**Exercises: Decision-trees**

**Exercise 1 (Feature selection):** Consider the following training examples for a binary classification problem, where we want to predict the value of Class attribute (Co or C1?)

**Table 4.1.** Data set for Exercise 2.

| 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 |

(a) Compute the Gini index for the overall collection of training examples.
(b) Compute the Gini index for the Gender attribute.
(d) Compute the Gini index for the Car Type attribute using multi-way split.
(e) Compute the Gini index for the Shirt Size attribute using multi-way split.
(f) Which attribute is better, Gender, Car Type, or Shirt Size? Why?
(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

**Exercise 2 (Building decision tree):** Given the following training data: *Buy Computer* data

| ID | age | income | student | credit_rating | buy |
|---|---|---|---|---|---|
| 1 | <=30 | high | no | fair | no |
| 2 | <=30 | high | no | excellent | no |
| 3 | 31 . . . 40 | high | no | fair | yes |
| 4 | >40 | medium | no | fair | yes |
| 5 | >40 | low | yes | fair | yes |
| 6 | >40 | low | yes | excellent | no |
| 7 | 31 . . . 40 | low | yes | excellent | yes |
| 8 | <=30 | medium | no | fair | no |
| 9 | <=30 | low | yes | fair | yes |
| 10 | >40 | medium | yes | fair | yes |
| 11 | <=30 | medium | yes | excellent | yes |
| 12 | 31 . . . 40 | medium | no | excellent | yes |
| 13 | 31 . . . 40 | high | yes | fair | yes |
| 14 | >40 | medium | no | excellent | no |

a) Step by step, build a decision tree using Information Gain based on Entropy
b) Use the constructed tree to predict the class of the following new example: age<=30, income=medium, student=yes, credit-rating=fair.

**Exercise 3 (Building decision tree using software):** Use Rapid-Miner (or Python) for building decision tree with the same parameters as Exercise 2. Compare the two trees.

- ⇨ Show your parameter settings
- ⇨ Submit HW#5_exercise3_student_ID.xml or .py