Sea of Questions

Questions

Overview of Data Science

Which statement is the most closely related to "The curse of dimensionality"?

a. The high dimensionality may pose difficulties for storage and computation

b. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance.

c. When the dimensionality increases, the difficulty of data analysis may not be affected significantly

Is Business understanding a crucial step in the product-driven data science process?

a. No, it does not relate to Data Science
b. No, we can ignore that step

c. Yes, of course

Data crawling and preprocessing

Which of the following accurately describes XPath?

a. XPath is the same as an XML file.

b. XPath is a query language.

c. XPath is a programming language.
d. XPath can be read using a Word document.

Does Scrapy natively support incremental scrawling strategy?

Yes
No

In Scrapy way, how to store crawled data into databases?

a. Write a hook into item pipelines

b. Write a hook into downloader
c. Write a hook into spider middleware

What is the main difference between Web-Scraper and Scrapy?

a. Scrapy is a library whereas Web-Scraper is stand-alone
b. Scrapy relies on XPath, whereas Web-Scraper does not

c. Scrapy is a library whereas Web-Scraper is a web-browser plugin

d. Web-Scraper is more refined than Scrapy, because it relies on a selector hierarchy

Can robots.txt practically stop unwanted web crawlers?

a. Yes

b. No

Data cleaning and integration

Can Google OpenRefine import data on remote URL?

Yes

No

Not a problem of data quality at value level?

Synonym

Missing value

Syntax violation

What is not a cause of noises in data?

a. Different considerations between the time when the data was collected and When it is analyzed

b. Faulty data collection instruments

c. Human error at data entry

Why data in real world is dirty?

a. Incompete

b. Integrated

c. Noisy

d. Inconsistent

Exploratory data analysis

What is the goal of exploratory data analysis?

a. Get a summary of the data, visualize and understand about the data

b. Visualize and make the data clean
c. Make the data clean, optimize a model, increase the predictiveness
d. Understand about data and transform data into some forms

Which Libs in Python should we use for exploratory data analysis?

a. SciPy and Numpy

b. NLTK, Spacy

c. Requests, Scrapy, BeautifulSoup
d. Tensorflow, Keras, Scikit-learn

e. Pandas
f. Matplotlib

What conclusions can be drawn from a box plot in exploratory data analysis?

a. Is variability different between subgroups?
b. Is the location concentration different between subgroups?
c. Is there outliers?
d. Is there any important feature (variable)?

What conclusions can be drawn from a histogram in exploratory data analysis?

a. Is the distribution of the data symmetric or skewed.
b. The dispersion of the data.
c. The distribution of the set of observations.
d. The data centralization.

e. Is there outliers in the data?

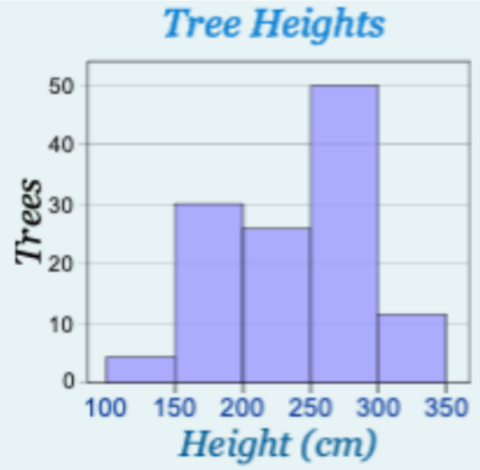
Info gainable from box plot?

Skewness

Probability distribution

Lower/upper quartile
Gap

Correct statement?



Histogram of tree heights data

What kind of bar chart? `q.plot(kind='barh')`

Bar graph

Data visualization

Which method shows hierarchical data in a nested format?

a. Bar chart

b. Treemap

c. Population pyramid
d. None of the other options

The three layers that make up the architecture are the backend, the artist, and the scripting layers?

Select one:

a. Matlab

b. Matplotlib

c. Pyplot

d. Seaborn

Temperature is of which type?

a. Unordered continuous data

b. Ordered discrete data

c. Unordered discrete data

d. Ordered continuous data

Machine learning

Difference between supervised- and unsupervised learning?

a. From the type of the output which is often a real number in supervised learning

b. From the aim of the algorithm, unsupervised learning often does not do prediction

c. From the training data for which supervised learning often requires labels/responses for the training phase

d. From the way we train a model, supervised learning means that we have to provide detailed steps for a machine to learn

What is the role of a loss function?

a. To measure the loss/error when making future prediction

b. To measure the error in some senses and to play as the objective function for learning from data

c. No role in the data science process

Learning a decision tree by the ID3 algorithm will stop if

a. The tree is big enough
b. The tree cannot classify correctly all the training data
c. The tree classifies correctly all the training data

d. The tree classifies correctly all the training data, or at any path all the attributes are used

Overfitting may refer to the situation where

a. Too few training data for a machine to learn
b. A method can predict inaccurately the behaviour of another method

c. A method makes small error rate on the training data while having significantly larger error rate for future data

d. Too many training data so that a machine can learn easily

Big data analysis

Which of the following scenario may not be a good fit for HDFS?

a. Storing enormous small files.

b. Storing data related to applications requiring low latency data access
c. Scenarios requiring random writes to the same file

d. None of the mentioned.

Velocity is a challenge of the era of big data, and refers to

a. The speed of analysis
b. The data that vary heavily
c. The computation it requires massively

d. The data that come continuously and fast

Variety is a challenge related to big data, and refers to

a. The data that comes in continuously and fast
b. The computation power that big data requires
c. The data with high uncertainty due to the presence of fake/noisy information in some sources (particularly on the internet)

d. The different kinds of data that must be handled: structured/unstructured data

Point out the correct statement:

a. Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data.

b. Hbase is a not relational database but it supports SQL.
c. Pig is a relational database with SQL support.
d. All of the mentioned.

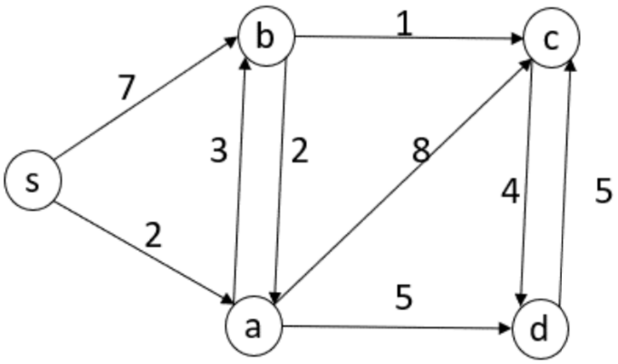
… function is responsible for consolidating the results produced by each of the Map0 functions/tasks.

a. Map
b. All of the mentioned
c. Reducer

d. Reduce

Text, image, graph analysis

Dijkstra, shortest path from s to c?



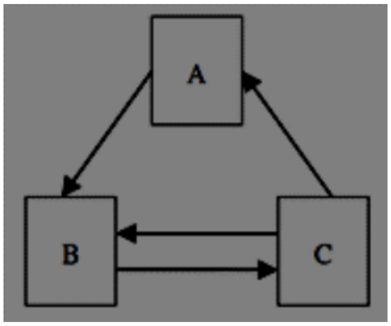
6 (s → a → b → c)

What is the most famous algorithm to rank web pages in the search engine results?

a. Textrank
b. Webrank

c. Pagerank

PageRank of A, damping factor = 0,7



0.2314

What is the purpose of histogram equalization?

a. To reduce noise from images.
b. To represent image content.
c. To increase the brightness of an image.

d. To enhance the contrast of an image.

Given an uncompressed grayscale image of 256 levels, how many byte(s) per pixel does it need?

a. 1

b. 3
C. 24
d. 8

Evaluation of analysis results

Is Hold-out a method for data preprocessing and understanding?

a. No, it is a method for training a model from a given dataset.

b. No, it is a strategy for model assessment and selection.

c. Yes, of course.

You made a system to predict network attacks and you sure that it has a testing accuracy of 99%. However your boss says that your system is useless in practice. What may be the reasons?

a. Your boss does not have enough knowledge to understand your hard work and system.
b. You are unlucky.

c. The training set may be problematic.
d. Accuracy may not reflect what your boss wants in this domain.
e. Your evaluation of the system may be done incorrectly.

Assume that you train a classifier on 10,000 training points and obtain a training accuracy of 99%. However, when you submit it to Kaggle, your accuracy is only 67%. Which of the following has a good chance of improving your performance on Kaggle?

a. Train on more data.

b. Set your regularization coefficient (if any) to 0 .

c. Use a validation set to tune your hyperparameters.

d. Remove randomly some parts of the training set when training your classifier.

What does Evaluation in the data science process include?

a. The evaluation of a system deployment in real life

b. The analysis, assessment, comparison of the results from both offline and reallife scenarios if any

Which is the most suitable statement about model selection?

a. The other statements are wrong.

b. Model selection concerns on the best setting of the parameters for a model when learning from a training dataset. Sometimes it refers to selecting one from many models.

c. Model selection only concerns on selection of the best one amongst different models when working with a given problem.