ML assignment-1

Part 1 – Algorithm overview

For each of the algorithms below, write a brief overview that includes:

- How the algorithm works.
- Two key strengths and two limitations.

1. Logistic Regression

How the algorithm works.

Logistic regression is used to solve binary classification problems.

It models the probability of a given input belongs to a certain category.

It uses logistic function like sigmoid to output values between 0 and 1.

For example, to predict the class of given fruit from different classes of fruits.

Strength

Efficiency: Logistic Regression is fast to train even on large datasets

Less prone to over fitting: with regularization it is less prone to overfitting

limitations

Assumes linearity: not suitable for liner data

Sensitive to outliers: as it's a linear model, extreme values skew the model learned weights.

2. K-Nearest Neighbours (KNN)

How the algorithm works.

K-Nearest Neighbours is used for both classification and regression.

It finds k closest neighbours in the feature space.

It makes decision based on majority classes(classification) or average value (regression).

KNN makes no assumption about data distribution.

Strength

Simple: Easy to understand and implement

Works well with non-linear data: handle non-linear decision boundaries, does not rely on underlying model.

limitations

choice of K: performance depends on choice of k, poorly chosen values lead to underfitting and overfitting

computationally expensive: compares new data against every point to make prediction.

3. Decision Tree

How the algorithm works.

Decision Tree is used for both classification and regression.

It recursively split the dataset into subsets based on feature values, creating a tree like model of decision

Root node represents the full dataset, branches represents decision based on features, leaf node represents predicted class label or numeric value.

Strength

Visual and interpretable: model can be visualized graphically

Capture non-linear relations: model complex data and non-linear relations by repeatedly splitting the data based on features.

limitations

prone to overfit: easily overfit, if grow deep with many branches. Pruning is required to mitigate this.

Hard to optimize: only looks at immediate improvements, greedy nature limits it to find global optimal tree.

4. Support Vector Machine

How the algorithm works.

SVM finds the optimal hyperplane that best separate the data points of different classes in feature space.

The hyperplane maximizes the margin between the nearest data points in every class.

It finds the line or plane or hyperplane that separates the data with largest gap possible.

If data in not linearly separable, it uses the kernel trick to project it to higher dimension, where a gap might exist which separates the data.

Strength

Handle high dimensional data: performs well when number of features is large

Less prone to overfit: with proper tuning, it achieves unique and optimal solution.

limitations

expensive to train: complexity increase significantly with high number of samples.

Difficult to tune and interpret: involves several hyper parameters, which must be carefully tuned. Incorrect choice will slow down the performance.

Part 2: Application Scenarios

For each of the following dataset scenarios, recommend the most suitable algorithm (Logistic Regression, KNN, Decision Tree, or SVM).

Provide a brief explanation for your choice.

1. High-Dimensional Data (e.g., text or gene expression data)

Ans. SVM is most effective in high dimensional data, SVM finds maximum margin hyperplane using support vectors. The kernel trick allows it map features into higher dimensional space without requiring any transformation to the data. SVM handles sparse data well.

2. Imbalanced Dataset (e.g., fraud detection, rare disease prediction)

Ans. Logistic regression is good algorithm to handle imbalanced datasets. As this algo can directly output the probabilities. Which can be tuned during the training, for making it more sensitive towards under-represented data (here can be fraud scenarios or rare disease conditions).

3. Small Dataset with Many Features (e.g., medical or genetic data)

Ans. Logistic Regression is also work good in small dataset when there are linear relationships between features and outputs. With regularization it can reduce dimensions and which increase data interpretability which can be useful in finding issues in medical data/images.

4. Non-linear Data Separation (e.g., complex shapes like spirals or circles)

Ans. For non-linear data separation, SVM is also best choice, as it can find the hyperplane with large gap in higher dimension. As it can use kernel trick which transforms data into higher dimension, where a boundary can be found with sufficient gap, so it can separate the data and extract information.

5. Dataset with Noise (e.g., data with many irrelevant or misleading features)

Ans. Easy to apply, Random forest is good choice, as decision tree perform feature section based on variables which provides most information gain and split at that point. This allows them to ignore other not important and noisy features.