Machine Learning Internship
Assignment Report

Assignment Title: Algorithm Comparison

Submitted by: Sai Manikanta

Domain: Machine Learning

Date: May 2025

Assignment: Algorithm Comparison

Name: Sai Manikanta

Assignment: Algorithm Comparison

Part 1: Algorithm Overview

Logistic Regression

Logistic Regression is a classification algorithm that estimates the probability of a binary outcome using a logistic (sigmoid) function. It models the relationship between input features and the log-odds of the target class.

Strengths:

- Easy to implement and interpret.
- Performs well with linearly separable data.

Limitations:

- Struggles with non-linear relationships.
- Sensitive to multicollinearity among features.

K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies a data point based on how its neighbors are classified. It calculates distances (like Euclidean) to find the 'k' closest points and uses majority voting.

Strengths:

- Simple and intuitive.
- No training phase (lazy learning).

Limitations:

- Computationally expensive for large datasets.
- Sensitive to noisy or irrelevant features.

Decision Tree

Decision Trees split data based on feature values to form a tree structure, aiming to create the purest class subsets. It uses metrics like Gini Index or Information Gain for splits.

Strengths:

- Easy to visualize and interpret.
- Handles both categorical and numerical data.

Limitations:

- Prone to overfitting.
- Unstable-small data changes can change the tree.

Support Vector Machine (SVM)

SVM finds the optimal hyperplane that best separates data into classes, maximizing the margin between support vectors. It can also handle non-linear data using kernel tricks.

Strengths:

- Effective in high-dimensional spaces.
- Works well with clear margin separation.

Limitations:

- Computationally intensive with large datasets.
- Requires careful tuning of kernel and parameters.

Part 2: Application Scenarios

High-Dimensional Data

Recommended Algorithm: Support Vector Machine (SVM)

SVM performs exceptionally well in high-dimensional spaces due to its ability to create optimal decision boundaries even with many features. It is also effective when the number of features exceeds the number of samples, making it suitable for text classification and genetic datasets.

Imbalanced Dataset

Recommended Algorithm: Logistic Regression

Logistic Regression can be adapted to handle imbalanced data using techniques like class weighting or resampling. It provides probabilistic outputs and is interpretable, making it ideal for sensitive applications like

fraud detection and disease prediction.

Small Dataset with Many Features

Recommended Algorithm: SVM

SVM handles high-dimensional spaces well and is effective with small datasets, especially when paired with kernel tricks. Its margin-based optimization reduces overfitting risk, which is critical in domains like healthcare with limited samples and many features.

Non-linear Data Separation

Recommended Algorithm: SVM (with non-linear kernels)

For datasets with non-linear boundaries, SVM with kernels like RBF or polynomial can capture complex patterns effectively. Its ability to transform data into higher dimensions allows it to create clear boundaries even with intricate data distributions.

Dataset with Noise

Recommended Algorithm: Decision Tree

Decision Trees can ignore irrelevant features during splitting and are inherently robust to noise to some extent. While they may still overfit, they offer interpretability and can be pruned to improve generalization on noisy datasets.