

## Fall 2023: CS 4435/5435 and DASE 4435 Data Mining

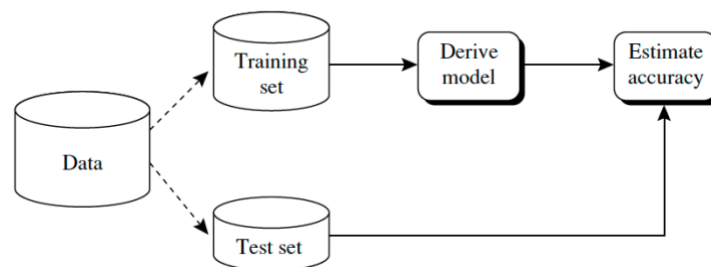
### Homework 3, Due on **October 26, 2023**

#### NOTE:

- (a) This is not a programming assignment. This problem can be solved **MANUALLY**, using paper and pencil or any other way (e.g., using a drawing software).
- (b) Since the problem below is a paper and pencil exercise, make sure that your work is presented neatly, and your answers can be read clearly.

**A decision tree** is a flowchart-like tree structure with non-leaf nodes denoting a test on an attribute, the branch representing the result of the test, and the leaf node denoting the class label. It is an example of a supervised learning classifier where a model is trained using a training dataset, and the resulting model is then used to make prediction on new datasets called the test data.

In decision tree classification, a set of training examples with known classes is used to generate a model, which is used to classify the test dataset. The process is illustrated below. (*Jiawei et al, Fig. 8.17*)



To get started, consider the problem of learning *whether or not* to play golf based on certain attributes. We will limit the number of training examples, since we are doing this manually.

#### **A. Construct a decision tree.**

Construct a decision tree using the attribute selection measurement method used in **ID3** article described in class. Show all your work including calculations. Each process is worth some points. **(70 points)**

**Note:** If multiple attributes tie for the best one, choose the one whose name appears earliest in alphabetical order (e.g., Humidity *before* Outlook *before* Temp. *before* Wind).

## B. Estimate the Accuracy

Using the decision trees produced above, classify the TEST examples. Report the accuracy (i.e., percent correct classification), and error rate on these examples. Briefly discuss your results. Show the Confusion matrix used to determine the accuracy. (30 points)

Here is the **TRAIN** Dataset:

TABLE I. DATA SET S

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sun	Hot	High	Low	No
D2	Sun	Hot	High	High	No
D3	Overcast	Hot	High	Low	Yes
D4	Rain	Sweet	High	Low	Yes
D5	Rain	Cold	Normal	Low	Yes
D6	Rain	Cold	Normal	High	No
D7	Overcast	Cold	Normal	High	Yes
D8	Sun	Sweet	High	Low	No
D9	Sun	Cold	Normal	Low	Yes
D10	Rain	Sweet	Normal	Low	Yes
D11	Sun	Sweet	Normal	High	Yes
D12	Overcast	Sweet	High	High	Yes
D13	Overcast	Hot	Normal	Low	Yes
D14	Rain	Sweet	High	High	No

Here is the **TEST** Dataset:

Day	Outlook	Temperature	Humidity	Wind	Play
D15	Sun	Hot	High	High	No
D16	Overcast	Sweet	High	Low	Yes
D17	Rain	Cold	Normal	Low	Yes
D18	Sun	Sweet	Normal	High	No
D19	Sun	Hot	High	Low	No