Week 6

1. Point out the wrong statement.

   a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

   b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode
   c) User data is distributed across multiple DataNodes in the cluster and is managed by the NameNode.
   d) DataNode is aware of the files to which the blocks stored on it belong to

   a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

- **Correct**. Replication Factor can indeed be set at both the cluster level and the file level in distributed file systems like HDFS.

  b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode

- **Correct**. A Block Report includes a list of all blocks stored on a DataNode.

  c) User data is distributed across multiple DataNodes in the cluster and is managed by the NameNode.

- **correct**. In a distributed file system like HDFS, user data is stored in a distributed manner across the cluster and managed by the HDFS, not just the local file system of each DataNode.

  d) DataNode is aware of the files to which the blocks stored on it belong to

- **Incorrect**. DataNodes manage blocks of data and are not aware of the higher-level file structure; this information is managed by the NameNode in HDFS.

2. What is the primary technique used by Random Forest to reduce overfitting?

   a) Boosting

   b) Bagging

   c) Pruning

   d) Neural networks

a. Incorrect - Boosting:

- **Not Used in Random Forest:** Boosting is a different technique used in methods like Gradient Boosting, where trees are built sequentially to correct errors from previous trees. It's not used by Random Forest, which relies on bagging.

b. Correct - Bagging (Bootstrap Aggregating):

- **Primary Technique Used by Random Forest:** Bagging is the key technique used by Random Forest to reduce overfitting. In Random Forest, multiple decision trees are trained on different random subsets of the data, created through bootstrapping (sampling with replacement). Each tree is trained independently on its subset of data, and the predictions from all trees are aggregated (typically by voting or averaging) to produce the final result. This approach helps to reduce variance and improves the model's ability to generalize by averaging out the errors from individual trees.

c. Incorrect - Pruning:

- **Not a Primary Technique in Random Forest:** Pruning is a technique used to reduce the size of decision trees by removing parts that are not contributing to the prediction accuracy. While pruning helps to control overfitting in individual decision trees, Random Forest primarily relies on bagging for overfitting reduction.

d. Incorrect - Neural Networks:

- **Not Related to Random Forest:** Neural networks are a different class of models and are not related to the ensemble method of Random Forest.

3. What statements accurately describe Random Forest and Gradient Boosting ensemble methods?

   S1: Both methods can be used for classification task

   S2: Random Forest is use for regression whereas Gradient Boosting is use for Classification task

   S3: Random Forest is use for classification whereas Gradient Boosting is use for regression task

   S4: Both methods can be used for regression

A) S1 and S2
B) S2 and S4
C) S3 and S4
D) S1 and S4

S1: Both methods can be used for classification tasks

- **Correct**. Both Random Forest and Gradient Boosting can be used for classification problems.

S2: Random Forest is used for regression whereas Gradient Boosting is used for Classification task

- **Incorrect**. Random Forest and Gradient Boosting can both be used for both regression and classification tasks.

S3: Random Forest is used for classification whereas Gradient Boosting is used for regression task

- **Incorrect**. As with S2, both methods can be used for both types of tasks.

S4: Both methods can be used for regression

- **Correct**. Both Random Forest and Gradient Boosting can be used for regression tasks as well.

4. In the context of K-means clustering with MapReduce, what role does the **Map** phase play in handling very large datasets?

A) It reduces the size of the dataset by removing duplicates

B) It distributes the computation of distances between data points and centroids across multiple nodes

C) It initializes multiple sets of centroids to improve clustering accuracy

D) It performs principal component analysis (PCA) on the data

A) It reduces the size of the dataset by removing duplicates

- **Incorrect**. The Map phase does not focus on removing duplicates but rather on distributing and processing the data.

B) It distributes the computation of distances between data points and centroids across multiple nodes

- **Correct**. The Map phase is responsible for calculating distances between data points and centroids and distributing this task across nodes.

C) It initializes multiple sets of centroids to improve clustering accuracy

- **Incorrect**. Initialization of centroids is generally done before the Map phase starts and is not part of its functionality.

D) It performs principal component analysis (PCA) on the data

- **Incorrect**. PCA is not typically done in the Map phase; it is a preprocessing step for dimensionality reduction.

5. What is a common method to improve the performance of the K-means algorithm when dealing with large-scale datasets in a MapReduce environment?

A) Using hierarchical clustering before K-means

B) Reducing the number of clusters

C) Employing mini-batch K-means

D) Increasing the number of centroids

A) Using hierarchical clustering before K-means

- **Incorrect**. While hierarchical clustering can be used to initialize centroids, it is not specific to improving K-means in a MapReduce environment.

B) Reducing the number of clusters

- **Incorrect**. Reducing the number of clusters might not improve performance and could lead to less meaningful clustering.

C) Employing mini-batch K-means

- **Correct**. Mini-batch K-means is a method used to handle large-scale datasets efficiently by processing small, random subsets of the data.

D) Increasing the number of centroids

- **Incorrect**. Increasing the number of centroids might not improve performance and could complicate the clustering process.

6.Which similarity measure is often used to determine the similarity between two text documents by considering the angle between their vector representations in a high-dimensional space?

A) Manhattan Distance

B) Cosine Similarity

C) Jaccard Similarity

D) Hamming Distance

A) Manhattan Distance

- **Incorrect**. Manhattan Distance is not used for text document similarity in this context.

B) Cosine Similarity

- **Correct**. Cosine Similarity measures the cosine of the angle between two vectors, making it ideal for text documents in high-dimensional space.

C) Jaccard Similarity

- **Incorrect**. Jaccard Similarity is used for comparing sets and is not based on vector angles.

D) Hamming Distance

- **Incorrect**. Hamming Distance is used for comparing strings of equal length and is not applicable to text document similarity in vector space.

7.Which distance measure calculates the distance along strictly horizontal and vertical paths, consisting of segments along the axes?

A)Minkowski distance

B) Cosine similarity

c) Manhattan distance

D) Euclidean distance

A) Minkowski distance

- **Incorrect**. Minkowski distance generalizes Euclidean and Manhattan distances but is not specific to axis-aligned paths.

B) Cosine similarity

- **Incorrect**. Cosine similarity measures the angle between vectors and does not involve distance calculation.

C) Manhattan distance

- **Correct**. Manhattan distance measures distance along axis-aligned paths (horizontal and vertical segments).

D) Euclidean distance

- **Incorrect**. Euclidean distance measures the straight-line distance between points, not restricted to axis-aligned paths.

8.What is the purpose of a **validation set** in machine learning?

A) To train the model on unseen data

B) To evaluate the model's performance on the training data

C) To tune hyperparameters and prevent overfitting

D) To test the final model's performance

A) To train the model on unseen data

- **Incorrect**. The validation set is not used for training but for model evaluation during training.

B) To evaluate the model's performance on the training data

- **Incorrect**. Evaluation on training data is not the purpose of a validation set; it is used for hyperparameter tuning and model selection.

C) To tune hyperparameters and prevent overfitting

- **Correct**. The validation set is used to tune hyperparameters and monitor performance to avoid overfitting.

D) To test the final model's performance

- **Incorrect**. Testing the final model's performance is done using a separate test set, not the validation set.

9. In **K-fold cross-validation**, what is the purpose of splitting the dataset into K folds?

A) To ensure that every data point is used for training only once

B) To train the model on all the data points

C) To test the model on the same data multiple times

D) To evaluate the model's performance on different subsets of data

A) To ensure that every data point is used for training only once

- **Incorrect**. In K-fold cross-validation, every data point is used for training multiple times, not just once.

B) To train the model on all the data points

- **Incorrect**. The dataset is split into folds, and only K-1 folds are used for training each time.

C) To test the model on the same data multiple times

- **Incorrect**. Each fold is used for testing once, and training is done on the remaining folds.

D) To evaluate the model's performance on different subsets of data

- **Correct**. K-fold cross-validation ensures that the model is evaluated on different subsets of the data, providing a robust measure of its performance.

10. Which of the following steps is NOT typically part of the machine learning process?

A) Data Collection

B) Model Training

C) Model Deployment

D) Data Encryption

A) Data Collection

- **Incorrect**. Data Collection is a fundamental step in the machine learning process.

B) Model Training

- **Incorrect**. Model Training is a core step in machine learning.

C) Model Deployment

- **Incorrect**. Model Deployment is part of the machine learning lifecycle, as it involves putting the model into production.

D) Data Encryption

- **Correct**. Data Encryption is not typically a part of the machine learning process itself, though it may be relevant for data security and privacy.