ML

1. Neural Networks and Relevant Theory

A **neural network** is a computational model inspired by the structure of the human brain. It consists of interconnected layers of nodes (also called neurons), where each connection represents a weight. Neural networks are the foundation of deep learning, enabling computers to recognize patterns, such as images, audio, or text.

- **Architecture**: Neural networks typically have three types of layers:
  - **Input Layer**: Receives the input data.
  - **Hidden Layers**: Process the data through a series of transformations and activations.
  - **Output Layer**: Produces the final output (e.g., classification label or prediction).
- **Activation Functions**: Decide whether a neuron should be activated based on the weighted sum of inputs. Common functions include ReLU, sigmoid, and softmax.
- **Backpropagation**: A method used to train the network by adjusting weights in reverse order from output to input. It minimizes the error by calculating gradients.
- **Gradient Descent**: An optimization algorithm that updates weights to minimize the error or loss function.

Neural networks require large amounts of data and computational power, making them suitable for deep learning tasks.

---

2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, lazy learning algorithm used for classification and regression. KNN classifies data points based on the "k" closest training examples in the feature space.

- **How it Works**:

1. Calculate the distance (e.g., Euclidean, Manhattan) between the target point and all points in the training set.
2. Select the "k" closest neighbors.
3. For classification, the most common class among the neighbors is chosen. For regression, the average of the neighbors' values is taken.

- **Advantages**: Simple and interpretable, no training phase.
- **Disadvantages**: High computational cost for large datasets, affected by irrelevant features and outliers.

---

3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression. SVM aims to find the optimal hyperplane that separates classes with the maximum margin.

- **Key Concepts**:
  - **Hyperplane**: A decision boundary separating different classes.
  - **Margin**: The distance between the hyperplane and the nearest data points from each class (support vectors).
  - **Kernel Trick**: For non-linear data, SVM can apply kernel functions (like polynomial or radial basis functions) to project data into higher dimensions, where it becomes linearly separable.

SVM is effective in high-dimensional spaces and is commonly used for text classification, image recognition, and more.

---

4. K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm used to partition a dataset into "k" clusters based on feature similarity.

- **How it Works**:
  1. Initialize "k" centroids randomly.
  2. Assign each data point to the nearest centroid.

3. Update centroids by computing the mean of all points assigned to each cluster.
4. Repeat steps 2 and 3 until centroids no longer change significantly.

- **Applications**: Image compression, customer segmentation, and anomaly detection.
- **Limitations**: Requires the number of clusters "k" to be predefined, sensitive to initial centroids and outliers.

---

5. Elbow Method

The **Elbow Method** helps determine the optimal number of clusters (k) in K-means clustering.

- **How it Works**:
  - For each "k", calculate the sum of squared distances between each point and its centroid (within-cluster sum of squares).
  - Plot the cost against the number of clusters.
  - Look for the "elbow" point where adding more clusters results in a smaller decrease in cost, suggesting that this is the optimal "k."

This method provides a visual approach to finding the ideal number of clusters, though interpretation can be subjective.

---

6. Normalization and Standardization

- **Normalization**: Rescales features to a range, often [0, 1]. It’s commonly used when the data has different ranges.Xnormalized=Xmax−XminX−Xmin

  Xnormalized=X−XminXmax−XminX_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}

- **Standardization**: Centers data around the mean with a standard deviation of 1, creating a distribution with mean = 0 and variance = 1. It's useful when features have different units.Xstandardized=σX−μ

  Xstandardized=X−μσX_{\text{standardized}} = \frac{X - \mu}{\sigma}

Both techniques are important for algorithms sensitive to feature scaling, like KNN, SVM, and neural networks.

---

7. Haversine Distance Calculation

Haversine Distance is used to calculate the great-circle distance between two points on a sphere, particularly useful for geographic coordinates (latitude and longitude).

- **Formula**:
  $d = 2r \cdot \arcsin(\sqrt{\sin^2(2\Delta lat) + \cos(lat1) \cdot \cos(lat2) \cdot \sin^2(2\Delta lon)})$
  where:

  $$d = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta \text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta \text{lon}}{2}\right)}\right)$$

  - $\Delta lat = lat2 - lat1$ $\Delta \text{lat} = \text{lat}_2 - \text{lat}_1$
  - $\Delta lon = lon2 - lon1$ $\Delta \text{lon} = \text{lon}_2 - \text{lon}_1$
  - r is the radius of the Earth, approximately 6,371 km.

---

8. Model Evaluation Metrics

- **Accuracy**: Percentage of correctly predicted instances over total instances. Not ideal for imbalanced datasets.
- **Precision**: Ratio of correctly predicted positive instances to total predicted positives.
- **Recall (Sensitivity)**: Ratio of correctly predicted positives to all actual positives.
- **F1 Score**: Harmonic mean of precision and recall, balancing the two.
- **ROC-AUC**: Measures the area under the receiver operating characteristic curve, indicating model performance across classification thresholds.

These metrics help evaluate a model's performance and are chosen based on the context and balance of the dataset.

---

9. Outliers and Detection Methods

Outliers are data points that deviate significantly from other observations and may indicate noise, data entry errors, or anomalies.

- **Z-Score Method**: Calculates the standard deviation from the mean for each data point. Points with a Z-score above a threshold (often 3) are considered outliers.
  $Z = \sigma X - \mu$

  $Z = X - \mu \sigma Z = \frac{X - \mu}{\sigma}$

Other methods include IQR (Interquartile Range) and DBSCAN clustering.

---

10. Correlation Matrix

A **Correlation Matrix** shows the correlation coefficients between pairs of variables in a dataset, helping identify relationships between them. The values range from -1 to 1:

- **+1**: Strong positive correlation.
- **0**: No correlation.
- **1**: Strong negative correlation.

This matrix is useful in feature selection and to understand multicollinearity between variables in a dataset.

- **Bias:** Error introduced by making simplifying assumptions about the model. High bias models underfit the training data.
- **Variance:** Sensitivity of a model to small fluctuations in the training data. High variance models overfit the training data.
- **Generalization:** A model's ability to perform well on unseen data. Good models balance bias and variance to generalize well.

Underfitting and Overfitting

- **Underfitting:** A model that is too simple to capture the underlying patterns in the data.

- **Overfitting:** A model that is too complex and fits the noise in the training data rather than the underlying pattern.

Linear Regression

- A statistical method to model the relationship between a dependent variable and one or more independent variables.
- The goal is to find the best-fitting line through the data points.

Regression: Lasso, Ridge

- **Lasso Regression:** A regularization technique that adds a penalty term to the loss function to shrink the coefficients of less important features.
- **Ridge Regression:** Another regularization technique that adds a penalty term to the loss function to reduce the magnitude of coefficients.

Gradient Descent Algorithm

- An optimization algorithm used to find the minimum of a function.
- It iteratively adjusts the parameters of a model in the direction of the steepest descent.

Evaluation Metrics

- **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual values.
- **Root Mean Squared Error (RMSE):** Square root of the average squared difference between predicted and actual values.
- **R-squared (R^2):** Proportion of the variance in the dependent variable explained by the independent variables.

- **K-Nearest Neighbors (KNN):** Classifies new data points based on the majority class of their K nearest neighbors.
- **Support Vector Machines (SVM):** Finds the optimal hyperplane to separate data points into different classes.

Ensemble Learning

- **Bagging (Bootstrap Aggregating):** Combines multiple models (e.g., decision trees) trained on different subsets of the training data.
- **Boosting:** Sequentially trains models, focusing on misclassified examples from previous iterations.
    - **Adaboost:** A popular boosting algorithm that assigns weights to training examples, giving more weight to misclassified examples.
- **Random Forest:** An ensemble of decision trees, where each tree is trained on a random subset of features and samples.

Binary vs. Multiclass Classification

- **Binary Classification:** Classifies data into two classes (e.g., spam or not spam).
- **Multiclass Classification:** Classifies data into more than two classes (e.g., different types of animals).

Balanced and Imbalanced Multiclass Classification

- **Balanced:** Each class has roughly the same number of samples.
- **Imbalanced:** One or more classes have significantly fewer samples than others.

Variants of Multiclass Classification

- **One-vs-One:** Trains binary classifiers for each pair of classes.
- **One-vs-All:** Trains a binary classifier for each class against all other classes.

Evaluation Metrics

- **Accuracy:** Proportion of correct predictions.
- **Precision:** Proportion of positive predictions that are actually positive.
- **Recall:** Proportion of actual positive cases that are correctly identified.
- **F1-score:** Harmonic mean of precision and recall.
- **Cross-Validation:** A technique to assess model performance by dividing the data into training and testing sets multiple times.
- **Micro-Average:** Calculates metrics globally across all classes.

- **Macro-Average:** Calculates metrics for each class individually and then averages them.

Clustering Algorithms

- **K-Means:** Partitions data into K clusters based on the mean of each cluster.
- **K-Medoids:** Similar to K-Means, but uses medoids (actual data points) as cluster centers.
- **Hierarchical Clustering:** Creates a hierarchy of clusters, either by agglomerative (bottom-up) or divisive (top-down) approach.
- **Density-Based Clustering (DBSCAN):** Groups together points that are closely packed together.
- **Spectral Clustering:** Converts data into a similarity graph and performs spectral analysis to identify clusters.

Outlier Analysis

- **Isolation Forest:** Isolates outliers by randomly selecting features and splitting data until each data point is isolated.
- **Local Outlier Factor (LOF):** Compares the local density of a data point to the density of its neighbors.

Evaluation Metrics

- **Elbow Method:** Used to determine the optimal number of clusters in K-Means by plotting the within-cluster sum of squares (WCSS) against the number of clusters.
- **Intrinsic Evaluation:** Assesses the quality of clustering without external labels.
    - Silhouette Coefficient: Measures how similar a data point is to its own cluster compared to other clusters.
    - Calinski-Harabasz Index: Measures the ratio of the sum of between-clusters dispersion and within-cluster dispersion.
- **Extrinsic Evaluation:** Assesses the quality of clustering using external labels.
    - Confusion Matrix: Shows the number of correct and incorrect predictions.

- Accuracy, Precision, Recall, F1-score: Standard classification metrics.

Artificial Neural Networks (ANNs)

ANNs are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes called neurons, which process information.

- **Single-Layer Perceptron:** A simple ANN with a single layer of neurons. It can only classify linearly separable data.
- **Multilayer Perceptron (MLP):** A more complex ANN with multiple layers of neurons. It can learn complex patterns and make nonlinear decisions.
- **Backpropagation Learning:** A supervised learning algorithm used to train MLPs. It involves calculating the error at the output layer and propagating it back through the network to adjust the weights and biases.
- **Functional Link Neural Network:** A type of ANN that uses nonlinear transformations to increase the input space dimensionality, making it capable of learning complex patterns.
- **Radial Basis Function (RBF) Network:** An ANN that uses radial basis functions as activation functions. It is well-suited for interpolation and function approximation tasks.

Activation Functions

Activation functions introduce nonlinearity into ANNs, enabling them to learn complex patterns. Common activation functions include:

- **Sigmoid:** Maps input values to a range of 0 to 1.
- **Tanh:** Maps input values to a range of -1 to 1.
- **ReLU (Rectified Linear Unit):** Maps negative input values to 0 and positive values to themselves.

Recurrent Neural Networks (RNNs)

RNNs are designed to process sequential data, such as time series or text. They have feedback connections that allow them to remember past information, making them suitable for tasks like language modeling and speech recognition.

Convolutional Neural Networks (CNNs)

CNNs are specifically designed for image and video recognition tasks. They use convolutional layers to extract features from the input data and pooling layers to reduce the dimensionality of the feature maps.