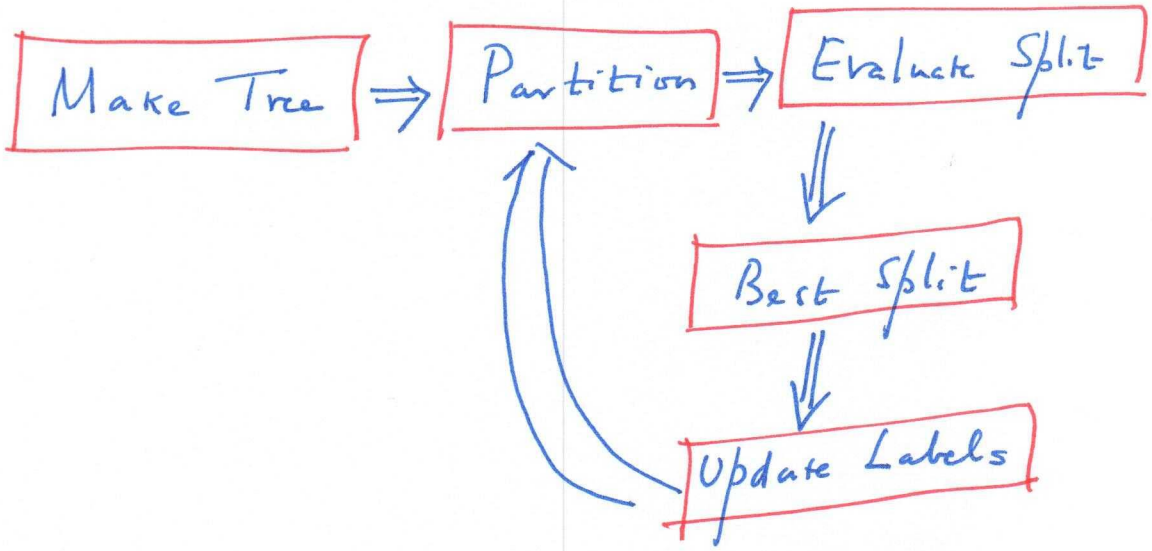
13/10/2021

Assignment 03

"Supervised Learning – classification".

Preamble (as discussed in the class)



Refer SLIQ paper as Attached.

Consider the Toy Data Set (Training Data)

Age	Salary	class
30	65	G
23	15	B
40	75	G
55	40	B
55	100	G
45	60	G

Question 1

Show the detailed Calculation of Computing Gini index and thus the Best split Considering class as label and

(i) Taking Age and Salary as Attributes (Best split among Age & Salary)

(ii) With Root partition keeping Age ≤ 35

Question 2

Write a MATLAB program to generate data from the classification tree of Question 1 (ii) and taking this generated data set as Training Data write a MATLAB program for building classification Tree of SLIQ.

Question 3

(Question 2)

Compute the accuracy of the classifier at every node using class histogram.

Consider the following Training Data for an Insurance firm:

Age	Car Type	Risk
23	family	High
17	sports	High
43	sports	High
68	family	Low
32	truck	Low
20	family	High

Question 4

Construct the classification tree for the above Training Data considering Risk as label (class).

Question 5

Using MATLAB generate Synthetic data by randomly filling Age (17 to 70) and Car Type (among family, sports, truck). Take this as unseen data and Test the classifier of Question 4.

Question b

From the Table 5.1, Generate the Synthetic data (Training Data) taking the attributes: Age, Weight, Gender, Sports, Salt, Sleep, Income, Drink with Hypertension as class Labels. Generate 100 examples (Records). Out of which take 70% as Training Data to build the tree model. Compute the error % and the classifier accuracy in %.

Does Pruning improved the accuracy? If so, which branches have been pruned.

~

Table 5.1 → Next Page

Column name	*Values*	*Explanation*
TypeOfMilk	Integer: 1-5	Type of milk person drinks. Integer values labeled to denote whether milk is whole milk, 2%, skim, powder, or no milk.
DeepFriedLastWeek	Integer: 0-7	Number of times person had deep fried food last week
BeefLastWeek	Integer: 0-7	Number of times person had beef last week
PorkLastWeek	Integer: 0-7	Number of times person had pork last week.
PoultryLastWeek	Integer: 0-7	Number of times person had poultry last week
FishLastWeek	Integer: 0,1,2	Number of times person had fish last week
LambLastWeek	Integer:0-7	Number of times person had lamb last week
OtherMeatLastWeek	Integer:0,1,2, 3,7	Number of times person had other meat last week
CheeseLastWeek	Integer:0-7	Number of times person had cheeses last week
EggsLastWeek	Integer:0-7,9	Number of times person had eggs last week
Meat2MealsLsWk	Integer:0-7,9	Number of times person had meat in two meals last week
Salt	Integer: 1-5,9	Does a person use salt in food? Numbers are labeled to denote a lot, moderate, very little, or none
Butter food	Integer: 1-3	Does a person butter their food? Numbers are labeled to denote frequently, some times, and never.
Sports	Integer: 1- 5,9	Does a person exercise ? Numbers are labeled to denote number of hours a person spends in a week.
Sleep	Integer: in hours 0 -24	The number of hours a person sleeps, on an average.
Smoking	Integer:1-4	Does a person smoke? Numbers are labeled to denote regular, occasional, former, never.
Drink	Integer:1-5	Does a person drink every day ?,Numbers are labeled to denote daily, weekly, occasionally, rarely and never.
Age	Integer:	Age of a person.
YearsEducation	Integer:1-5	How many years education has a person had?
Income	Integer:	What is persons income ?
Gender	Categorical: M,F	What is gender of person? M,F are labeled to denote male and female.
Weight	Integer: in kilograms	What is the weight of the customer?
Height	Integer: in inches	What is the height of the customer?
Hypertension	Categorical: L,H,N	What is the level of hypertension of person? **L,N,H** are labeled to denote low,normal, and High.

Table 5.1: Data set on a hypertension study.