CS5402 Homework – 2

(Deadline: 03-02-2025 11:59pm)

Instructor: Huiyuan Yang

- ❑ Theory questions [45%] :

Q1. [20%] Given the confusion matrix for 5 classes as below:

1. What is the Accuracy and F1-score for class c_1, c_2, c_3, c_4, c_5?

(**Please provide detailed steps, incldluing the process of converting a multi-class confusion matrix into a binary confusion matrix, as well as the calculations for precision and recall**)

Ground Truth \ Predicted label	c_1	c_2	c_3	c_4	c_5
c_1	52	3	7	2	2
c_2	2	28	0	2	3
c_3	5	2	25	7	7
c_4	1	3	6	40	5
c_5	1	2	1	3	17

Q2: [25%] Given the following sample dataset that represents whether or not to play tennis based on different features:

- Calculate the Information Gain for each feature: $(G(D, Outlook),\ G(D, Temperature),\ G(D, Humidity),\ G(D, Wind)$.
- Based on the Information Gain for each feature, which feature should be used first to split the dataset when creating a decision tree?

(**Please follow the same steps we covered in class to calculate information gain. For example, use one circle to represent the entire dataset before splitting, and then use two circles to represent the two subgroups after the split**).

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	No
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

- ❑ Implementation: [50%] Check details in the attached Firstname_Lastname.py file.

  In this implementation, you need to implement the find_best_cutpoint() function based on the algorithm discussed in class. Specifically, for each continuous attribute:

  1. Sort the feature values in ascending order.
  2. Iterate through the successive midpoints of the sorted values.
     a. Calculate the information gain for current midpoint
  3. Finally, return the cutpoint(midpoint) with the largest information gain.

- ❑ Submission requirements: [5%]

Put the theory part and code into a folder with the name: [Firstname_LastName_HW*].zip

- Following all the submission requirements [5%]
- Theory part:
  - You solution should be clear, concise but also contain enough details.
  - Only PDF format is allowed.
  - Please make sure the scanned PDF file is high-resolution and easily readable if you wrote your solution on paper.
- Programming part:
  - A sanity check to test the correctness of your algorithm.