Machine Learning Algorithms: Definitions, Merits and Demerits

1. Linear Regression

Definition: Linear Regression is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. It predicts a continuous outcome by finding the best-fitting straight line through the data points.

Mathematical Form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

Merits:

- Simple to implement and efficient to train
- Overfitting can be reduced by regularization
- Performs well when the dataset is linearly separable

Demerits:

- Assumes that the data is independent which is rare in real life
- Prone to noise and overfitting
- Sensitive to outliers

2. Logistic Regression

Definition: Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable. Despite its name, it's a classification algorithm rather than a regression algorithm, used to predict the probability of an observation belonging to a certain class.

Mathematical Form: $P(y=1|X) = 1/(1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)})$

Merits:

- Less prone to overfitting but it can overfit in high dimensional datasets
- Efficient when the dataset has features that are linearly separable
- Easy to implement and efficient to train

Demerits:

- Should not be used when the number of observations are lesser than the number of features
- Assumption of linearity which is rare in practice
- Can only be used to predict discrete function (classification)

3. Decision Tree

Definition: A Decision Tree is a flowchart-like tree structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents an outcome. It splits the data into subsets based on the value of input features to make predictions.

Mathematical Basis: Uses criteria like Gini impurity or information gain to make optimal splits.

Merits:

- Can solve non-linear problems
- Can work on high-dimensional data with excellent accuracy
- Easy to visualize and explain

Demerits:

- Overfitting (might be resolved by random forest)
- A small change in the data can lead to a large change in the structure of the optimal decision tree
- Calculations can get very complex

4. K-Nearest Neighbor (KNN)

Definition: KNN is a non-parametric, lazy learning algorithm that classifies new cases based on a similarity measure (e.g., distance functions). It stores all available cases and classifies new cases by a majority vote of its k neighbors.

Mathematical Basis: Uses distance metrics (usually Euclidean) to find similar data points.

Merits:

- Can make predictions without training (lazy learner)
- Time complexity is O(n)
- Can be used for both classification and regression

Demerits:

- Does not work well with large datasets
- Sensitive to noisy data, missing values, and outliers
- Needs feature scaling

- Choosing the correct K value is challenging

5. K-Means Clustering

Definition: K-Means is an unsupervised learning algorithm that partitions data into K distinct clusters based on distance to the centroid of a cluster. The algorithm aims to minimize the within-cluster variance (sum of squares).

Mathematical Objective: Minimize the sum of squared distances from each point to its assigned cluster center.

Merits:

- Simple to implement
- Scales to large datasets
- Guarantees convergence
- Easily adapts to new examples
- Generalizes to clusters of different shapes and sizes

Demerits:

- Sensitive to outliers
- Choosing the k values manually is tough
- Dependent on initial values
- Scalability decreases when dimension increases

6. Support Vector Machine (SVM)

Definition: SVM is a supervised learning algorithm that finds the optimal hyperplane that maximally separates different classes in the feature space. For non-linear problems, it uses kernel functions to transform the data into higher-dimensional spaces.

Mathematical Basis: Maximizes the margin between classes while minimizing classification errors.

Merits:

- Good at high dimensional data
- Can work on small datasets
- Can solve non-linear problems through kernel tricks

Demerits:

- Inefficient on large data
- Requires picking the right kernel
- Can be computationally intensive

7. Principal Component Analysis (PCA)

Definition: PCA is a dimensionality reduction technique that transforms the data into a new coordinate system where the greatest variances lie on the first coordinate (principal component), the second greatest variance on the second coordinate, and so on.

Mathematical Basis: Finds eigenvalues and eigenvectors of the covariance matrix of features.

Merits:

- Reduces correlated features
- Improves performance by reducing dimensionality
- Reduces overfitting by removing noise

Demerits:

- Principal components are less interpretable than original features
- Information loss when components are discarded
- Must standardize data before implementing PCA

8. Naive Bayes

Definition: Naive Bayes is a probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. It calculates the probability of each class and the conditional probability of each class given each feature.

Mathematical Basis: $P(y|X) = P(X|y) \times P(y) / P(X)$

Merits:

- Training period is less compared to complex models
- Better suited for categorical inputs
- Easy to implement

Demerits:

- Assumes that all features are independent which is rarely happening in real life

- Zero Frequency problem (when a category variable has a category in test data that wasn't observed in training data)
- Estimations can be wrong in some cases due to the independence assumption

9. Artificial Neural Network (ANN)

Definition: ANNs are computing systems inspired by biological neural networks. They consist of layers of interconnected nodes (neurons) that process information using dynamic state responses to external inputs. Each connection has a weight that adjusts as learning proceeds.

Mathematical Basis: $y = f(\sum(\text{weight} \times \text{input}) + \text{bias})$, where f is the activation function.

Merits:

- Have fault tolerance
- Have the ability to learn and model non-linear and complex relationships
- Can generalize on unseen data

Demerits:

- Long training time
- Non-guaranteed convergence
- Black box (hard to explain solution)
- Hardware dependence
- Requires user's ability to translate the problem properly

10. AdaBoost

Definition: AdaBoost (Adaptive Boosting) is an ensemble learning method that combines multiple "weak classifiers" into a single "strong classifier." It adjusts the weights of instances after each iteration, focusing more on incorrectly classified instances.

Mathematical Basis: Weighted combination of weak learners where each learner's weight depends on its accuracy.

Merits:

- Relatively robust to overfitting
- High accuracy through ensemble approach
- Easy to understand and to visualize

Demerits:

- Sensitive to noisy data
- Affected by outliers
- Not optimized for speed

11. Gradient Boosting

Definition: Gradient Boosting is an ensemble technique that builds models sequentially, with each new model correcting errors made by the combined existing models. It uses gradient descent to minimize a loss function.

Mathematical Approach: Fits a new model to the residual errors made by the previous model.

Merits:

- Generally more accurate compared to other models
- Trains faster especially on larger datasets
- Most implementations provide support for handling categorical features
- Some implementations handle missing values natively

Demerits:

- Can overemphasize outliers and cause overfitting
- Requires many trees (>1000) which can be time and memory exhaustive
- More parameters to tune than simpler models

12. Random Forest

Definition: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

Mathematical Approach: Combines multiple decision trees through bootstrap sampling of data and random feature selection.

Merits:

- Capable of performing both classification and regression tasks
- Capable of handling large datasets with high dimensionality
- Enhances the accuracy of the model and prevents the overfitting issue

Demerits:

- Requires much computational power for large datasets
- Requires much time for training
- Can't describe relationships within data (lacks interpretability)

13. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Definition: DBSCAN is a density-based clustering algorithm that groups together points that are closely packed in regions of high density, while marking points in low-density regions as outliers.

Key Concepts: Core points, border points, and noise points based on density parameters.

Merits:

- Does not require a-priori specification of number of clusters
- Able to handle outliers within the dataset
- Able to find arbitrarily size and arbitrarily shaped clusters

Demerits:

- Struggles with clusters of similar density
- Struggles with high dimensionality data
- Sensitive to parameter settings (eps and min_samples)