CS771 Assignment 2

Geetika
Dept. of Comp. Science and Engineering
geetika21@iitk.ac.in

Apoorva Gupta
Dept. of Comp. Science and Engineering
apoorvag21@iitk.ac.in

Girik Maskara
Dept. of Comp. Science and Engineering
girikm20@iitk.ac.in

Amrit Kumar
Dept. of Materials Science and Engineering
amritkr@iitk.ac.in

Akshat Goyal
Dept. of Materials Science and Engineering
akshatgo@iitk.ac.in

Sourabh Mundhra
Dept. of Chemical Engineering
msourabh@iitk.ac.in

Abstract

This document is the submission of group Tompers Squad for Assignment 2.

Preprint. Under review.

1 Part 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 to Part 1:

Through our solution code, we have presented a decision tree algorithm which we built by initial experiments to maximize information gain using entropy and the support of a Gini Index. To calculate the values of these variables, the probabilities were computed as the number of same-length words to the total number of words in the split at a specific node.

$$p_i = \frac{\text{Number of words of the same word length}}{\text{Total number of words in the split at the node}}$$

$$\text{Entropy} = -\sum_{i=0}^{n} p_i \log_2 p_i$$

$$\text{Gini Index} = 1 - \sum_{i=0}^{n} (p_i)^2$$

The training time, model size, win rate and the average number of queries made can be found in the table below.

Table 1: Comparison of models trained using Gini Index and Entropy

Variable	Training Time (in sec)	Model Size (Bytes)	Win Rate	Avg. Queries
Entropy	9.782	1120812	1.0	4.985
Gini Index	9.527	1120812	1.0	4.985

Initial experiments yielded similar outputs for both the variables as observable. To further refine our experimentation, we employed a combination of the two variables to enable splits at each node of the tree. Results for this experimentation indicated no further need to optimize the average number of queries as it was safely placed in the aforementioned limits as stated in the problem statement.

Table 2: Model trained using both Gini Index and Entropy

Variable	Training Time (in sec)	Model Size (Bytes)	Win Rate	Avg. Queries
Both (Gini & Entropy)	7.980	1056819	1.0	4.215

Using both the variables simultaneously gives a better model with slightly smaller filesize and significant decrease in training time. Thus, the decision to choose the above model as the solution.

The decision tree stops expanding once the number of queries reach their limit i.e., 15 queries per round or if the split at the node has less words than the minimum leaf size (taken 1).

2 Part 2

Code submitted.