Supervised Learning

Enrique Alifio Ditya

Questions and Answers

What is Supervised Learning?

Supervised learning is a branch of machine learning that deals with learning a function from labeled training data. The goal of supervised learning is to approximate the mapping function from input variables to output variables, given a training dataset consisting of input-output pairs. Supervised learning tasks are largely categorized into two types: classification and regression. Classification tasks are tasks where the output variable is a category, such as True or False. Regression tasks, on the other hand, are tasks where the output variable is a real value, such as prices.

How does each of the implemented algorithms work?

Most machine learning algorithms (including one implemented here) require the dataset to be of numerical values; therefore, the data needs to be preprocessed accordingly before being fed to the model. The following is a step-by-step explanation of how each of the implemented algorithms works:

K-Nearest Neighbors

1. Initialize the number of neighbors (K)
2. Loop over all the samples of the dataset
3. For each sample, calculate its distance to all other samples in the dataset
4. Sort the distances in ascending order
5. Select the top K samples with the smallest distance
6. Assign the class of the sample to the class that is most frequent among the K samples

Logistic Regression

1. Initialize the weights and bias randomly
2. Initialize the learning rate and number of epochs
3. Loop over the number of epochs
4. Calculate the linear combination of the weights and the input features
5. Pass the linear combination to the sigmoid function
6. Update the weights and bias using gradient descent for the chosen loss function

Iterative Dichotomiser 3 (ID3)

1. Initialize the root node
2. Set the base case for the recursion, that is:
    - If all samples in the node belong to the same class, return the node
    - If the node contains no features, return the most frequent class in the node

3. Find the best feature and threshold to split the data:
    - For each feature, find all possible thresholds to split the data (midpoint between two consecutive values)
    - For each threshold, calculate the information gain:
        - Calculate the parent entropy
        - Calculate the child entropy for the left and right splits
        - Calculate the weighted average of the child entropy
        - Calculate the information gain by subtracting the child entropy from the parent entropy
4. Split the data into two nodes based on the selected feature and threshold
5. Recursively call the function on the two nodes

Decision Tree

The decision tree generally works in a similar fashion to ID3, with the difference being that decision trees can handle numerical data, while ID3 can only handle categorical data. The implementation of the decision tree in this repository uses the CART algorithm, which is a generalization of ID3 that can handle numerical data.

1. Initialize the root node
2. Set the base case for the recursion, that is:
    - If all samples in the node belong to the same class, return the node
    - If the node contains no features, return the most frequent class in the node
3. Find the best feature and threshold to split the data:
    - For each feature, find all possible thresholds to split the data (midpoint between two consecutive values)
    - For each threshold, calculate the Gini impurity:
        - Calculate the parent Gini impurity
        - Calculate the child Gini impurity for the left and right splits
        - Calculate the weighted average of the child Gini impurity
        - Calculate the Gini impurity by subtracting the child Gini impurity from the parent Gini impurity
4. Split the data into two nodes based on the selected feature and threshold
5. Recursively call the function on the two nodes

What are the advantages and disadvantages of each of the implemented algorithms?

K-Nearest Neighbors

Advantages:

- Simple to implement
- Can be used for both classification and regression tasks
- Can be used for multi-class classification tasks
- Performs well on non-linear decision boundaries

Disadvantages:

- Computationally expensive, with $O(n^2)$ time complexity
- Scales poorly with the number of features
- Requires a lot of memory to store the entire dataset
- Sensitive to noisy data, since it uses the entire dataset to make predictions
- Requires feature scaling, since it uses the Euclidean distance to calculate the distance between samples

Logistic Regression

Advantages:

- Simple to implement
- Fast to train
- Outputs probabilities
- Can be used for both binary and multi-class classification tasks (depending on the loss function)
- Performs well on linear decision boundaries

Disadvantages:

- Requires feature scaling, since gradient calculations are sensitive to the scale of the input features
- Poorly handles non-linear decision boundaries
- Sensitive to noisy data, since it uses the entire dataset to make predictions

Iterative Dichotomiser 3 (ID3)

Advantages:

- Simple to implement
- Outputs a decision tree that is easy to interpret
- Can handle multi-class classification tasks

Disadvantages:

- Cannot handle numerical data
- Scales poorly with the number of features
- Prone to overfitting, since it can create a large number of branches

Decision Tree

Advantages:

- Simple to implement
- Outputs a decision tree that is easy to interpret
- Can handle numerical data
- Can handle multi-class classification tasks

Disadvantages:

- Prone to overfitting, since it can create a large number of branches. Can be mitigated by pruning the tree or with ensemble methods such as random forest.
- Proven to be NP-complete, meaning that it is computationally expensive to find the optimal tree

What are the applications of supervised algorithms in real life?

Supervised learning algorithms are used in a wide variety of applications, such as:

- Image classification in detecting cancer cells
- Speech recognition in virtual assistants to classify speech into text
- Spam detection in email services to classify emails into spam or not spam
- Medical diagnosis in detecting breast cancer from mammograms
- Credit scoring in banks to classify customers into low-risk or high-risk
- Sentiment analysis in social media to classify text into positive or negative
- And many more!