CS771 Assignment 2

Introduction To Machine Learning

Submitted by

Jigyashu Garg (200478)
Manish Meena (200560)
Saurav Kumar (200906)
Harsh Saroha (200419)
Deepesh Pratap (200313)

April 5, 2023

PROBLEM 1

Give detailed calculations explaining the various design decisions you took to develop your decision tree algorithm. This includes the criterion to choose the splitting criterion at each internal node (which essentially decides the query word that Melbo asks when that node is reached), criterion to decide when to stop expanding the decision tree and make the node a leaf, any pruning strategies and hyperparameters etc.

Answer : The decision tree algorithm designed is a recursive binary tree-based algorithm that splits the data based on certain criteria at each internal node until a certain stopping criterion is met. The following are the key design decisions and calculations involved in developing this algorithm:

Split Criterion: Two splitting criteria were used at different levels in this algorithm:

- Split by length: This criterion was used only at the root node of the tree. It splits the data based on the length of the words in the data. All words of the same length are grouped together and form a node in the tree. This splitting criterion is used at the root node to take advantage of the fact that most English words are of a certain length range, so it can help to reduce the search space and make the tree smaller and more efficient.

- Split by overlap: This criterion was used at all other levels of the tree. It chooses a character from the word at the current node and splits the data into two groups based on whether that character is present in that position in the words in the group. This criterion is used at other levels to find the best attribute to split on, where the best attribute is the one that results in the lowest entropy, to calculate entropy, the suboptimal split method was used, iterated through each attribute and tried to find the attribute that results in the minimum entropy. For each attribute, we looped over each sample in the current node and compare the attribute's value with the sample's value. If the value matches, we add the sample to the corresponding split. Computed the entropy for each split and add it to the overall entropy for that attribute. Finally, return the attribute with the minimum overall entropy.

  $Entropy = -\sum_{i=1}^{n} p_i log_2 p_i$

where n is the number of classes, and p_i is the probability of an item belonging to class i.

Stopping Criterion: The tree-building process is stopped when one of the following stopping criteria is met:

- Minimum Leaf Size: If the number of data points in a node is less than or equal to the minimum leaf size, the node is made a leaf node, and no further splitting is performed. This is to prevent overfitting and reduce the size of the tree.
- Maximum Depth: If the depth of the tree reaches the maximum depth, the node is made a leaf node, and no further splitting is performed. This is to prevent overfitting and reduce the size of the tree.
- Purity: If all the data points in a node belong to the same class, the node is made a leaf node, and no further splitting is performed. This is to prevent overfitting and reduce the size of the tree.

Hyperparameters: The algorithm has two hyperparameters:

- Minimum leaf size (min_leaf_size): determines the minimum size of the dataset that can form a leaf node.
- Maximum depth (max_depth): determines the maximum depth of the decision tree.

The design decisions for this decision tree algorithm are primarily based on a trade-off between accuracy and computational efficiency. For example, the use of entropy-based splitting criteria and stopping criterion based on purity can lead to more accurate models, but this can be computationally expensive for large datasets. On the other hand, the use of minimum leaf size and maximum depth can make the algorithm computationally efficient but may lead to lower accuracy.