CS 229 Machine Learning, spring 2019

Homework 5: Decision Tree

Due Saturday March 16, 11:59pm

Submit by the **blackboard system**

1. (**50 points**) Table 1 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.
2. (**30 points**) Construct a decision tree from the given data by using the C4.5 algorithm. (Note: C4.5 uses **Gain Ratio** for evaluating the split quality. Here, you should modify the calculation of gain ratio a little bit to consider the *count* of the generalized data tuple)

Suggestion: *Age* and *Salary* have been discretized into intervals. You can consider them like ordinal attributes. When trying multi-splitting, you can merge values by their closeness. For example, if you have a three-way split of age, you can have [26-30] at one branch, [31 35] at one , and [36 40] [41 45] [46 50] at one. Same for the *Salary.*  You can choose either three-way or four-way split for *Age* and *Salary.*)

Table 1 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 |

1. (**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?
2. (**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?