If already registered, click to check your payment status

Course outline

About NPTEL

How does an NPTEL online course work?

Week-0

Practice: Week 0 : Assignment 0

Week-1

Week-2

Week-3

Week-4

Week-5

Week-6

Week-7

Week-8

Text Transcripts

DOWNLOAD VIDEOS

Books

Week 0 : Assignment 0

Your last recorded submission was on 2025-10-11, 09:05 IST

1) According to the CAP theorem, a distributed system can guarantee at most how many of the following three properties: Consistency, Availability, and Partition Tolerance? ***1 point***

- (•) 1
- ( ) 2
- ( ) 3
- ( ) None

No, the answer is incorrect.
Score: 0
Accepted Answers:
2

2) In the context of CAP theorem, which of the following best describes "Consistency"? ***1 point***

- ( ) Every request receives a response, without guarantee it contains the latest data
- ( ) All nodes see the same data at the same time
- (•) The system continues to function despite network partitions
- ( ) All writes are acknowledged immediately

No, the answer is incorrect.
Score: 0
Accepted Answers:
All nodes see the same data at the same time

3) What happens in a distributed system when a network partition occurs, based on the CAP theorem? ***1 point***

- ( ) The system becomes faster
- (•) The system loses all data
- ( ) A choice must be made between consistency and availability
- ( ) None of the above

No, the answer is incorrect.
Score: 0
Accepted Answers:
A choice must be made between consistency and availability

4) What is the default block size in HDFS (as of Hadoop 2.x)? ***1 point***

- ( ) 64 KB
- (•) 128 MB
- ( ) 256 MB
- ( ) 1 GB

Yes, the answer is correct.
Score: 1
Accepted Answers:
128 MB

5) What is the default block size in HDFS (as of Hadoop 2.x)? ***1 point***

- ( ) NameNode
- (•) JobTracker
- ( ) DataNode
- ( ) TaskTracker

No, the answer is incorrect.
Score: 0
Accepted Answers:
DataNode

6) What is the role of the NameNode in HDFS? ***1 point***

- ( ) Executes MapReduce programs
- (•) Stores actual data
- ( ) Manages the file system namespace and metadata
- ( ) Compresses the data blocks

No, the answer is incorrect.
Score: 0
Accepted Answers:
Manages the file system namespace and metadata

7) In HDFS, if a DataNode fails, what happens to its data? ***1 point***

- ( ) It is lost permanently
- (•) NameNode replicates it from other nodes
- ( ) System crashes
- ( ) JobTracker handles the recovery

Yes, the answer is correct.
Score: 1
Accepted Answers:
NameNode replicates it from other nodes

8) ZooKeeper is primarily used for which of the following in a distributed system? ***1 point***

- ( ) File storage
- (•) Data analysis
- ( ) Coordination and configuration management
- ( ) Data visualization

No, the answer is incorrect.
Score: 0
Accepted Answers:
Coordination and configuration management

9) What is the purpose of the "Znode" in ZooKeeper? ***1 point***

- ( ) A storage location for Hadoop blocks
- (•) A processing unit
- ( ) A node that stores metadata and configuration information
- ( ) A container for YARN jobs

No, the answer is incorrect.
Score: 0
Accepted Answers:
A node that stores metadata and configuration information

10) Which of the following ensures high availability in ZooKeeper? ***1 point***

- ( ) DataNode replication
- (•) Leader election among ZooKeeper nodes
- ( ) HDFS balancer
- ( ) TaskTracker backup

Yes, the answer is correct.
Score: 1
Accepted Answers:
Leader election among ZooKeeper nodes

Check Answers and Submit

Your score is: 3/10

Week 1

1. What does the 'Variety' aspect of Big Data refer to?

(a) The amount of data being generated
(b) The speed at which data is produced
(c) The types and formats of data
(d) The correctness of data

Correct Answer: (c) The types and formats of data

Explanation:
Variety indicates different types of data—structured, semi-structured, and unstructured—such as text, images, audio, and video.

2. Which of the following is used in Hadoop for distributed storage?

(a) Hive
(b) HDFS
(c) YARN
(d) Spark

Correct Answer: (b) HDFS

Explanation:
HDFS (Hadoop Distributed File System) is the storage layer of Hadoop, designed to store large datasets across multiple machines.

3. Which technology enables resource management in a Hadoop cluster?

(a) MapReduce
(b) YARN
(c) HDFS
(d) Pig

Correct Answer: (b) YARN

Explanation:
YARN (Yet Another Resource Negotiator) manages resources and schedules jobs across the Hadoop cluster.

4. What is Apache Spark primarily known for?

(a) Real-time processing using batch jobs
(b) Disk-based computation
(c) Resource management
(d) In-memory computation for fast analytics

Correct Answer: (d) In-memory computation for fast analytics

Explanation:
Apache Spark processes data in-memory, allowing performance up to **100x faster** than traditional MapReduce.

5. What does MapReduce do in the Hadoop ecosystem?

(a) Manages job execution
(b) Provides data security
(c) Splits and processes large data sets in parallel
(d) Stores data

Correct Answer: (c) Splits and processes large data sets in parallel

Explanation:
MapReduce is a **programming model** for processing large datasets by breaking them into smaller chunks (Map) and then combining results (Reduce).

6. What is the purpose of Apache Zookeeper?

(a) Coordinates and manages distributed applications
(b) Stores massive unstructured data
(c) In-memory computation
(d) Provides SQL support

Correct Answer: (a) Coordinates and manages distributed applications

Explanation:
Zookeeper helps **manage configuration, synchronization, and metadata** across distributed systems.

7. Why is traditional RDBMS not suitable for Big Data?

(a) It lacks GUI
(b) It cannot support SQL
(c) It fails to handle large volume, variety, and velocity of data
(d) It is open-source

Correct Answer: (c) It fails to handle large volume, variety, and velocity of data

Explanation:
RDBMS is not designed for the **scale, speed, and heterogeneity** of Big Data.

8. A research lab is storing high-resolution satellite images, videos, and sensor data from different instruments. What Big Data characteristic does this scenario highlight?

(a) Volume
(b) Variety
(c) Veracity
(d) Viscosity

Correct Answer: (b) Variety

Explanation:
This scenario involves **multiple data types** (images, videos, sensor logs), showing **Variety**.

9. You are developing a healthcare monitoring system using wearable sensors that stream data continuously. Which Big Data technologies should you consider for processing this stream?

(a) Spark Streaming and Kafka
(b) Hive and Pig
(c) HDFS and MapReduce
(d) Cassandra and ZooKeeper

Correct Answer: (a) Spark Streaming and Kafka

Explanation:
Kafka handles high-throughput streaming data; **Spark Streaming** processes it in real-time.

10. Which of the following is a NoSQL database suitable for handling unstructured data?

(a) Oracle
(b) Hive
(c) Cassandra
(d) MySQL

Correct Answer: (c) Cassandra

Explanation:
Cassandra is a **distributed, highly scalable NoSQL database** designed for handling huge volumes of unstructured data.

Week 2

1. Which of the following best describes the reason for data locality in HDFS?

A. To reduce disk I/O latency
B. To improve CPU utilization
C. To increase network throughput
D. To reduce data transfer latency and improve performance

Explanation:

A is incorrect: Disk I/O is local to nodes and not the key reason for data locality.
B is incorrect: CPU usage is not directly optimized by HDFS.
C is incorrect: HDFS aims to minimize network use, not maximize it.

D is correct: Data locality brings computation close to data, reducing network transfer and improving performance.

2. Which feature was introduced in HDFS Federation to overcome scalability issues?

A. Block-level striping
B. Multiple NameNodes with independent namespaces
C. Centralized metadata server
D. Heterogeneous replication

Explanation:

A is incorrect: Block striping is not a federation feature.

B is correct: Federation allows multiple NameNodes, each managing its namespace to improve scalability.
C is incorrect: Centralization reduces scalability.
D is incorrect: Replication strategy is not federation-specific.

3. Why is the HDFS default block size larger than traditional file systems?

A. It supports better file security
B. It enables faster metadata lookup

C. It minimizes disk seek time and maximizes throughput
D. It reduces replication overhead

Explanation:
A is incorrect: Security is handled separately.
B is incorrect: Metadata lookup isn't directly tied to block size.

C is correct: Larger blocks reduce seeks and improve throughput.
D is incorrect: Larger blocks mean fewer blocks but replication still applies.

4. What happens when a DataNode fails to send heartbeat to the NameNode in time?

A. DataNode is upgraded to active mode
B. NameNode increases replication factor
C. DataNode is marked dead, and blocks are re-replicated
D. System restarts the failed DataNode automatically

Explanation:

A is incorrect: There's no "active mode" for DataNodes.
B is partially true but misleading; replication is adjusted, not increased.

C is correct: Heartbeats are vital; missed ones trigger replication elsewhere.
D is incorrect: Manual or admin-defined processes handle node recovery.

5. Which MapReduce feature makes it highly fault tolerant?

A. Re-execution of failed tasks on other nodes
B. Heartbeat mechanism
C. Vertical scaling
D. Immediate job termination on failure

Explanation:
A is correct: Failed tasks are retried on other machines.
B is a feature of HDFS, not directly tied to MapReduce fault tolerance.

C is incorrect: Hadoop uses horizontal scaling.
D is incorrect: Jobs are not terminated on single task failures.

6. What would be the outcome if the replication factor is set to 1 in HDFS?

A. No data access will be allowed
B. Data will be more robust
C. There will be no tolerance to node failure
D. Performance will drastically increase

Explanation:
A is incorrect: Access is allowed until the node fails.
B is incorrect: Robustness is reduced.

C is correct: With only one replica, any node failure leads to data loss.
D is incorrect: Performance may degrade due to lack of locality.

7. Which of the following is NOT a component of Hadoop 2.x's YARN architecture?

A. Node Manager
B. Application Master
C. Resource Manager
D. Task Tracker

Explanation:

A, B, and C are key components of YARN.

D is correct: Task Tracker existed in Hadoop 1.x; it was replaced in YARN.

8. In HDFS, the parameter `dfs.blocksize` is primarily tuned to:

A. Reduce network latency
B. Increase memory size of DataNode
C. Control the number of blocks a file is divided into
D. Manage CPU allocation per task

Explanation:

A is partially true but not primary.
B is unrelated.

C is correct: Larger block size results in fewer blocks per file.
D is incorrect: Not related to block size.

9. Which MapReduce phase is responsible for combining all values with the same key?

A. Map
B. Shuffle
C. Reduce
D. Partition

Explanation:

A generates intermediate key-value pairs.
B organizes and transfers data.

C is correct: Reduce aggregates values by key.
D assigns key groups to reducers but doesn't merge them.

10. Why are many small files problematic in HDFS?

A. They increase disk seek time
B. They cause NameNode memory overhead
C. They slow down the replication
D. They reduce the number of map tasks

Explanation:

A is less relevant in HDFS context.

B is correct: Each file/block metadata consumes NameNode memory.
C: Replication is more of a function of number of blocks.
D: They actually increase the number of map tasks.

Week-3

1.A retail analytics company processes daily sales logs in Spark. They want transformations to execute **only when an action is called**, avoiding unnecessary computation. Which Spark feature makes this possible?

A. In-memory caching
B. Lazy evaluation
C. DAG scheduling
D. Broadcast variables

Answer: B
Explanation:

- **A**: Incorrect — Caching speeds up reuse but doesn't delay execution.
- **B**: Correct — Lazy evaluation delays execution until an action is invoked.
- **C**: Incorrect — DAG scheduling determines execution order but doesn't delay it.
- **D**: Incorrect — Broadcast variables optimize shared data, not evaluation timing.

2.In Spark, which of the following is an **action** and not a transformation?

A. map()
B. filter()
C. collect()
D. flatMap()

Answer: C
Explanation:

- **A**: Incorrect — map() is a transformation.
- **B**: Incorrect — filter() is a transformation.
- **C**: Correct — collect() triggers computation and returns data to the driver.
- **D**: Incorrect — flatMap() is a transformation.

3. Which transformation will result in a **narrow dependency** in Spark?

A. groupByKey()
B. reduceByKey()
C. join()
D. sortByKey()

Answer: B
Explanation:

- **A**: Incorrect — groupByKey() shuffles all data with the same key → wide dependency.
- **B**: Correct — reduceByKey() can combine values locally before shuffle → narrow dependency.
- **C**: Incorrect — join() causes wide dependency.
- **D**: Incorrect — sortByKey() causes shuffle → wide dependency.

4.Which Spark feature ensures **lineage information** is used for fault recovery?
A. DAG Scheduler
B. RDD abstraction
C. Executor memory
D. Checkpointing

Answer: B
Explanation:

- **A**: Incorrect — DAG Scheduler handles job stages, not lineage.
- **B**: Correct — RDD maintains lineage to recompute lost partitions.
- **C**: Incorrect — Executor memory is for task execution, not recovery.
- **D**: Incorrect — Checkpointing stores RDD to storage, not just lineage use.

5.Which is a **narrow transformation** in Spark?
A. map()
B. groupByKey()
C. sortByKey()
D. join()

Answer: A
Explanation:

- **A**: Correct — map() processes each partition independently.
- **B**: Incorrect — groupByKey() shuffles data.
- **C**: Incorrect — sortByKey() shuffles data.
- **D**: Incorrect — join() shuffles data.

6.Which of these transformations triggers a **shuffle**?
A. map()

B. filter()
C. reduceByKey()
D. mapValues()

Answer: C
Explanation:

- **A**: Incorrect — map() is local.
- **B**: Incorrect — filter() is local.
- **C**: Correct — reduceByKey() may shuffle grouped data by key.
- **D**: Incorrect — mapValues() is local.

7.In Spark, which component coordinates the execution of tasks across executors?
A. DAG Scheduler
B. Cluster Manager
C. Task Scheduler
D. Driver Program

Answer: C
Explanation:

- **A**: Incorrect — DAG Scheduler divides jobs into stages.
- **B**: Incorrect — Cluster Manager allocates resources, not tasks.
- **C**: Correct — Task Scheduler sends tasks to executors for execution.
- **D**: Incorrect — Driver defines the main program, doesn't assign tasks.

8.An e-commerce site is processing real-time orders in Spark Streaming. To avoid recomputation in case of node failure, they save intermediate RDDs to HDFS. Which technique are they using?
A. Lineage recovery
B. Persistence
C. Checkpointing
D. Broadcast variables

Answer: C
Explanation:

- **A**: Incorrect — Lineage recovery recomputes, doesn't save to storage.
- **B**: Incorrect — Persistence keeps data in memory/disk, not external storage.
- **C**: Correct — Checkpointing saves RDDs to reliable storage for recovery.

- **D**: Incorrect — Broadcast variables are for distributing static data.

9.Which statement is **true** about Spark actions?
A. They transform one RDD into another
B. They trigger execution of transformations
C. They cannot return results to the driver
D. They always involve shuffle

Answer: B
Explanation:

- **A**: Incorrect — Transformations convert RDDs, not actions.
- **B**: Correct — Actions trigger the execution plan built by transformations.
- **C**: Incorrect — collect() returns results to driver.
- **D**: Incorrect — Not all actions require shuffle.

10.In Spark, which persistence level uses **serialization** to reduce memory usage?
A. MEMORY_ONLY
B. MEMORY_AND_DISK
C. DISK_ONLY

D. MEMORY_ONLY_SER

Answer: B
Explanation:

- **A**: Incorrect — MEMORY_ONLY keeps objects in memory in deserialized form.
- **B**: Incorrect — MEMORY_AND_DISK stores in deserialized form first.
- **C**: Incorrect — DISK_ONLY stores data on disk.
- **D**: Correct — MEMORY_ONLY_SER stores serialized objects to save space.

Week-4

1.What does the "soft state" in BASE mean?
A. Data remains in permanent state always
B. State can change over time even without new input
C. State is soft-deleted
D. State is unchangeable once written

Answer: B
Explanation:

- **A**: Incorrect — Soft state means changeable.
- **B**: Correct — System state may change without input due to eventual consistency updates.
- **C**: Incorrect — Not about deletion.
- **D**: Incorrect — That's immutability.

2.Which CAP property is most likely reduced when a system uses asynchronous replication?
A. Availability
B. Consistency
C. Partition tolerance
D. Throughput

Answer: B
Explanation:

- **A**: Incorrect — Availability is unaffected, may even improve.
- **B**: Correct — Asynchronous replication can cause temporary inconsistency.
- **C**: Incorrect — Partition tolerance remains unaffected.
- **D**: Incorrect — Throughput is not directly a CAP property.

3.Which of the following systems is an example of CP under CAP theorem?
A. DNS
B. MongoDB (default)
C. HBase
D. Cassandra

Answer: C
Explanation:

- **A**: Incorrect — DNS is AP (eventual consistency).

- **B**: Incorrect — MongoDB can be tuned, but default is AP.
- **C**: Correct — HBase prioritizes consistency over availability during partitions.
- **D**: Incorrect — Cassandra is AP.

4. What does BASE in distributed databases stand for?
A. Basically Available, Soft state, Eventual consistency
B. Basic Availability, Secure transactions, Eventual consistency
C. Basic Analysis, Soft transactions, Eventual consistency
D. Basically Available, Strong consistency, Eventual consistency

Answer: A
Explanation:

- **A**: Correct — BASE trades strong consistency for availability.
- **B**: Incorrect — "Secure" is not part of BASE.
- **C**: Incorrect — Not related to analysis.
- **D**: Incorrect — Strong consistency contradicts eventual consistency.

5.A stock trading platform must ensure that once a transaction is confirmed, all users see the same updated balance immediately, even if it delays responses during network issues. Which CAP property is being prioritized?
A. Availability
B. Consistency
C. Partition tolerance
D. Durability

Answer: B
Explanation:

- **A**: Incorrect — Availability ensures quick responses, but here delay is accepted.
- **B**: Correct — Consistency ensures all nodes return the same latest data.
- **C**: Incorrect — Partition tolerance deals with handling network splits.
- **D**: Incorrect — Durability is from ACID properties, not CAP.

6. How does CAP theorem impact the design of distributed systems?

A) It emphasizes data accuracy over system availability

B) It requires trade-offs between consistency, availability, and partition tolerance

C) It prioritizes system performance over data security

D) It eliminates the need for fault tolerance measures

Answer: B

Explanation:

- **A**: Incorrect — CAP doesn't always emphasize data accuracy (consistency) over availability; the choice depends on system goals.
- **B**: **Correct** — CAP theorem states that in the presence of network partitions, a distributed system can only fully guarantee **two** of the three: **Consistency (C)**, **Availability (A)**, and **Partition tolerance (P)**, requiring trade-offs.
- **C**: Incorrect — CAP is unrelated to performance vs. security.
- **D**: Incorrect — Fault tolerance is still required; CAP doesn't remove that need.

7. If a distributed system must always be available, which property will it have to sacrifice during partitions?
A. Consistency
B. Partition tolerance
C. Durability
D. Reliability

Answer: A
Explanation:

- **A**: Correct — Availability + Partition tolerance means sacrificing consistency (AP).
- **B**: Incorrect — Partition tolerance is mandatory in real networks.
- **C**: Incorrect — Durability is ACID, not CAP.
- **D**: Incorrect — Reliability is broader than CAP.

8. What does the "soft state" in BASE mean?
A. Data remains in permanent state always
B. State can change over time even without new input
C. State is soft-deleted
D. State is unchangeable once written

Answer: B
Explanation:

- **A**: Incorrect — Soft state means changeable.
- **B**: Correct — System state may change without input due to eventual consistency updates.
- **C**: Incorrect — Not about deletion.
- **D**: Incorrect — That's immutability.

9.Which CAP theorem property ensures the system continues to operate even if some nodes cannot communicate?
A. Availability
B. Partition tolerance
C. Consistency
D. Fault tolerance

Answer: B
Explanation:

- **A**: Incorrect — Availability is about response readiness.
- **B**: Correct — Partition tolerance allows system to work during network splits.
- **C**: Incorrect — Consistency ensures same data across nodes.
- **D**: Incorrect — Fault tolerance is broader than partitions.

10. Why is it impossible to achieve all three CAP properties simultaneously in a distributed system?
A. Due to hardware limitations
B. Because of network latency
C. Because network partitions are inevitable in distributed systems
D. Because data replication is slow

Answer: C
Explanation:

- **A**: Incorrect — Not about hardware.
- **B**: Incorrect — Latency impacts speed, not impossibility.
- **C**: Correct — Partitions can always occur; a system must choose between C and A during them.
- **D**: Incorrect — Replication speed is not the fundamental reason.

Week 5

1. HBase is primarily designed to run on top of:

A. MySQL
B. Hadoop Distributed File System (HDFS)
C. MongoDB
D. PostgreSQL

Correct Answer: B

- A & D: MySQL and PostgreSQL are relational databases, not distributed file systems.
- B: HBase is designed to work on HDFS for scalability, fault tolerance, and high availability.
- C: MongoDB is a NoSQL database but independent, not built on HDFS.

2. HBase prefers _____ over availability in the CAP theorem.

A. Availability
B. Consistency
C. Partition Tolerance
D. Durability

Correct Answer: B

- A: Cassandra emphasizes availability.
- B: HBase prioritizes strong consistency over availability (unlike Cassandra)
- C: Partition tolerance is always required in distributed systems.
- D: Durability is provided, but the key design choice is consistency.

3. Kafka guarantees message order within:

A. A Topic
B. A Partition
C. A Consumer Group
D. Entire Cluster

Correct Answer: B

- A: Topics can have multiple partitions, so global order not guaranteed.

- B:Messages are strictly ordered within a partition.
- C: Consumer group ensures delivery, not ordering.
- D: Cluster-wide ordering is impossible at scale.

4. Kafka was originally developed by:

A. Google
B. Facebook
C. LinkedIn
D. Twitter

Correct Answer: C

- C: Kafka was created by LinkedIn before becoming Apache project.
- A, B, D: Incorrect history.

5. Which of the following is a feature of Spark Streaming?

A. Cannot integrate with batch processing
B. High latency
C. Second-scale latency with fault tolerance
D. Works only with Storm

Correct Answer: C

- A: It integrates with batch and interactive Spark jobs.
- B: It provides low latency, not high.
- C: Spark Streaming achieves near real-time (second-scale) latency with fault tolerance.
- D: Storm is a competitor, not dependency.

6. DStream in Spark Streaming is:

A. Distributed Database
B. A sequence of RDDs
C. A Hadoop file
D. A Kafka broker

Correct Answer: B

- A: Not a database.
- B: DStream is a high-level abstraction representing a continuous stream as RDD batches.
- C: Hadoop file is storage, not stream.
- D: Kafka broker is messaging infrastructure.

7. Zookeeper in HBase is used for:

A. Data Storage
B. Distributed Coordination
C. Caching
D. Query Execution

Correct Answer: B

- A: Data is stored in HDFS, not Zookeeper.
- B: Zookeeper manages coordination among region servers, leader election, etc.
- C: Not a cache.
- D: Queries are handled by RegionServers, not Zookeeper.

8. In HBase, a Cell contains:

A. Only a Row Key
B. Row Key + Column Family + Column Qualifier + Value + Timestamp
C. Column Family + Row Key
D. Region + Store + Memstore

Correct Answer: B

- A: Row key alone is insufficient.
- B:A Cell in HBase is defined by RowKey, Column Family, Column Qualifier, Value, and Timestamp.
- C: Missing qualifier and timestamp.
- D: These are architectural components, not cell structure.

9. A company stores billions of sensor readings in HBase. Each reading is identified by a unique sensor ID and timestamp. The application frequently queries by sensor ID and requires strong consistency.

Which HBase feature makes this possible?
A. Bloom Filters
B. Row Key
C. HFile Compression
D. MemStore

Correct Answer: B

- A: Bloom filters optimize reads but don't guarantee unique identification.
- B:Row Key ensures quick, consistent lookups just like a primary key.
- C: Compression reduces storage size but not lookup consistency.
- D: MemStore is for temporary storage, not retrieval mechanism.

10. A bank uses Spark Streaming to detect fraudulent transactions in real-time. They need results within a few seconds and must combine both streaming (transactions as they occur) and historical batch data (customer profile).

Which Spark feature enables this?
A. Micro-batching
B. HDFS-only storage
C. Column Families
D. Replication Factor

Correct Answer: A

- A: Spark Streaming processes data in micro-batches, integrating batch + streaming workloads.
- B: HDFS storage helps persistence, but not real-time fraud detection.
- C: Column Families belong to HBase, not Spark.
- D: Replication Factor relates to fault tolerance, not hybrid processing.

Week 6

1. **Which is an example of classification?**

a) Predicting next day's temperature in °C
b) Identifying whether a tumor is malignant or benign
c) Estimating sales revenue for next month
d) Predicting rainfall in millimeters

Answer:

a & d) Incorrect: These are numeric predictions → regression.

b) Correct: Classification predicts categories, like malignant/benign

c) Incorrect: Sales revenue is numeric → regression.

2. **In regression tasks, the output is:**

a) A discrete label
b) A probability distribution
c) A continuous numeric value
d) A similarity score

Answer:

a) Incorrect: Classification task.
b) Incorrect: Sometimes used in classification (probabilities), not regression.
c) Correct: Regression predicts continuous values like stock prices.

d) Incorrect: Similarity score is used in clustering, not regression.

3. **In K-means clustering, the elbow method is used for:**

a) Choosing the number of clusters (k)
b) Reducing dimensionality
c) Selecting features
d) Measuring classification accuracy

Answer:

a) Correct: Elbow method finds the optimal k using WSSE curve
b) Incorrect: Dimensionality reduction uses PCA etc.

c) Incorrect: Feature selection step, unrelated to elbow.

d)Incorrect: Accuracy is for classification, not clustering.

4. **Which similarity measure is not a proper distance metric but efficient for sparse vectors?**

a) Euclidean distance
b) Manhattan distance
c) Cosine similarity
d) Jaccard distance

Answer:

a, b, d)Incorrect: All are proper distance metrics.

c) Correct: Cosine similarity is not a true distance metric but efficient for high-dimensional sparse data

5. **Which evaluation metric measures exactness of predictions?**

a) Recall
b) Precision
c) Accuracy
d) F-measure

Answer:

a) Incorrect: Recall = completeness (true positives found).
b) Correct: Precision = proportion of correctly predicted positives among all predicted positives

c) Incorrect: Accuracy = overall correctness.

d) Incorrect: F-measure balances precision & recall.

6. **Which method divides data into k partitions and uses each once for validation?**

a) Holdout method
b) Leave-one-out validation
c) K-fold cross-validation
d) Random subsampling

Answer:

a) Incorrect: Simple train/validation split.

b) Incorrect: Special case with k = N.

c)Correct: K-fold cross-validation uses k partitions
d) Random splits, less structured than k-fold.

7. **What technique helps to avoid overfitting in decision trees?**

a) Increasing tree depth indefinitely
b) Using pre-pruning or post-pruning
c) Using noisy data intentionally
d) Removing validation set

Answer:

a) Incorrect: Leads to overfitting.

b) Correct: Pruning methods control complexity and reduce overfitting

c) Noise increases errors.

d) Validation set is crucial to detect overfitting.

8. **An e-commerce company wants to recommend related items when a customer buys something (e.g., "Customers who bought X also bought Y"). Which ML technique should be used?**

a) Classification
b) Regression
c) Association analysis
d) Clustering

Answer:

a) Incorrect: no category is being predicted.

b) Incorrect: it's not numeric prediction.

c) Association analysis

Correct: Market basket analysis (association rules) is used to find items that frequently occur together.

d) Clustering could group customers, but not directly find item associations.

9. **A weather department wants to group weather patterns into categories like monsoon, snowy, dry, and humid without prior labels. Which technique is best?**

a) Regression
b) Classification
c) Clustering
d) Precision-recall analysis

Answer:

a) Incorrect: Regression predicts numeric values like rainfall.

b) Incorrect: Classification requires labels (not available here).

c) Clustering

Correct: Clustering groups unlabeled data based on similarity (weather conditions)

d) Incorrect: Precision/recall are evaluation metrics, not learning techniques.

10. **A company is using K-means clustering to segment customers. After plotting WSSE for different values of k, they notice the curve bends (forms an elbow) at k = 4. What does this suggest?**

a) The best number of clusters is likely 4
b) They should always increase k for better results
c) Clustering is not possible with this data
d) WSSE does not apply to K-means

Answer:

a) The best number of clusters is likely 4

Correct: The elbow point suggests an optimal trade-off between cluster accuracy and complexity

b) Incorrect: Increasing k always reduces WSSE but may overfit.

c) Incorrect: Clustering is possible; the elbow helps interpret results.

d) Incorrect: WSSE is a standard K-means evaluation measure

Week -7

1.Which is generally considered the best predictive method for classification/regression?

a) Single Decision Tree
b) Bagging
c) Gradient Boosted Trees
d) Random Forest

Answer:

a)Incorrect: Weak predictor.

b)Incorrect: Variance reduction but weaker than boosting.

c) Gradient Boosted Trees

Correct: GBT usually outperforms others in predictive power.

d) Incorrect:Good but slightly worse than GBT.

2. Why do Random Forests lose interpretability compared to single trees?

a) They prune nodes
b) They contain hundreds of trees
c) They use PCA internally
d) They are only regression models

Answer:

a)Incorrect: Pruning doesn't cause loss of interpretability.

b) They contain hundreds of trees

Correct: Large ensembles cannot be easily interpreted by humans.

c) Incorrect:PCA is not used internally.
d) Incorrect:Random Forests handle both regression & classification.

3. Cross-validation in Spark ML is used for:

a) Building deeper trees
b) Improving feature scaling
c) Selecting best model parameters
d) Reducing overfitting by pruning

Answer:

a, b, d) Incorrect:Not main purposes of CV.

c) Selecting best model parameters

Correct: Cross-validation helps choose parameters like tree depth, minInstances.

4. Which of the following is true for Gradient Boosting?

a) Combines weak classifiers iteratively
b) Uses bagging of decision trees
c) Selects features randomly
d) Cannot be used for regression

Answer:

a) Combines weak classifiers iteratively

Correct: Boosting builds strong models from weak learners.

b) Incorrect:That's bagging.

c) Incorrect:Random selection is Random Forest, not boosting.

d) Incorrect:Gradient Boosting works for both classification and regression.

5. Random Forest differs from Bagging by:

a) Selecting random features for each split
b) Using pruning
c) Building only shallow trees
d) Handling missing values

Answer:

a) Selecting random features for each split

Correct: Random Forest decorrelates trees by random feature selection.
b) Incorrect: Pruning is separate.

c) Incorrect: Depth is not limited.

d) Incorrect: Missing value handling is not its defining feature.

6. What is Bagging mainly used for?

a) Reducing bias
b) Reducing variance
c) Improving interpretability
d) Feature selection

Answer:

a) Incorrect: Boosting reduces bias.

b) Reducing variance

Correct: Bagging stabilizes models by averaging and reducing variance.

c) Incorrect: Bagging does not improve interpretability.

d) Incorrect: Feature selection is unrelated.

7. What is the purpose of entropy in decision trees?

a) Measure the number of attributes
b) Measure dataset purity/impurity
c) Reduce dimensionality
d) Increase interpretability

Answer:

a) Incorrect: attributes count is unrelated.

b) Measure dataset purity/impurity

Correct: Entropy quantifies impurity in a dataset. Lower entropy = purer set

c) Incorrect: Dimensionality reduction is PCA's role.

d) Incorrect: Interpretability comes from tree visualization, not entropy.

8. In regression trees, what happens if no split is made?

a) Model predicts the median of targets
b) Model predicts the average of targets
c) Model predicts the mode of targets
d) Model predicts randomly

Answer:

a) Incorrect: Median is not used.

b) Model predicts the average of targets

Correct: Without splits, regression trees predict mean value to minimize squared error.

c) Incorrect: Mode applies to classification, not regression.

d)Incorrect: Random prediction is not valid.

9. A bank wants to predict whether a loan applicant will **default or not**. They use a decision tree, but the model is overfitting badly. Which strategy should they apply?

a) Allow the tree to grow deeper
b) Use pruning or set maxDepth
c) Remove the validation set
d) Train on fewer features

Answer:

a) Incorrect: Deeper tree → more overfitting.
b) Use pruning or set maxDepth

Correct: Controlling tree growth (maxDepth, minInfoGain, pruning) prevents overfitting.

c) Incorrect: Validation is necessary to detect overfitting.

d) Incorrect: Reducing features blindly may lose useful information.

10. A retail chain uses **Spark ML** to analyze the Breast Cancer dataset. They split data into training and test sets (70/30). After training a Decision Tree, accuracy is low. What is the **next best step**?

a) Use cross-validation to tune hyperparameters
b) Reduce dataset size
c) Train without a test set
d) Randomly drop features

Answer:

a) Use cross-validation to tune hyperparameters

Correct: Spark ML supports cross-validation to optimize parameters like maxDepth, minInstancesPerNode.

b) Incorrect: Smaller dataset reduces accuracy further.

c) Incorrect: Test set is essential to evaluate generalization.

d) Incorrect: Random feature dropping can harm model performance.

Week 8

1. **Which of the following is the main purpose of a Parameter Server in machine learning?**
   a) Store raw datasets for training
   b) Store and update distributed model parameters
   c) Visualize the training process
   d) Replace gradient descent

Answer:

a) Incorrect – Datasets are stored in HDFS or other storage, not PS.

b) correct- Store and update distributed model parameters

Parameter servers store model parameters in a distributed manner and support push/pull operations.

c) Incorrect – Visualization is done via monitoring tools.
d) Incorrect – Gradient descent is still used; PS just manages updates.

2. **Which operation in a Parameter Server retrieves the latest parameters?**
   a) Add
   b) Push
   c) Pull
   d) Update

Answer:

a) Incorrect – `add` modifies a parameter.
b) Incorrect – `push` sends updates to server.

c) Correct- Pull – Workers pull parameters from the server to get the latest version.
d) Incorrect – Update is not a defined API keyword.

3. **Which of the following is a problem with asynchronous execution in parameter servers?**
   a) Requires GPUs
   b) Lacks theoretical convergence guarantee
   c) Needs MapReduce
   d) Cannot handle large datasets

Answer:

a) Incorrect – GPUs are optional.

b) Correct- Lacks theoretical convergence guarantee – Because updates may be arbitrarily delayed.
c) Incorrect – MapReduce is unrelated.
d) Incorrect – It can still handle large datasets.

4. **If a page has fewer outgoing links, its PageRank contribution to each link is**:
   a) Higher
   b) Lower
   c) Zero
   d) Independent of number of links

Answer:
a) Correct- Higher – Contribution is divided among outgoing links.
b) Incorrect – Fewer links → higher contribution per link.
c) Incorrect – Not zero unless no links exist.
d) Incorrect – Clearly dependent.

5. **Which framework is specifically designed for large-scale graph analytics and integrates with Spark?**
   a) GraphX
   b) Pregel
   c) GraphLab
   d) Hadoop

Answer:
a) Correct- GraphX – A Spark-based distributed graph processing system.
b) Incorrect – Pregel is from Google, standalone.
c) Incorrect – GraphLab is another framework.
d)Incorrect – Hadoop is data-parallel, not graph-specific.

6. **GraphX introduces which two main abstractions?**
   a) VertexRDD and EdgeRDD
   b) Map and Reduce
   c) Datasets and DataFrames
   d) Triplets and Tuples

Answer:

a)Correct- VertexRDD and EdgeRDD – These represent vertices and edges in GraphX.
b) Incorrect – MapReduce is an older abstraction.
c) Incorrect – Belongs to Spark SQL.
d) Incorrect – Triplets are operators, not core abstractions.

7. **Why is MapReduce inefficient for PageRank?**
   a) It cannot store graphs
   b) It redundantly shuffles graph structure each iteration
   c) It cannot perform joins
   d) It does not support parallelism

Answer:

a) Incorrect – Graphs can be stored.

b) Correct- It redundantly shuffles graph structure each iteration
c) Incorrect – Joins are possible in MapReduce.
d) Incorrect – Parallelism is supported but not efficient for iterative ML.

8. **A tech company trains a deep learning model across multiple GPUs in a data center. They need a way for each worker to update and retrieve model parameters efficiently. Which architecture should they use?**
   a) Parameter Server
   b) MapReduce
   c) GraphX
   d) Pregel

Answer:

a) Correct- Parameter Server – Used for distributed training to push/pull parameters across workers.
b) Incorrect –MapReduce – Inefficient for iterative ML tasks.
c) Incorrect – GraphX – Graph processing, not parameter management.
d) Incorrect – Pregel – Vertex-centric graph framework, not ML parameter sharing.

9. **A flight network graph represents airports as vertices and flights as edges. A company wants to find the most influential airport (hub) in the network. Which algorithm in GraphX should they use?**
   a) K-means
   b) PageRank
   c) Logistic Regression
   d) Linear Regression

Answer:

a) Incorrect – K-means is clustering, not ranking.

b) Correct- PageRank – Identifies importance of nodes in a network, suitable for airport hubs.

c) Incorrect – Logistic regression is for classification.

d) Incorrect – Linear regression predicts numeric values, not graph importance.

10. Which algorithm can GraphX execute for graph analytics?

a) PageRank

b) Triangle Counting

c) Shortest Path

d) All of the above

Answer:

a), b), c)Incorrect- individually since GraphX does more than just one.

d) Correct- All of the above – GraphX supports multiple graph algorithms.