7. What does the term "seasonality" refer to in the context of time series data?

a) The time interval of data collection
b) The cyclic pattern within a time series
c) The size of the dataset

8. How is time series accumulation different from other forms of aggregation?

a) It focuses on reducing computational complexity.
b) It involves hierarchical structure aggregation.
c) It is specifically designed for financial data.
d) It is done at irregular time interval.

9. What is an important modelling decision related to time series data?

a) The choice of data colours
b) The choice of frequency (time interval)
c) The choice of data encryption method
d) The choice of hardware for data storage
e) The diversity of data sources

10. What is the purpose of saving BTC data to a CSV file in the code example?

a) To reduce the size of the dataset
b) To improve data security
c) To avoid repeatedly pulling data
d) To create a backup of the data

11. In the context of time series data, what does the term "frequency" refer to?

a) The pitch of a cyclic pattern
b) The time interval of data collection
c) The size of the dataset
d) The diversity of data sources

Lambda

1. Executing (lambda x, y: x-y) (2, 3) in Python produces

a) 0

[✓ b) -1]

c) 5

d) 6

2. Executing print (map (lambda x: x*x, [0,1,2])) in Python produces

a) [0,0,0]

b) [0,1,2]

[✓ c) [0,1,4]]

d) 3

3. Executing print (filter (lambda x: x > 2 and x < 8, [-1,0,5,3])) in Python produces

[✓ a) [5,3]]

b) [3,5]

c) [-1,0,5,3]

d) 8

4. Executing print (reduce (lambda x, y: x*y, [1,2,3,4])) in Python produces

[✓ a) 24]

b) 2

c) [1,2,3,4]

d) 7

5. Executing print (reduce (lambda x, y: x + y, [1,2,3,4])) in Python produces

a) 24

b) 3

c) [1,2,3,4]

[✓ d) 10]

6. The output of the following code is

```
 dist = get_metric('euclidean')
 X = [[2, 2]]
 Y = [[2, 2]]
dist. pairwise(X, Y)
```

a) 0 b) 4 c)2 d) 1

7. Let x = {a, b, d} and y = {b, c} then Sjaccord. (x, y) is ---------

a) 0.2 b) 0.25 c) 1 d) 0.5

8. If x = [0 1 0 1] and y = [1 0 1 0] then dhamming (x, y) is ---------- while the squared ecludian (d_2(x,y))2 is -------

a) 4/4
b) 2/4
c) 4/2
d) 2/2

9. Which of the following is not an example for NOSQL database

a) MongoDb
b) Neo4j
c) Cassandra
d) SQLite

10. Which module in Python supports regular expressions?

a) String b) regex c) pyregex d) sklearn

11. What does the function "match" in the regular expressions package do?

a) matches a pattern at the start of the string.
b) matches a pattern at any position in the string.
c) such a function does not exist.
d) none of the mentioned

12. Which module in Python supports XML?

a) BeautifulSoup
b) numpy
c) pyxmlex
d) xmlrequest

13. Answer the following question regarding the state after the execution of the following code,

```
df = DataFrame ([(1, 'Kolter', 'Zico'), (2, 'Manek', 'Gaurav'), (3, 'Rice',
'Leslie')],
columns = ["Person ID", "Last Name", "First Name"])
df. drop (1, inplace=True, axis=1)
```

how many records will be in the dataframe df

a) 6 b) 3 c) 9 d) 2

14. To normalize the following dataset using three different techniques
10,40,50,10,50, 70,90,30

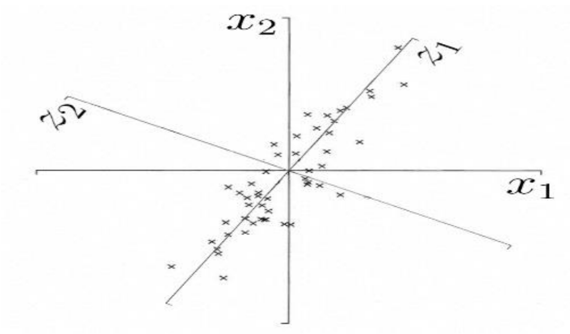
b) subtract each value from the standard deviation of the data.
c) A followed by B
d) Divide each value by the median.
e) Replace each value (x by x-mean of the data)/standard deviation.

15. which of the following is true regarding the filter model

a) Separating feature selection from classifier learning
b) Relying on general characteristics of data (information, distance, dependence, consistency)
c) No bias toward any learning algorithm
d) Fast

16. Which axis you may select reduce the dimension

a) z1
b) z2
c) z1 or z2
d) z1 and z2



17. Given the following term frequencies in a corpus D that contains 3 documents D1...D3, answer the following questions: -

Document 1 (D1)		Document 2 (D2)		Document 3 (D3)	
Term	Term count	Term	Term count	Term	Term count
Caw	2	Sudan	3	Egypt	2
Sudan	1	Caw	2	Nile	2
Camel	1	Nile	1	Cow	1

Answer several questions regarding the normalized tf, DF, IDF, tflogidf, binary representation, data matrix distance matrix

18. What the output of the following code:

```
iris = datasets. load_iris()
df = pd.DataFrame(iris['data'], columns = iris['feature_names'])
scalar = StandardScaler()
scaled_data = pd.DataFrame(scalar.fit_transform(df))
pca = PCA(n_components = 2)
pca.fit(scaled_data)
data_pca = pca.transform(scaled_data)
```

you should be able to answer any questions about the code in handouts such as the above code.

Additional resources:

- 10 questions on Data Science Basics Data Science Basics Questions and Answers - Sanfoundry

- 10 questions on Lambda Python Questions for Campus Interviews - Sanfoundry

- REFERENCE FOR FIRST 20 QUESTION FROM LEC 1 and LEC 2 Theories

1. (Database SQL Only Included(1,2,3,4,5,6,7,8,9,10)) : SQL Queries - Database Questions & Answers - Sanfoundry
2. (MongoDB, Only Included (1,2,3,8,9,10)) NoSQL Databases - MongoDB Questions and Answers - Sanfoundry

Midterm

Program: General/Intelligent Systems/Cybersecurity
Level: Third **Term: Fall 2023/2024**
Course Code: 02-24-01203 **Course Title: Data Science Tools &Software**
Time Allowed: 60minutes **Total points: 20**
Professor name: Dr. Mohamed Abd El-Hafeez

جامعة الإسكندرية
ALEXANDRIA
UNIVERSITY
كلية الحاسبات وعلوم البيانات

Attempt ALL the following 53 questions

Choose the MOST APPROPRIATE answer for the following statements.

You may choose E (=ALL) if all answers (A, B, C and D) are correct or choose F (=NONE) if none of the answers fits.

Please write your answers on the ANSWER SHEET ONLY

In the designated answer sheet, mark your choice (ⓐ, ⓑ, ⓒ, ⓓ, ⓔ, or ⓕ) in front of the question number.

Be sure that you have filled the appropriate bubbles carefully as in the example below.

Example: if the choice for question 300 is "C" then your answer sheet should look like this:

300. ⓐ ⓑ ● ⓓ ⓔ ⓕ

1. Which of the following is not true regarding Data Science?
   a) Concerned only with big data
   b) Heavy focus on machine learning algorithms
   c) Concerned only with small data
   d) Concerned with theories in statistics
2. Which module in Python supports regular expressions?
   a) String **b) re** c) pyregex d) sklearn
3. What does the function "search" in the regular expressions package do?
   a) matches a pattern at the start of the string
   b) matches a pattern at any position in the string
   c) replace all matched
   d) delete all matched
4. Which of the following HTTP methods never modifies a server's state?
   a) response = requests.put(...) b) response = requests.post(...)
   c) response = requests.delete(...) d) **response = requests.get(...)**
5. Which module in Python supports parsing HTML and XML documents?
   a) BeautifulSoup b) numpy c) pandas d) sklearn
6. What is the library that corresponds to the alias "ps" in the following code

```
df = ps.DataFrame([(1, 'Kolter', 'Zico')])
```

   a) pandas
   b) panorama
   c) pymatplots
   d) scipy

Answer the following two questions regarding the state after the execution of the following code:-

```
df = DataFrame([(1, 'Kolter', 'Zico'),(2, 'Manek', 'Gaurav'), (3, 'Rice', 'Leslie')],
columns=["Person ID", "Last Name", "First Name"])
df.drop(1, inplace=True, axis=0)
df.drop(2, inplace=True, axis=1)
```

7. how many records (rows) will be in the dataframe df, after executing the above code?
   a) 1 b) 3 **c) 2** d) 4
8. how many columns the dataframe df will have, after executing the above code?
   a) 6 **b) 2** c) 3 d) 1

9. Which of the following is not an example of unordered data?
   a) Employee records b) Documents c) Bank transactions **d)Time Series**

10. What is the primary purpose of the Request module?

a) **Send HTTP requests to a server and retrieve web page content**
b) Manage database connections for data storage
c) Execute complex algorithms for data analysis
d) Control graphical user interface interactions

11. What will be the output of the following Python code?

```
CarName = 'Porche'
WordName = 'World'
print('{0} is the fastest car in the {2}'.format(CarName, WordName))
```

a) **Porche is the fastest car in the World**
b) Porche is the fastest car in the
c) Porche is the fastest car in the 2
d) IndexError: tuple index out of range

12. What does the term "ACID" stand for in the context of databases?

a) All-Comprehensive Isolation and Durability
b) **Atomicity, Consistency, Isolation, Durability**
c) Advanced Configuration for Isolated Databases
d) Association of Concurrent Information and Data

13. How is the _id field automatically created if not provided in MongoDB?

a) Integer b) Timestamp **c) ObjectId** d) AutoID

14. What is MongoDB?

a) Relational database b) Document-oriented database c) NoSQL database d) **Both B and C**

15. In MongoDB, what is a document equivalent to in a SQL database?

a) Table b) **Record** c) Field d) Column

16. Which method is used to find documents in a MongoDB collection based on a specific condition?

a) get_one() b) search() **c) find_one()** d) query_one()

17. The hamming distance between two binary vectors is equivalent to :-

a) Jaccard Index b) Euclidean Distance **c) Squared Euclidean Distance** d) cosine similarity

18. Question: What does setting New_max=1 and New_min=0 achieve in data normalization?

a) Increases data complexity b) Reduces the impact of outliers
c) Adds noise to the dataset d) **Standardizes data within a specific range**

19. Which of the following is common technique to replace missing data in a dataset?

a) Mean **b) Median** **c) Mode** **d) Random Value**

20. What is the cosine similarity between the vectors (1, 0) and (0, 1)?

a) 1 **b) 0** c) 0.5 d) 2

21. What is the primary purpose of converting an image to grayscale in machine learning algorithms?

a) To increase computational complexity
b) To introduce color variations
c) To reduce computational complexity
d) To improve image resolution

22. In the context of image normalization, what is the benefit of scaling all images to a common range such as [0,1]?

a) It increases computational complexity
b) It ensures fairness across all images
c) It introduces colour variations
d) It reduces the need for data augmentation

23. What is data augmentation in the context of image processing?

a) Increasing the size of an image dataset
b) Making minor alterations to existing data to increase diversity
c) Reducing the diversity of a dataset
d) Converting images to grayscale

24. What is the assumed seasonality for a monthly time series?

a) 7 **b) 12** c) 30 d) 365

25. Executing print((lambda x, y: x//y)(4, 3)) in Python produces
a) 0 **b) 1** c) 4/3 d) 7

26. Executing print(map(lambda x: x**3 , [0,1,2])) in Python produces
a) [0,0,0] b) [0,1,2] **c) [0,1,8]** d) [0,1,3]

27. Executing print(list(filter(lambda x: x > 2 and x < 8, [-1,0,5,3]))) in Python produces
a) [5,3] b) [3,5] c) [-1,0,5,3] d) 8

28. Executing print(functools.reduce(lambda x, y: x+y, [1,2,3,4])) in Python produces
a) 24 **b) 10** c) [1,3,6,10] d) 1

29. Which of the following database is not a relational database?
a) SQLite b) MySQL c) Oracle d) MS Access

30. Which of the following library is used for data visualization?
a) TensorFlow b) Scrapy c) Scikit Learn **d) Matplotlib**

31. Which of the following is not a tool for data processing, machine learning algorithm implementation, and visualization.
a) SAS b) Weka c) RapidMiner d) SAS and WEKA

32. Complex data streams can be analyzed and visualized dynamically using
a) Apache Spark b) Scrapy c) MS Excel d) MS Powerpoint

33. metrics.DistanceMetric.get_metric is a function defined in
a) Pandas b) Scrapy **c) Sklearn** d) Matplotlib

34. The output of the following code is

```
 dist = get_metric('euclidean')
 X = [[2, 3]]
 Y = [[2, 2]]
dist.pairwise(X,Y)
```

a) 1 b)5 c) 9 d) 0

35. Let x={a,b,d} and y={b,c} then $S_{jaccard}$.(x,y) is ---------
a) 0.25 b) 1 c) 0.5 d)

36. To normalize the following dataset
10,40,50,10,50, 70,90,30
a) Divide each value by the mean of the data
b) Replace each value x by (x - the mean of the data / the standard deviation of the data)
c) A followed by B
d) Replace each value x by (x - the standard deviation of the data/ the median of the data)

37. The missing value in the following data may be replaced by
10,40,50,10,50,?,10,60,10,30
a) the mean of the remaining data **b) a value of 10**
c) the median of the remaining data **d) a value of 30**

38. Which of the following distances has three components (distance due to span, content and position)
a) D("Hello","HALA") b) D({a,b,c,d},{a,d})
c) D(10:30,5:20) d) D(10101,01111)

39. Which of the following is an example of Application of Dimensionality Reduction
a)Microarray data analysis **b) Protein classification**
c) Face recognition **d) Handwritten digit recognition**

40. Which of the following is not a Challenge in the analysis of the data:-
a) Few samples b) Mixed data types and unbalanced data
c) Very high dimensionality d) Noise

41. To use PCA in python you should import it using " From ------------ import PCA"
a) sklearn b) sklearn.models **c) sklearn.decomposition** d) sklearn.metrics

42. MRMR ------------ the relevancy while ------------- the redundancy
a) maximizes /minimizes b) minimizes/ minimizes c) minimizes/ maximizes d) maximizes / maximizes

43. which of the following is true regarding the Wrapper model
a) Relying on a predetermined classification algorithm
b) Using predictive accuracy as goodness measure
c) High accuracy
d) computationally expensive

Regarding the following code, answer the following two questions:-

```
iris = datasets.load_iris()
df = pd.DataFrame(iris['data'], columns = iris['feature_names'])
scalar = StandardScaler()
scaled_data = pd.DataFrame(scalar.fit_transform(df))
pca = PCA(n_components = 2)
pca.fit(scaled_data)
data_pca = pca.transform(scaled_data)
```

44. What is the purpose of using StandardScaler() in the following code
    a) reduce the dimension b) fill missing data **c) normalize the data** d) remove noise
45. the number of columns of the data_pca
    a) 4 **b) 2** c) 3 d) 1

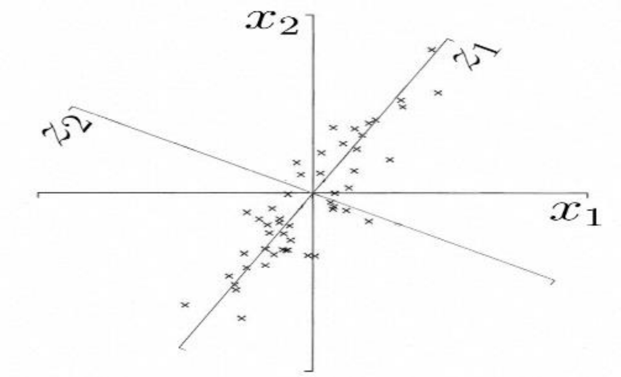
Given the following term frequencies in a corpus D that contains 3 documents D1..D3, answer the following questions:-

Document 1 (D1)	
Term	Term
Caw	2
Sudan	1
Camel	1

Document 2 (D2)	
Term	Term
Sudan	3
Caw	2
Nile	1

Document 3 (D3)	
Term	Term Count
Egypt	2
Nile	2
Caw	1

46. The resulting data matrix will be of size
    a) 3×5 b) 4 × 4 c) 5×5 d) 5×4
47. The normalized term frequency of tf (“camel”,D1) is
    a) 0.20 b) 3 c) 4 **d) 0.25**
48. The inverse document frequency idf(“Camel”,D)
    a) 3 b) 1 c) 1/3 d) 0
49. what is the tflogidf(“caw”,D)
    a) 0 b) 1 c) 3 d) 5
50. The resulting distance matrix will be of size
    a) 3×5 b) 4 × 4 c) 5×5 **d) 3×3**
51. The corresponding feature vector of document D1 using binary term frequency is
    a) [1 1 1 0 0] b) [1 0 0 0 1] c) [1 0 1 1] d) [2 1 1]
52. The correlation between the data using the new axes z1,z2 is ---------- than the correlation between the same data with respect to the axes x1,x2
    a) Higher **b) lower** c) equals d) higher or equals
53. Which axis you may neglect to reduce the dimension
    a) z1 **b) z2** c) z1 or z2 d) z1 and z2



Best Wishes

DST Revision After Midterm

Feature Selection& Reduction Techniques& Applications (Handout5)

1-Dimensionality reduction techniques are primarily used for:

a) Data visualization

b) Data compression

c) Noise removal

d) All of the above

Answer: d) All of the above

2-Which of the following is an application of dimensionality reduction?

a) Customer relationship management

b) Image retrieval

c) Face recognition

d) All of the above

Answer: d) All of the above

3-Document classification involves:

a) Classifying unlabeled documents

b) Storing and retrieving documents efficiently

c) Removing noise from documents

d) None of the above

Answer: a) Classifying unlabeled documents

4-Gene expression microarray analysis deals with:

a) Classifying novel samples into disease types

b) Analyzing high-dimensional microarray data

c) Reducing noise in gene expression data

d) All of the above

Answer: d) All of the above

5-Feature selection differs from feature reduction in that:

a) Feature selection uses all original features

b) Feature reduction uses a subset of original features

c) Feature selection considers correlation between features

d) Feature reduction is computationally expensive

Answer: b) Feature reduction uses a subset of original features

6-Which model of feature selection relies on a predetermined classification algorithm?

a) Filter model

b) Wrapper model

c) MRMR model

d) Unsupervised model

Answer: b) Wrapper model

7-Principal Component Analysis (PCA) is used for:

a) Data compression

b) Dimensionality reduction

c) Feature selection

d) Image retrieval

Answer: b) Dimensionality reduction

8-The MRMR feature selection algorithm aims to:

a) Maximize redundancy between features

b) Minimize relevance of target features

c) Maximize relevance and minimize redundancy

d) None of the above

Answer: c) Maximize relevance and minimize redundancy

9-Which type of feature reduction algorithm is based on linear transformations?

a) Unsupervised

b) Supervised

c) Semi-supervised

d) Nonlinear

Answer: a) Unsupervised

10-The geometric interpretation of principal components involves:

a) Finding the line of best fit in X space

b) Finding the plane perpendicular to the first principal component

c) Minimizing the distance between data points and principal components

d) All of the above

Answer: d) All of the above

11-Which library in Python can be used for PCA?

a) NumPy

b) Pandas

c) Scikit-learn

d) Matplotlib

Answer: c) Scikit-learn

12-What is the purpose of standardizing features before applying PCA?

a) To normalize the data

b) To remove outliers

c) To ensure all features have equal importance

d) None of the above

Answer: a) To normalize the data

13-Information gain is a measure used in:

a) Wrapper model

b) Filter model

c) MRMR model

d) Principal Component Analysis

Answer: b) Filter model

14-The minimum redundancy and maximum relevance (MRMR) feature selection algorithm uses:

a) Heuristic search

b) Complete search

c) Nondeterministic search

d) Simple heuristic algorithm

Answer: d) Simple heuristic algorithm

15-Which feature reduction algorithm is based on nonlinear transformations?

a) Latent Semantic Indexing (LSI)

b) Independent Component Analysis (ICA)

c) Manifold learning

d) Linear Discriminant Analysis (LDA)

Answer: c) Manifold learning

16-What does the filter model of feature selection rely on?

a) General characteristics of data

b) A predetermined classification algorithm

c) Heuristic search strategies

d) Sequential forward selection

Answer: a) General characteristics of data

17-The algebraic derivation of principal components involves:

a) Calculating the covariance matrix

b) Solving a system of linear equations

c) Applying matrix factorization techniques

d) None of the above

Answer: b) Solving a system of linear equations

18-Which feature reduction algorithm is based on unsupervised learning?

a) Latent Semantic Indexing (LSI)

b) Linear Discriminant Analysis (LDA)

c) Canonical Correlation Analysis (CCA)

d) Partial Least Squares (PLS)

Answer: a) Latent Semantic Indexing

The Influence of the Lambda Calculus on Programming Languages Handout6

1-What is a lambda function in Python?

a) A built-in function

b) An anonymous function defined with the lambda keyword

c) A special type of recursive function

d) A function that can only be used once

Answer

b) An anonymous function defined with the lambda keyword

2. What is the correct syntax for a lambda function that adds two numbers, a and b?

a) lambda a, b: a + b

b) lambda (a, b): a + b

c) function (a, b): return a + b

d) (lambda a, b: a + b)

Answer:

a) lambda a, b: a + b

3- How do you call a lambda function that multiplies two numbers?

a) (lambda a, b: a * b)(5, 3)

b) lambda a, b: a * b(5, 3)

c) call(lambda a, b: a * b, 5, 3)

d) lambda(5, 3, a * b)

Answer:

a) (lambda a, b: a * b)(5, 3)

4- Which of the following is true about lambda functions?

a) They can contain multiple expressions

b) They can only have one parameter

c) They return the result of the expression automatically

d) They must contain a return statement

Answer:

c) They return the result of the expression automatically

5-How do you use a lambda function with the map() function in Python?

a) map(lambda x: x * 2, [1, 2, 3])

b) lambda x: x * 2, map([1, 2, 3])

c) map([1, 2, 3], lambda x: x * 2)

d) lambda map(x: x * 2, [1, 2, 3])

Answer:

a) map(lambda x: x * 2, [1, 2, 3])

6-What does this lambda function do? lambda x: x > 10

a) Adds 10 to x

b) Multiplies x by 10

c) Checks if x is greater than 10

d) Reduces x by 10

Answer:

c) Checks if x is greater than 10

Explanation: This lambda function returns True if x is greater than 10, else False

7- How do you use a lambda function as a key for sorting a list of tuples by the second element?

a) sorted(my_list, key=lambda x: x[1])

b) lambda x: x[1], sorted(my_list)

c) sorted(my_list, lambda x: x[1])

d) sort(my_list, key=lambda x: x[1])

Answer:

a) sorted(my_list, key=lambda x: x[1])

8-Can lambda functions capture variables from the enclosing scope?

a) Yes

b) No

c) Only global variables

d) Only if passed as parameters

Answer:

a) Yes

9- How would you filter out all negative numbers from a list using a lambda function?

a) filter(lambda x: x > 0, my_list)

b) lambda x: x > 0, filter(my_list)

c) filter(my_list, lambda x: x > 0)

d) lambda filter(x: x > 0, my_list)

Answer:

a) filter(lambda x: x > 0, my_list)

10- How is a lambda function typically defined?

a) Using the keyword "define"

b) Using the keyword "function"

c) Using the keyword "lambda"

d) Using the keyword "anonymous"

Answer:

c) Using the keyword "lambda"

11-How many arguments can a lambda function take?

a) Only one argument

b) Exactly two arguments

c) Any number of arguments

d) Only keyword arguments

Answer:

c) Any number of arguments

12- What is the purpose of lambda functions in Python?

a) To define complex functions with multiple expressions

b) To create anonymous functions for simple operations

c) To define functions with default arguments

d) To create global functions

Answer:

b) To create anonymous functions for simple operations

13-Which built-in Python functions are often used with lambda functions?

a) max() and min()

b) sum() and average()

c) map() and filter()

d) sort() and reverse()

Answer:

c) map() and filter()

14What does a lambda function return?

a) Multiple values

b) None

c) A single value

d) A list of values

Answer:

c) A single value

Which of the following is a valid use of a lambda function?

a) Defining a complex sorting algorithm

b) Creating a function with multiple expressions

c) Writing a function with a docstring

d) Passing a simple operation as an argument to another function

Answer:

d) Passing a simple operation as an argument to another function

Classification& Regression Handout 7

1. ………..a form of data analysis that extracts models describing important data

classes.

a. Classification

b. Analysis of data

c. Extraction of data

d. Dataset

2. Data classification is a two-step process, consisting of:

a. Data, Information

b. Learning, Classification

c. Knowledge, Information

d. Data, Knowledge

3. Data that used to check the model is working correctly

a. training set

b. test set

c. raw data

d. none of the above

4. If the test set is used to select models, it is called ………

a. training set

b. validation set

c. none of the above

5. The test set is dependant on the training set.

a. True

b. False

6. In ________ the groups are not predefined.

a. Association rules.

b. Summarization.

c. Clustering.

d. Prediction

7. The number of classes is known in.....

a. classification

b. clustering

c. none of the above

8. Dissimilarities and similarities are assessed based on the attribute ......

describing the objects and often involve distance ......

(a) Values, Measures

(b) Measures, Values

(c) Data, Measures

(d) Measures, Data

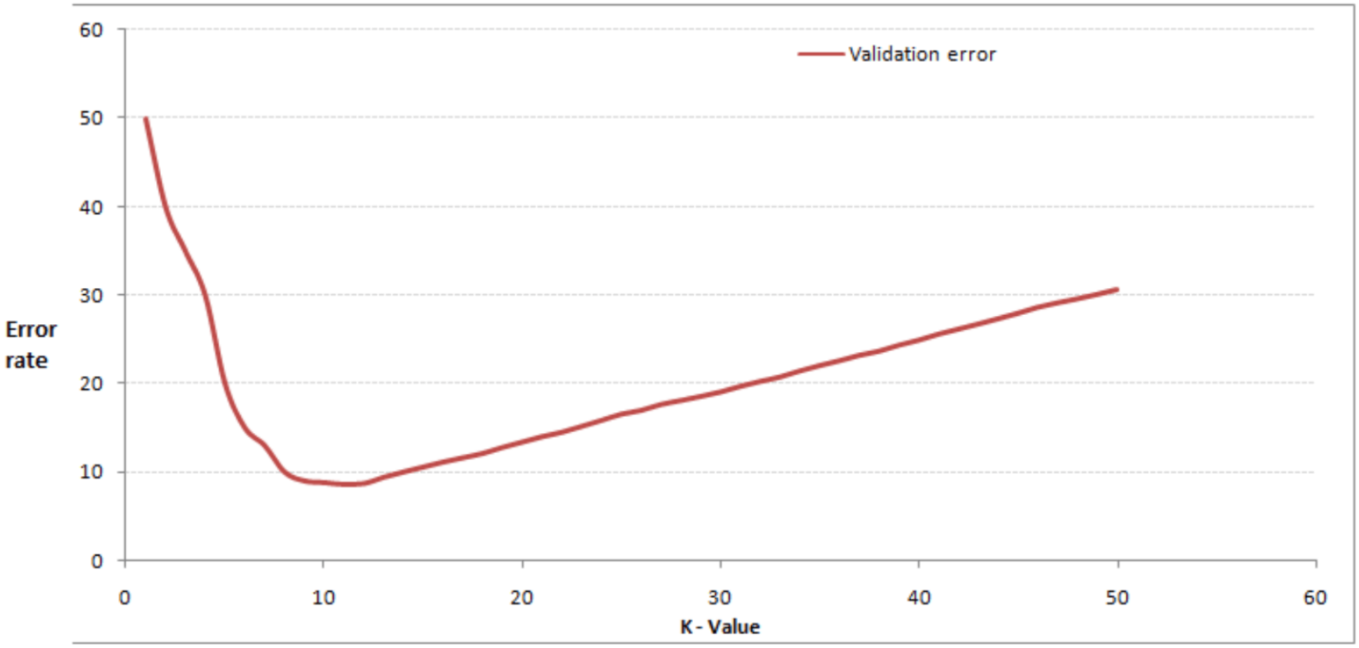
9- In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.

a-3

b-10

c-20

d-50



10-Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

a-Linear Regression

b-K-NN

c-Logistic Regression

d-NN

11- Which of the following statement is true about k-NN algorithm?

1- k-NN performs much better if all of the data have the same scale.

2-k-NN works well with a small number of features (X's), but struggles when the number of inputs is very large

3-k-NN makes no assumptions about the functional form of the problem being solved

a-1 and 2

b-1 and 3

c-Only 1

d-All of the above

12-When you find noise in data which of the following option would you consider in k-NN?

a-I will increase the value of k

b-I will decrease the value of k

c-Noise can not be dependent on value of k

d-None of these

13- The basic distinction between a linear regression model and generalised (1)

linear regression model is the following

1. **The errors in the linear regression are normally distributed while they can have a

more general distribution for the generalised linear model

2. The errors in the linear regression model are homoskedstic while they are

heteroskedstic in a generealised linear model

3. The generalised linear model is not used for continuous dependent variable while that is

not the case with the linear regression model

4. The linear regression model is easy to estimate while the generalised linear regression

model is not easy to estimate.

14- The process of training a predictive model with well defined target values is known as (1)

1. Unsupervised learning

2. **Supervised learning

3. Model estimation

4. Model testing

15- The following is true of the Knn algorithm (2)

1. It has slow training phase

2. It has a fast classification phase

3. **Makes no assumptions about the data distribution

4. Produces a predictive model

16- The trade off between over fitting and under fitting training data is called (3)

1. **The bias-variance tradeoff

2. The residual sum of squares

3. The tradeoff curve

4. The null deviance

17- The following transformation is called the z-score transformation (2)

1. (X-max(X))/(Max(X)-Min(X)

2. (X-Mean(X))/(Max(X)-Min(X))

3. **(X-Mean(X))/Standard Deviation of X

4. Mean(X)/Max(X)

18- How do you deal with Euclidean distance for nominal data in the context of Knn (3)

classification?

1. **Using dummy coding

2. Ignoring such data

3. Deleting those observations

4. Replacing these observations by 0

19- The following is the correct code for a function that normalizes the data (1)

1. normalize <- function(x) { return ((x - max(x)) / (max(x) - min(x))) }

2. **normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }

3. normalize <- function(y) { return ((x - min(x)) / (max(x) - min(x))) }

4. normalize <- function(x) { return ((x - min(x)) / (min(x) - max(x))) }

20- The following is the correct code to execute a Knn model (where train is the (2)

training data, test is the testing data, labels are stored in train_labels and we have a 7 nearest neighbour classifiction

1. wbcd_test_pred <- knn(train =train, test = test, cl =train_labels, k=21)

2. **wbcd_test_pred <- knn(train =train, test = test, cl = train_labels, k=07)

3. wbcd_test_pred <- knn(train = wbcd_train, test = wbcd_test, cl = wbcd_train_labels,k=21)

4. wbcd_test_pred <- knn(train = test, test = train, cl = test, k=21)

Clustering algorithms Handout 8

1-Which of the following is finally produced by Hierarchical Clustering?

a) final estimate of cluster centroids

b) tree showing how close things are to each other

c) assignment of each point to clusters

d) all of the mentioned

2-…………. method works by grouping data objects into a hierarchy or “tree” of the cluster.

a. Hierarchical

b. K-Means

c. K-Medoids

d. None of the above

3-There are …… styles of hierarchical clustering algorithms to build a tree from the input set S

a. 1

b. 2

c. 3

d. 4

4-Agglomerative is ……. tree

a. Top-Down

b. Bottom-Up

c. Both a & b

d. None of the above

5.Divisive is ……. Tree

a. Top-Down

b. Bottom-Up

c. Both a & b

d. None of the above

6.Consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

a. Single linkage

b. Complete linkage

c. Average linkage

d. None of the above

7.A general version of K-means……….

a. Hierarchical

b. K-Means

c. K-Medoids

d. None of the above

8.k-medoids is a:

a. Partitioning methods

b. Hierarchical Methods

c. Model-based clustering

d. None of the above

9. which of the following is not a clustering Method:

a. K-means

b. CLARANS

c. k-medoids

d. None of the above

10.The …… methods can be integrated to cluster data with mixed numeric and nominal values.

a. K-Modes

b. K-Means

c. K-Medoids

d. A &B

11.____________ is a self-learning technique in which system has to explore data.

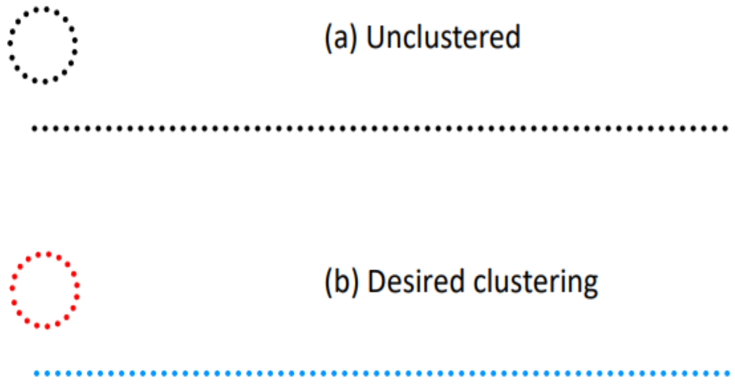
a. Supervised Learning

b. Unsupervised Learning

c. semi-supervised Learning

d. Reinforcement Learning

12Which of the following methods will cluster the data in panel (a) of the figure below into the two clusters (red circle and blue horizontal line) shown in panel (b)? Every dot in the circle and the line is a data point. In all the options that involve hierarchical clustering, the algorithm is run until we obtain two clusters.



(a) Unclustered

(b) Desired clustering

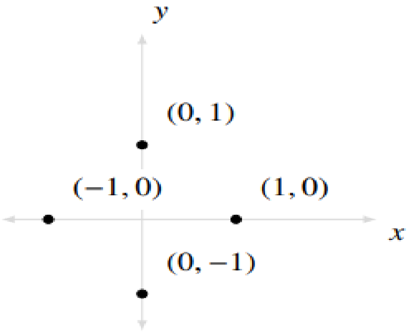
A- Hierarchical agglomerative clustering with Euclidean distance and complete linkage

B- Hierarchical agglomerative clustering with Euclidean distance and single linkage

C- Hierarchical agglomerative clustering with Euclidean distance and centroid linkage

D- k-means clustering with k = 2

13- Consider running the hierarchical agglomerative clustering algorithm on the following set of four points in R 2 , breaking ties arbitrarily. If we stop when only two clusters remain, which of the following linkage methods ensures the resulting clusters are balanced



A: Complete linkage

B: Centroid linkage

C: Average linkage

D: All the above