**Assignment 3**

**Note:**

**For all the questions, if you use CODES or attach the CODE file as your computation and reasoning, then this part will be ZERO.**

Q1. Entropy, information gain, information gain rate, Gini index, classification error rate computation (40 points)

The following table is a mini data set about some loan\_granting records of a bank (Let “loan\_granting” be the class label attribute)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **rID** | **age** | **taxable\_income  (K yuan)** | **credit\_rating** | **marital\_status** | **car\_type** | **loan\_granting** |
| 1 | senior | 15.0 | good | married | luxury | yes |
| 2 | middle\_aged | 25.0 | good | divorced | luxury | yes |
| 3 | middle\_aged | 12.0 | good | married | family | no |
| 4 | youth | 6.5 | fair | single | sports | no |
| 5 | youth | 8.5 | good | single | sports | yes |
| 6 | senior | 10.0 | fair | divorced | family | no |
| 7 | middle\_aged | 14.5 | fair | married | family | yes |
| 8 | senior | 13.5 | good | married | family | yes |
| 9 | youth | 11.0 | fair | married | luxury | yes |
| 10 | middle\_aged | 10.0 | fair | divorced | family | no |
| 11 | senior | 9.5 | good | divorced | family | no |
| 12 | youth | 7.5 | fair | married | sports | no |
| 13 | middle\_aged | 13.5 | good | married | family | yes |
| 14 | middle\_aged | 8.5 | fair | single | sports | no |
| 15 | youth | 10.0 | good | single | sports | yes |
| 16 | senior | 10.5 | good | divorced | family | no |
| 17 | youth | 5.0 | fair | single | sports | no |
| 18 | youth | 5.5 | fair | single | sports | no |
| 19 | senior | 20.0 | fair | married | luxury | yes |
| 20 | middle\_aged | 27.0 | good | married | luxury | yes |

1. Compute the entropy of this collection training examples. (4 points. Computation and reasoning: 3 points; result: 1 points)
2. For “Taxable\_income”, which is a continuous attribute, compute the Gini index for every possible split. (6 points. Computation and reasoning: 5 points; conclusion: 1 points)
3. Show which one is the best split (among “age”, “taxable\_income”, “credit\_rating”, “marital\_status”, “car\_type”) by calculating the Gini index. (6 points. Computation and reasoning: 5 points; conclusion: 1 points)
4. Show which one is the best split (among “age”, “credit\_rating”, “marital\_status”, “car\_type”) by calculating the entropy. (6 points. Computation and reasoning: 5 points; conclusion: 1 points)
5. Show which one is the best split (among “age”, “credit\_rating”, “marital\_status”, “car\_type”) by calculating the information gain. (6 points. Computation and reasoning: 5 points; conclusion: 1 points)
6. Show the information gain rate of attribute “age”. (4 points. Computation and reasoning: 3 points; result: 1 points)
7. Show which one is the best split (among “age”, “credit\_rating”, “marital\_status”, “car\_type”) by calculating the classification error rate? (5 points. Computation and reasoning: 4 points; conclusion: 1 points)
8. Explain why “RID” should not be used as the attribute test condition even though it has the lowest Gini. (3 points)

Q2. Construct decision tree. (42 points)

The following table is some records of jade identification (Let “Quality” be the class label attribute).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RID** | **Color** | **Transparency** | **Knocking sound** | **Color uniformity** | **Thickness** | **Smoothness** | **Quality** |
| 1 | vivid green | high | clear | high | thick | smooth | Good |
| 2 | deep green | high | medium clear | high | thick | smooth | Good |
| 3 | deep green | high | clear | high | thick | smooth | Good |
| 4 | vivid green | high | medium clear | high | thick | smooth | Good |
| 5 | light green | high | clear | high | thick | smooth | Good |
| 6 | vivid green | medium | clear | high | medium thick | rough | Good |
| 7 | deep green | medium | clear | low | medium thick | rough | Good |
| 8 | deep green | medium | clear | high | medium thick | smooth | Good |
| 9 | deep green | medium | medium clear | low | medium thick | smooth | Not good |
| 10 | vivid green | low | dull | high | thin | rough | Not good |
| 11 | light green | low | dull | low | thin | smooth | Not good |
| 12 | light green | high | clear | low | thin | rough | Not good |
| 13 | vivid green | medium | clear | low | thick | smooth | Not good |
| 14 | light green | medium | medium clear | low | thick | smooth | Not good |
| 15 | deep green | medium | clear | high | medium thick | rough | Not good |
| 16 | light green | high | clear | low | thin | smooth | Not good |

1. Compute a decision tree using ID3 algorithm. Use information gain as the criterion for splitting. (25 points. Computation and reasoning for each splitting: 20; result: 5 points.

Hints: you should work out the basis of each splitting)

1. Given a data tuple having the values {“*vivid green*”, “*medium*”, “*medium*”, “*high*”, “*medium thick*”, “*smooth*”}, show what a naïve Bayesian classification of the *Quality* should be? (5 points. Computation and reasoning: 4 points; conclusion: 1 points)
2. Show the shortcomings of the algorithm in a). And show your idea about how to improve. (3 points)
3. Show the difference between ID3, C4.5 and CART algorithm to construct the decision tree. (6 points)
4. What can we do to avoid over-fitting? (3 points)

Q3. Evaluation of Decision tree Prediction. (18 points. Computation: 12 points; diagram: 6 points.)

The data tuples in the following figure are sorted by decreasing probability value, as returned by a classifier. For each tuple, compute the values for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Compute the true positive rate (TPR) and false positive rate (FPR). Plot the ROC curve for the data.

|  |  |  |
| --- | --- | --- |
| **Tuple No.** | **Class** | **Probability** |
| 1 | P | 0.98 |
| 2 | N | 0.83 |
| 3 | P | 0.77 |
| 4 | P | 0.68 |
| 5 | N | 0.63 |
| 6 | P | 0.58 |
| 7 | P | 0.54 |
| 8 | N | 0.53 |
| 9 | N | 0.46 |
| 10 | P | 0.42 |