EE708: Fundamentals of Data Science and Machine Intelligence

Reading Assignment

Based on Decision Trees

1. A dataset contains 200 samples classified into two classes: 120 positive and 80 negative.
   a. Compute the Gini index before splitting.
   b. If a split results in subsets:
      Left: (50 positive, 10 negative)
      Right: (70 positive, 70 negative)
      Compute the weighted Gini index and determine whether the split improves purity.
2. Consider the given dataset with two independent variables (x_1, x_2) and one dependent variable (y):
   a. Use the sum of squared errors (SSE) to determine the best splitting point for x_1.
   b. Construct the first split of a regression tree using SSE as the impurity measure.

x_1	x_2	y
1	5	10
2	6	12
3	8	15
4	10	18
5	12	21
6	15	25
7	18	28
8	20	30

3. You are analysing past movie data to predict whether a film will be a hit (H = +) or flop (H = –) based on:
   Budget (B): 1 (High) or 0 (Low)
   Reviews (R): 1 (Positive) or 0 (Negative)
   Build the best decision tree using B and R to predict success.

B	R	H	Count
0	0	–	5
0	0	+	1
0	1	–	0
0	1	+	4
1	0	–	3
1	0	+	1
1	1	–	2
1	1	+	0

4. Consider the following data set:

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	yes
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

   a. Considering "profitable" as the binary values attribute we are trying to predict, which of the attributes would you select as the root in a decision tree with multi-way splits using the cross-entropy impurity measure?
   b. How would you decide on the next split using the same impurity measure?

5. Write a code to obtain a fully grown regression tree for the data given in Q2 and visualize the regression tree.
6. Write a code to obtain a fully grown classification tree for the data given in Q4 and visualize the classification tree. Verify your answers in Q4.
7. Binary classification tree:
    a. Train a fully grown binary classification tree based on Gini impurity using the dataset *A4_train.csv* and visualize it.
    b. Compute the Sum of Squared Errors (SSE) on the test dataset (*A4_test.csv*) at each depth and plot the variation of SSE with depth.
    c. Determine the optimal pruning depth by selecting the depth where the SSE change is minimal.
    d. Visualize the pruned tree.