SUPERVISED LEARNING ALGORITHMS

working and uses

pdf by Zia Alam

Supervised Learning is training a model by data that is labeled or having the correct answer for the given features.

For example, you show it pictures of cats and dogs and say, "This is a cat, and that is a dog."

The computer looks at these examples and tries to figure out how to tell cats and dogs apart. It learns by finding patterns, like noticing that cats have pointy ears, and dogs have floppy ears.

Once the computer has learned enough, you can show it a new picture, and it will try to guess if it's a cat or a dog based on what it learned.

These algorithms are frequently used in a wide range of applications and are often the first choices for various supervised learning tasks.

Linear Regression
Logistic Regression
Decision Trees
Random Forest
Support Vector Machines (SVM)
k-Nearest Neighbors (k-NN)
Naive Bayes
Neural Networks
Gradient Boosting Machines
Lasso and Ridge Regression

⭐When to use which algorithm depends on the problem at hand, the data, and the desired outcomes.⭐

Linear Regression

Linear regression is a simple and widely used algorithm for regression tasks. It models the relationship between a dependent variable (the target) and one or more independent variables (features) by fitting a linear equation. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of squared differences between the predicted and actual values.

It is used for predicting continuous numeric values, such as predicting house prices based on features like square footage, number of bedrooms, and location.

Logistic Regression

Logistic regression is used for binary classification tasks. It models the probability that a given input belongs to a particular class using a logistic (S-shaped) function.

It estimates the likelihood of a binary outcome and is commonly used in scenarios like spam detection (spam or not spam), disease diagnosis (positive or negative), or customer churn prediction (churn or no churn).

Decision Tree

Decision trees are a versatile algorithm used for both classification and regression tasks. They create a tree-like structure by splitting the dataset into subsets based on the values of different input features. The decision tree learns to make decisions based on these splits. In classification, it assigns data points to different classes, while in regression, it predicts a continuous numeric value.

Decision trees are used in various fields, including finance for credit risk assessment and biology for species classification.

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

It builds multiple decision trees and combines their predictions to achieve a more robust and accurate model.

Random Forest is commonly used in tasks like recommendation systems, credit scoring, and image recognition.

Support Vector Machines (SVM)

SVMs are used for binary classification tasks and aim to find the optimal hyperplane that maximizes the margin between different classes in the data. They are particularly effective in situations with clear class separation. SVMs can handle non-linear data by transforming it into a higher-dimensional space using kernel functions.

They are used in applications like text classification, image recognition, and bioinformatics.

k-Nearest Neighbors (k-NN)

k-NN is a simple and intuitive algorithm used for both classification and regression.

It classifies or predicts data points based on the majority class or average value of their k nearest neighbors in the feature space.

k-NN is suitable for tasks like recommendation systems (e.g. YouTube recommendations), anomaly detection, and pattern recognition.

Naive Bayes

Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to make predictions.

It's particularly effective for text classification tasks, such as spam email detection, sentiment analysis, and document categorization.

It assumes that features are conditionally independent, which simplifies the calculations

It has the following 3 types GaussianNB, MultinomialNB, and BernoulliNB.

Neural Networks

Neural networks, especially deep neural networks, are a family of algorithms inspired by the structure of the human brain.

They consist of multiple interconnected layers of artificial neurons and are used for complex tasks such as image recognition, natural language processing, and speech recognition.

Deep learning models can learn hierarchical representations of data by adjusting the weights and biases of the neurons through backpropagation.

Gradient Boosting Machines

Gradient boosting is an ensemble learning technique that combines weak learners (typically decision trees) to create a strong learner. Algorithms like XGBoost and LightGBM are used for regression and classification tasks. They iteratively build a sequence of decision trees to correct the errors of previous trees, leading to improved predictive performance.

They are popular in Kaggle competitions and are applied in various fields, including finance, healthcare, and marketing

Lasso and Ridge Regression

Lasso and Ridge are regularization techniques used in linear regression to prevent overfitting and improve model generalization. Lasso adds a penalty term to the linear regression cost function, which encourages feature selection by driving some feature weights to zero. Ridge adds a penalty term that reduces the magnitude of feature weights.

These techniques are used to control model complexity and reduce the impact of irrelevant features.

Besides these algorithm, there are some other supervised learning algorithm which are not so often used, but they have their own use and advantages

Ordinary Least Squares (OLS)
Ridge Classifier
Stochastic Gradient Descent (SGD)
Perceptron, Elastic Net
Extreme Learning Machines (ELM)
Bayesian Networks
Conditional Random Fields (CRF)
Principal Component Analysis (PCA)
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis (QDA)
Multilayer Perceptrons (MLPs)
Radial Basis Function Networks (RBFN)
Gaussian Process Regression
Multi-layer Perceptron (MLP)
Categorical Naive Bayes
Adaptive Boosting (AdaBoost)
Genetic Algorithms