Decision Tree Algorithm Assignment

K.N.Toosi University of Technology
Introduction to Data Mining

Fall 2024

Part I
Practical Assignment

Dataset

Age	Income	Married	Buys
22	High	No	No
35	Low	Yes	Yes
25	Medium	No	Yes
45	Medium	Yes	Yes
50	High	Yes	Yes
30	Low	No	No
40	High	No	Yes
20	Low	No	No
50	Low	Yes	Yes
35	Medium	No	Yes

Table 1: binary customer data

Task

Create a decision tree using the dataset above. Determine the best criteria for splitting the data at each node and illustrate the resulting tree. Discuss the process you followed and the reasoning behind your decisions.

Part II

Practical Assignment

Dataset

Bedrooms	Square_Feet	House_Age	Price
3	1500	10	300000
4	2000	5	450000
2	900	30	150000
3	1800	20	350000
5	2500	8	550000
4	2200	15	500000
2	1200	25	200000
3	1600	10	310000
4	2100	7	480000
5	3000	3	600000

Table 2: house pricing data

Task

Create a regression tree using the dataset above. Determine the best criteria for splitting the data at each node and illustrate the resulting tree. Discuss the process you followed and the reasoning behind your decisions.

Performance Metric

For regression trees, use the Mean Squared Error (MSE) to measure performance. The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value.

Part III

Implementation Assignment

Dataset

The dataset for this assignment is the COVID-19 dataset available on Kaggle: COVID-19 Dataset.

Task

Implement a decision tree classification algorithm in Python and compare it to the Sklearn implementation. Unlike Sklearn, do not use one-hot encoding; manually split multi-label categorical features. Compare your performance with the Sklearn decision tree. Performance should be measured via F1-score.

F1-score

The F1-score is the harmonic mean of precision and recall. It is calculated as follows:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

where:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Note

Any attempt to use AI tools for generating the code is strictly prohibited. Students will be asked to present and explain their code during a class session.