CS 229 ASSIGNMENT 4:

Decision Tree

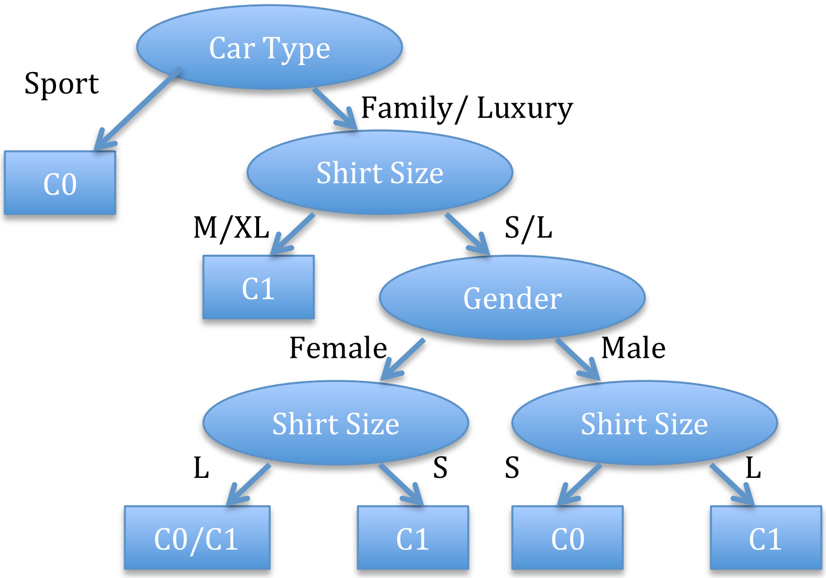
NAME: Xiaopeng Xu KAUST ID: 129052

1. (**30 points**) Consider the training data shown in Table 1:
Construct a decision tree by splitting based on the gain in the **Gini index or Gain Ratio**

Table 1 data set for decision tree classification

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

The final tree I constructed is:



2. (**55 points**)

Table 2 consists of training data from an employee database. The data have been generalized. For a given row entry, *count* represents the number of data examples having the values for *departments, status, age*, and *salary* given in that row. Let the *status* be the class label attribute.

(1) (**5 points**) How to modify C4.5 algorithm to take into consideration the *count* of each generalized data tuple (i.e. of each row entry)?

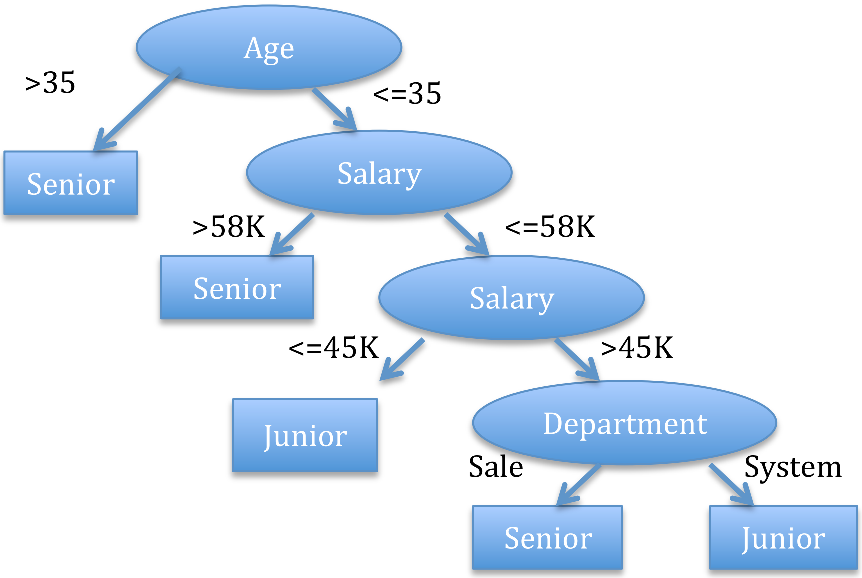
I should add a weight for each sample, when calculating the number of samples. The number of sample is not the number of samples but the summation of sample counts.

(2) (**30 points**) Construct a decision tree from the given data by using the modified C4.5 algorithm

Table 2 data set of an employee database

department	status	Age	salary	count
sales	senior	31…35	46K-50K	30
sales	junior	26...30	26K-30K	40
sales	Junior	31…35	31K-35K	40
systems	junior	21…25	46K-50K	20
systems	senior	31…35	66K-70K	5
systems	junior	26…30	46K-50K	3
systems	senior	41…45	66K-70K	3
marketing	senior	36…40	46K-50K	10
marketing	junior	31…35	41K-45K	4
secretary	senior	46…50	36K-40K	4
secretary	junior	26…30	26K-30K	6

The tree I constructed is:



(3) (**5 points**) use the tree you learned to classify a given example with the values "system", "26…30" and "46-50K" for the attributes *departments, age*, and *salary* . The *status* of this employee is?
Junior

(4) (**15 points**) Use the training data in Table 2 to learn a Naïve Bayes classifier, and classify the same given example with the values "system", "26…30" and "46-50K" for the attributes *departments, age*, and *salary.* The *status* of this employee is?
Junior

3. (**15 points**) Why is *tree pruning* useful in decision tree induction? What are the pros and cons of using a separate set of samples to evaluate pruning?

1) Tree pruning
   a. Helps avoid over-fitting;
   b. Reduces the tree size to generate a more robust and accurate classifier.

2) Using separate sample to evaluate pruning:
   Pros: If choosing suitable dataset to do this, this can be a cross-validation to the tree constructed. So the tree could be more accurate.
   Cons: a. Not fully use the information available to construct the Decision Tree, thus will miss some information.
   b. If the separate sample is not a good representative for the training data, then the pruned tree will be biased.