Indian Institute of Technology, Kharagpur

CS60050 - Machine Learning

Report of Assignment – 1

Decision Tree

Group Assigned : E

Group Members : Tamal Kanti Baksi – 22CS60R60 Pawan Singh Koranga – 22CS60R25

Date of Submission : 14th September 2022

Tasks For Assignment :

1. For our dataset, we have to Randomly divide Dataset E into 80% for training and 20% for testing, We have to Build a **Decision Tree Classifier** using ID3 algorithm and Train the classifier **using Information Gain (IG)** measure.

2. We have to Repeat (1) for 10 random splits and Print the best test accuracy and the depth of that tree.
3. Perform **reduced error pruning** operation over the tree obtained in (2). Plot a graph showing the variation in test accuracy with varying depths. Print the pruned tree obtained in hierarchical fashion with the attributes clearly shown at each level.

   Note the following points while creating the decision tree:

   a. Randomly split the feature matrix into train split and test split with 80- 20 ratio.
   b. To read data from dataset_E.csv, you can use the python package pandas.
   c. No packages to be used for Decision Tree Classifier.

Dataset Information :

The Dataset contains 1 file as follows :

Dataset_E.csv - Labelled dataset of 768 lines, with each line storing Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and the Outcome.

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The dataset contained a total of 9 columns as follows:

1. Pregnancies - Number of times pregnant
2. Glucose - Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. BloodPressure - Diastolic blood pressure (mm Hg)
4. SkinThickness - Triceps skin fold thickness (mm)
5. Insulin- 2-Hour serum insulin (mu U/ml)
6. BMI - Body mass index (weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction- Diabetes pedigree function
8. Age- Age (years)
9. Outcome- Class variable (0 or 1) 268 of 768 are 1, the others are 0

There were a total of 768 examples in the dataset and no examples had any of the attribute literals missing.

Implementing a Decision Tree Classifier :

1. Using Information Gain

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where *Values(A)* is the set of all possible values for attribute A, and *S*, is the subset of S for which attribute A has value v.

Explanation of procedure and Code implementation

1. Code Implementation Explanation :

Decision Tree –

1. We have split the provided dataset (Dataset_E.csv) into two parts- 80% as the training set and 20% as the testing set.
2. Then we have built the Decision tree Classifier following the **ID3 Algorithm**. Subsequently, we have trained the same using the highest **Information Gain** measures of the given features.
3. We have selected the training and the testing set randomly for 10 iterations following the above splitting measures. Each time we have trained our classifier followed by a testing with the remaining dataset. After each iteration we have measure the accuracy comparing with the outcome.
4. We have done reduced error pruning operation over the tree obtained and have done variation in test accuracy with varying depths.

Results –

Results are shown in separate output file.

2. Conclusion :

➢ This assignment used all the concepts about the Decision Tree Classifier.

➢ We implemented the Decision tree Classifier by using information gain using python.

➢ This concludes our report on first assignment of Machine learning based on decision trees.