MACHINE LEARNING TOPICS

Supervised Learning

In supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.

Unsupervised Learning

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.

Reinforcement Learning

Using reinforcement learning, the model can learn based on the rewards it received for its previous action.

Overfitting

Overfitting in machine learning occurs when a model learns the training data too well, including its noise and outliers, leading to poor performance on new, unseen data. Essentially, the model becomes too tailored to the specific training set and fails to generalize to other examples.
Several techniques can be used to mitigate overfitting:
Increase Training Data
Simplify the Model
Regularization
Cross-Validation
Feature Selection
Early Stopping

confusion matrix

A confusion matrix in machine learning is a table that summarizes the performance of a classification model. It shows how often the model correctly and incorrectly predicted the classes in a dataset. This table allows for a deeper understanding of the model's strengths and weaknesses, going beyond simple accuracy metrics.

Learning rate

In machine learning, the learning rate is a hyperparameter that controls the step size during the optimization process. It determines how much the model's parameters are adjusted in response to the error calculated during training. Essentially, it dictates how quickly or slowly a model learns.

Validation set

In machine learning, a validation set is a portion of your dataset that's used to evaluate how well a model is performing during the training process, specifically to fine-tune its hyperparameters

and prevent overfitting. It's like a separate "check" to see if the model is generalizing well to unseen data, even before you get to the final testing phase.

Decision Tree Classification

A decision tree builds classification models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

It works by recursively partitioning the data based on feature values, aiming to create homogeneous subsets with respect to the target variable. This process continues until a stopping criterion is met, resulting in a tree structure with decision nodes (internal nodes) and leaf nodes (terminal nodes). Decision trees are used for both classification and regression tasks.

In machine learning, decision trees use splitting criteria to determine how to divide data at each node. Common criteria include Gini impurity, entropy, and information gain, all of which aim to minimize impurity (or maximize information) in the resulting child nodes after a split.

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning can occur in:

Top-down fashion. It will traverse nodes and trim subtrees starting at the root

Bottom-up fashion. It will begin at the leaf nodes

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a type of deep learning algorithm, particularly effective for analyzing visual data like images and videos. It excels at automatically extracting features from input data and is widely used in computer vision tasks like image classification, object detection, and image segmentation. CNNs utilize a special architecture with convolutional layers and pooling layers to process data and learn hierarchical representations of features.

How it Works:

1. Input:

Images, videos, or other data with a grid-like structure are fed into the CNN.

2. Convolutional Layers:

The input data is convolved with learned filters to produce feature maps.

3. Pooling Layers:

The feature maps are downsampled through pooling, reducing their dimensionality.

4. Repeat:

Steps 2 and 3 are repeated through multiple convolutional and pooling layers, extracting increasingly complex features.

5. Fully Connected Layers:

The extracted features are flattened and passed through fully connected layers for classification or other tasks.

Convolutional Neural Networks (CNNs) differ from regular neural networks primarily in their architecture and how they process data, particularly for tasks involving visual or spatial data. CNNs are specifically designed to automatically learn features from data, while traditional neural networks often rely on manually crafted features.

Computer vision
Computer vision, in the context of machine learning, is a field focused on enabling computers to "see" and interpret visual data like images and videos, mimicking human vision. It uses machine learning models to analyze visual information, identify objects, understand scenes, and even make predictions or take actions based on what it sees.

TensorFlow
TensorFlow is an open-source software library primarily used for machine learning and deep learning applications. It provides a flexible architecture for building and deploying models, particularly those involving numerical computation and data flow graphs.

PyTorch
PyTorch is an open-source machine learning framework based on the Python programming language, primarily used for building and training deep neural networks.

There is a **three-step process** followed to create a model:
Train the model
Test the model
Deploy the model

Missing values
One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.

Choose a Classifier Based on a Training Set Data Size
When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit. For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

False positives are those cases that wrongly get classified as True but are False.
False negatives are those cases that wrongly get classified as False but are True

Deep learning

The Deep learning is a subset of machine learning that involves systems that think and learn like humans using artificial neural networks. The term 'deep' comes from the fact that you can have several layers of neural networks. One of the primary differences between machine learning and deep learning is that feature engineering is done manually in machine learning. In the case of deep learning, the model consisting of neural networks will automatically determine which features to use (and which not to use).

Applications of supervised machine learning include:
Email Spam Detection
Healthcare Diagnosis
Sentiment Analysis
Fraud Detection

Bias

Bias in a machine learning model occurs when the predicted values are further from the actual values. Low bias indicates a model where the prediction values are very close to the actual ones.
Underfitting: High bias can cause an algorithm to miss the relevant relations between features and target outputs.

Variance

Variance refers to the amount the target model will change when trained with different training data. For a good model, the variance should be minimized.
Overfitting: High variance can cause an algorithm to model the random noise in the training data rather than the intended outputs.

Logistic Regression.

Logistic regression is a classification algorithm used to predict a binary outcome for a given set of independent variables.

K Nearest Neighbor Algorithm.

K nearest neighbor algorithm is a classification algorithm that works in a way that a new data point is assigned to a neighboring group to which it is most similar.
In K nearest neighbors, K can be an integer greater than 1. So, for every new data point, we want to classify, we compute to which neighboring group it is closest.

Cross-Validation

Cross-validation is a technique used in machine learning to assess how well a model generalizes to new, unseen data. It involves partitioning the available data into multiple subsets, using some

for training and others for validation, and repeating this process multiple times to get a more robust estimate of model performance. This helps to prevent overfitting, where a model performs well on training data but poorly on new data, and provides a more reliable measure of how the model will perform in real-world scenarios.