Data Science and Python: Essential Concepts & Libraries

Python Libraries for Data Science

1. What is Pandas library in Python?

Pandas is a powerful Python library for data manipulation and analysis, providing data structures like DataFrames and Series that make working with structured data intuitive and efficient.

2. Key features of Pandas:

- DataFrame and Series data structures
- Handling of missing data
- Data alignment and integrated indexing
- Reading/writing different file formats (CSV, Excel, SQL, etc.)
- Powerful data merging and joining
- Time series functionality
- Flexible reshaping and pivoting of data

3. What is NumPy Library in Python?

NumPy (Numerical Python) is a fundamental library for scientific computing in Python, providing support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these elements.

4. What is matplotlib library?

Matplotlib is Python's primary 2D plotting library that produces publication-quality figures in various formats and interactive environments across platforms.

5. What is the difference between seaborn library and matplotlib lib?

Matplotlib is a low-level plotting library offering detailed control over visualizations, while Seaborn is a higher-level library built on matplotlib that provides more aesthetically pleasing default styles and specialized visualizations for statistical data analysis.

6. Are Sklearn and Scikit-learn the same library? What is its use in data science?

Yes, Sklearn and Scikit-learn are the same library. It's a comprehensive machine learning library providing tools for classification, regression, clustering, dimensionality reduction, model selection, and preprocessing.

7. What are functions in Pandas and NumPy libraries?

Pandas common functions:

- `read_csv()`, `read_excel()` - importing data
- `head()`, `tail()` - viewing data
- `describe()` - summary statistics
- `groupby()` - aggregating data
- `merge()`, `join()` - combining datasets
- `pivot_table()` - restructuring data

NumPy common functions:

- `array()` - creating arrays
- `zeros()`, `ones()` - creating special arrays
- `mean()`, `std()` - statistical operations
- `reshape()` - changing array dimensions
- `concatenate()` - combining arrays
- `dot()` - matrix multiplication

8. What is a data frame in Python?

A DataFrame in Python (primarily through Pandas) is a 2-dimensional labeled data structure with columns of potentially different types, similar to a spreadsheet or SQL table.

9. How to find duplicates in Python?

python

```python
# Identify duplicate rows
duplicates = df.duplicated()

# View duplicate rows
duplicate_rows = df[df.duplicated()]

# Find duplicates based on specific columns
duplicates_by_cols = df.duplicated(['column1', 'column2'])

# Get all duplicate rows (including first occurrences)
all_duplicates = df[df.duplicated(keep=False)]
```

10. What is the use of the describe command?

The `describe()` function in Pandas generates descriptive statistics of a DataFrame or Series, showing count, mean, standard deviation, min, quartiles, and max values for numerical columns, and provides a useful summary for categorical columns.

11. Which Naive Bayes classification algorithms are used in Python?

- `GaussianNB` - for continuous data
- `MultinomialNB` - for discrete count data (text classification)
- `BernoulliNB` - for binary/boolean features
- `ComplementNB` - an adaptation of Multinomial NB
- `CategoricalNB` - for categorical features

Machine Learning Evaluation Metrics

12. What is the significance of a Confusion Matrix?

A confusion matrix is a table showing the performance of a classification model, displaying counts of true positives, true negatives, false positives, and false negatives. It helps evaluate model accuracy beyond simple percentage metrics.

13. What is TP, TN, FP, FN in a confusion matrix?

- **TP (True Positive)**: Correctly predicted positive class
- **TN (True Negative)**: Correctly predicted negative class
- **FP (False Positive)**: Incorrectly predicted positive class (Type I error)
- **FN (False Negative)**: Incorrectly predicted negative class (Type II error)

14. What is recall?

Recall (sensitivity) measures the proportion of actual positives correctly identified: TP/(TP+FN). It answers: "Of all actual positive cases, how many did we catch?"

15. What is precision?

Precision measures the proportion of predicted positives that are actually positive: TP/(TP+FP). It answers: "Of all cases we predicted as positive, how many actually were?"

16. What is F1 score?

F1 score is the harmonic mean of precision and recall: 2 × (precision × recall)/(precision + recall). It provides a balance between precision and recall, especially useful for imbalanced datasets.

Data Visualization

17. What is the need for data visualization in data science?

Data visualization helps identify patterns, trends, and outliers quickly, communicates findings effectively, aids in exploratory data analysis, makes complex data understandable, and supports decision-making processes.

18. What is an outlier?

An outlier is a data point that differs significantly from other observations in a dataset, typically falling far from the mean or median, and potentially indicating measurement error, data corruption, or a genuine anomaly.

19. When to use histogram and pie chart?

- **Histogram**: Use for showing distribution of continuous data, frequency distributions, and identifying the shape of data distribution
- **Pie Chart**: Use for showing proportions of a whole (parts of 100%), when you have categorical data with relatively few categories

20. What are the challenges in Big Data visualization?

- Handling extremely large volumes of data
- Performance issues with rendering
- Visualizing multi-dimensional data
- Maintaining clarity with complex data relationships
- Balancing detail and overview
- Effective interactive exploration techniques
- Real-time visualization challenges

21. What is joint plot, dist plot?

- **Joint Plot**: Combines scatter plot with histogram/kernel density estimate on each axis, showing both distribution and relationship
- **Dist Plot**: Shows the distribution of a univariate set of observations using histograms, kernel density estimates, or rug plots

22. What are tools used for data visualization?

- **Python libraries**: Matplotlib, Seaborn, Plotly, Bokeh
- **BI tools**: Tableau, Power BI, Looker
- **JavaScript libraries**: D3.js, Chart.js
- **R packages**: ggplot2, Shiny
- **Other**: Google Data Studio, Excel, Flourish

Data Processing

23. What is Data Wrangling?

Data wrangling is the process of cleaning, structuring, and enriching raw data into a desired format for better decision-making in less time. It includes data cleaning, normalization, transformation, integration, and enrichment.

24. What is data transformation?

Data transformation is the process of converting data from one format or structure to another, which may include normalization, aggregation, feature engineering, discretization, or standardization to make it suitable for analysis.

25. What is the use of standard scalar function in Python?

The `StandardScaler` function in scikit-learn standardizes features by removing the mean and scaling to unit variance, which is important for many machine learning algorithms that assume data is normally distributed.

Big Data Technologies

26. What is Hadoop?

Hadoop is an open-source framework for distributed storage and processing of large datasets across clusters of computers using simple programming models.

27. What is HDFS and MapReduce?

- **HDFS (Hadoop Distributed File System)**: A distributed file system providing high-throughput access to application data
- **MapReduce**: A programming model for processing and generating large datasets with a parallel, distributed algorithm on a cluster

28. What are the components of the Hadoop Ecosystem?

- HDFS (storage)
- MapReduce (processing)
- YARN (resource management)
- Hive (SQL-like queries)
- Pig (data flow language)
- HBase (NoSQL database)
- Spark (in-memory processing)
- ZooKeeper (coordination service)
- Oozie (workflow scheduler)
- Sqoop (data transfer tool)
- Flume (log collection)

29. What is Scala?

Scala is a high-level programming language that combines object-oriented and functional programming paradigms, running on the Java Virtual Machine (JVM).

30. What are features of Scala?

- Object-oriented and functional programming
- Static typing
- JVM compatibility
- Concise syntax
- Pattern matching
- Immutability support
- Higher-order functions
- Lazy evaluation
- Type inference
- Actor-based concurrency model

31. How is Scala different from Java?

- Scala supports both OOP and functional programming; Java is primarily OOP
- Scala has more concise syntax with fewer boilerplate code
- Scala has native immutability, Java requires extra effort
- Scala supports traits with implementation; Java interfaces traditionally didn't
- Scala has pattern matching; Java uses switch statements
- Scala has better type inference
- Scala treats everything as an object, including functions

32. List applications of Scala:

- Big data processing with Apache Spark
- Web applications (Play Framework)
- Distributed systems
- Concurrent applications
- API development
- Data engineering pipelines
- Financial systems
- Streaming applications (Akka)

Data Science Fundamentals

33. What is data science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.

34. What is Big Data?

Big Data refers to extremely large and complex data sets that cannot be adequately processed using traditional data processing applications, typically characterized by the "5 Vs."

35. What are the characteristics of Big Data?

The 5 Vs of Big Data:

- **Volume**: Enormous amounts of data
- **Velocity**: High speed of data generation and processing
- **Variety**: Different types of data (structured, unstructured, semi-structured)
- **Veracity**: Uncertainty or reliability of data
- **Value**: Extracting meaningful insights from data

36. List phases in data science life cycle:

1. Business understanding
2. Data acquisition
3. Data preparation and cleaning
4. Exploratory data analysis
5. Feature engineering
6. Modeling
7. Evaluation
8. Deployment
9. Monitoring and maintenance

Statistics

37. What is Central tendency?

Central tendency is a statistical measure that identifies a single value as representative of an entire distribution, typically using mean, median, or mode.

38. What is dispersion?

Dispersion (or variability) measures how spread out data points are from the central value, using metrics like range, variance, standard deviation, and interquartile range.

39. What is mean, mode, mid-range, median? Calculate for 10,22,13,10,21,43,77,21,10

For the dataset [10,22,13,10,21,43,77,21,10]:

- **Mean**: (10+22+13+10+21+43+77+21+10)/9 = 227/9 = 25.22
- **Median**: Sorted [10,10,10,13,21,21,22,43,77], middle value = 21
- **Mode**: 10 (appears 3 times)
- **Mid-range**: (min + max)/2 = (10 + 77)/2 = 43.5

40. What is Variance?

Variance measures the average squared deviation from the mean, quantifying how far individual data points are from the mean value of the dataset.

41. What is Standard Deviation?

Standard deviation is the square root of variance, representing the average distance between each data point and the mean, in the original units of the data.

Machine Learning Concepts

42. What is meant by posterior probability in Naive Bayes theorem?

Posterior probability P(class|features) is the probability of a class given observed features, calculated using Bayes' theorem by combining prior probability and likelihood.

43. What is meant by likelihood probability in Naive Bayes theorem?

Likelihood probability P(features|class) represents the probability of observing specific features given a particular class, a key component in calculating the posterior probability.

44. How can we deal with Missing Values? What ways to deal with missing value or null value:

- Deletion (listwise/pairwise)
- Mean/median/mode imputation
- Predictive modeling imputation
- K-nearest neighbors imputation
- Forward/backward fill for time series
- Using algorithms that handle missing values
- Multiple imputation techniques

- Domain-specific imputation
- Creating "missing" category for categorical data
- Using indicator variables to mark missingness

Natural Language Processing

45. What is NLTK?

NLTK (Natural Language Toolkit) is a leading Python library for working with human language data, providing easy-to-use interfaces to over 50 corpora and lexical resources along with text processing libraries.

46. What is Tokenization in NLP?

Tokenization is the process of breaking down text into smaller units called tokens, which can be words, characters, or subwords, forming the building blocks for further NLP processing.

47. What is Stemming?

Stemming is the process of reducing words to their word stem or root form by removing affixes, often using heuristic rules (e.g., "running" → "run").

48. What is Lemmatization?

Lemmatization is similar to stemming but returns proper dictionary words (lemmas) by understanding context and using vocabulary and morphological analysis (e.g., "better" → "good").

49. What is Corpus in NLP?

A corpus is a large and structured collection of texts used to do statistical analysis, hypothesis testing, or model training in natural language processing.

50. What is Spark framework?

Apache Spark is a unified analytics engine for large-scale data processing, providing high-level APIs in Java, Scala, Python, and R, and supporting SQL queries, streaming data, machine learning, and graph processing.