Introduction to machine learning 1

September 6, 2023

1 Q1- Explain the following with an example

2 1) Artificial Intelligence

3 2) Machine Learning

4 3) Deep Learning

Artificial intelligence - Ai is a smart application that can perform own task without any human intervation example - self driving car , Robots

Machine learning - ML is a subset of Ai . it provides stats tool to learn , analyze , visualize and develop predictive models from the data example- Recommendation system.

Deep learning - Deeplearning is a subset of machine learning main aim is to mimic human brain. example - Object detection, image recognition. # Q2- What is supervised learning? List some examples of supervised learning. Supervised learning is a type of machine learning where an algorithm learns to map input data to output data based on a labeled dataset. example - recommendation system , image classification , finiancial fraud detection. # Q3 - What is unsupervised learning? List some examples of unsupervised learning In supervised machine learning you dont have an output sepecific such as you just have datsset for this dataset you need to form cluster or similar groups Unsupervised learning is a type of machine learning where an algorithm is trained on a dataset without labeled output. Example - customer segmentation , topic modeling

5 Q4- What is the difference between AI, ML, DL, and DS?

AI is a broad field of computer science that aims to create machines, systems, or software that can perform tasks that typically require human intelligence. It encompasses various subfields, including machine learning and deep learning, as well as natural language processing, computer vision, robotics, and more.

ML is a subset of AI that focuses on the development of algorithms and models that allow computers to learn from and make predictions or decisions based on data. ML algorithms improve their performance on a task as they are exposed to more data, without being explicitly programmed for that task. Examples of ML techniques include linear regression, decision trees

DL is a subfield of machine learning that deals with neural networks composed of multiple layers (deep neural networks). DL has gained prominence due to its ability to automatically learn hierarchical features from data, making it particularly effective in tasks like image and speech

recognition. Deep learning models, such as convolutional neural networks (CNNs) for images and recurrent neural networks (RNNs) for sequences, have achieved remarkable success in various AI applications.

Data Science is a multidisciplinary field that combines elements of statistics, computer science, domain knowledge, and data analysis to extract insights and knowledge from data. DS encompasses data collection, cleaning, exploration, feature engineering, modeling, and interpretation of results. Data scientists use various tools and techniques, including statistical analysis, machine learning, and data visualization, to solve complex data-related problems.

6 Q5 - What are the main differences between supervised, unsupervised, and semi-supervised learning?

Supervised Learning:

Labeled Data: In supervised learning, the training dataset consists of input data along with corresponding target labels or outputs. Each data point is associated with a known outcome. Objective: The primary goal is to learn a mapping from input data to output labels so that the algorithm can make accurate predictions or classifications for new, unseen data. Examples: Classification tasks (e.g., image classification, spam detection) and regression tasks (e.g., predicting house prices) are common examples of supervised learning.

Unsupervised Learning:

Unlabeled Data: Unsupervised learning operates on datasets that lack explicit target labels or outputs. The algorithm seeks to discover patterns, structures,or relationships within the data. Objective: The main objective is to uncover hidden insights, group similar data points, reduce dimensionality, or detect anomalies without predefined categories. Examples: Clustering tasks (e.g., customer segmentation), dimensionality reduction (e.g., PCA), anomaly detection, and topic modeling (e.g., identifying themes in text data) are examples of unsupervised learning.

Semi-Supervised Learning:

Mixed Data: Semi-supervised learning combines elements of both supervised and unsupervised learning. The training dataset includes a mixture of labeled and unlabeled data points. Objective: The goal is to leverage the limited labeled data to improve the model's performance on the larger pool of unlabeled data. Semi-supervised learning aims to benefit from the information contained in both labeled and unlabeled data. Examples: Semi-supervised learning can be applied to scenarios where obtaining labeled data is costly or time-consuming. For instance, in a medical diagnosis task, there may be a shortage of labeled patient data, but there is an abundance of unlabeled patient data from various sources.

7 Q6- What is train, test and validation split? Explain the importance of each term.

Training Set: Purpose: The training set is the largest portion of the dataset and is used to train the machine learning model. The model learns from the patterns and relationships within this data. Importance: The training set is crucial for building the model's parameters and making it capable

of making predictions or classifications. The model learns to generalize from the training data, so it's essential that this data is representative and diverse to capture underlying patterns.

Validation Set: Purpose: The validation set, sometimes called the development set or holdout set, is used during model training to tune hyperparameters and assess the model's performance. Importance: By evaluating the model's performance on the validation set, you can fine-tune hyperparameters (e.g., learning rates, regularization strength) and make decisions about the model's architecture and settings. This helps optimize the model's generalization and avoid overfitting (where the model fits the training data too closely but performs poorly on new data).

Test Set: Purpose: The test set is a separate, unseen portion of the data that is not used during model training or hyperparameter tuning. It is used to assess the final performance and generalization of the trained model. Importance: The test set provides an unbiased evaluation of how well the model is likely to perform on new, unseen data. It measures the model's ability to generalize to data it has never encountered before. Without a separate test set, you might mistakenly think your model performs better than it actually does because it has seen the validation data during training.

The importance of each term can be summarized as follows:

Training Set: It is essential for building the model's parameters and teaching it to make predictions based on the patterns in the data.

Validation Set: It is crucial for optimizing the model's hyperparameters and assessing its performance during development. It helps ensure that the model is not overfitting.

Test Set: It provides an unbiased estimate of the model's performance on unseen data, helping you determine if the model generalizes well to real-world scenarios.

Q7 - How can unsupervised learning be used in anomaly detection? Unsupervised learning is commonly used in anomaly detection because it excels at identifying patterns and structures within data without the need for labeled anomalies # Q8- List down some commonly used supervised learning algorithms and unsupervised learning algorithms. Supervised Learning Algorithms: 1. Linear Regression: 2. Logistics regression: 3. Decision trees: 4. Random forest: 5. Support Vector Machines(SVM): 6. K-Nearest Neighbors(K-NN): 7. Naive Bayes: 8. Neural Networks:

Unsupervised learning Algorithms: 1. K-means Clustering: 2. Hierarchical clustering: