# Decision Tree Simulation

# Sample Data: D

| ID | age | income | student | creditRating | buyComputer |
|---|---|---|---|---|---|
| 1 | youth | 100K | yes | fair | no |
| 2 | youth | 120K | no | excellent | no |
| 3 | middleaged | 150K | no | fair | yes |
| 4 | middleaged | 75K | no | fair | yes |
| 5 | senior | 50K | yes | fair | yes |
| 6 | senior | 48K | no | excellent | no |
| 7 | midleaged | 20K | yes | excellent | yes |
| 8 | youth | 75K | no | fair | no |
| 9 | youth | 30K | yes | fair | yes |
| 10 | senior | 88K | yes | fair | yes |
| 11 | youth | 25K | yes | excellent | yes |
| 12 | middleaged | 58K | no | excellent | yes |
| 13 | middleaged | 90K | yes | fair | yes |
| 14 | senior | 180K | yes | excellent | no |

# Sample Data: D with Non nominal Data

| ID | age | income | student | creditRating | buyComputer |
|---|---|---|---|---|---|
| 1 | youth | 100K | yes | fair | no |
| 2 | youth | 120K | no | excellent | no |
| 3 | middleaged | 150K | no | fair | yes |
| 4 | middleaged | 75K | no | fair | yes |
| 5 | senior | 50K | yes | fair | yes |
| 6 | senior | 48K | no | excellent | no |
| 7 | midleaged | 20K | yes | excellent | yes |
| 8 | youth | 75K | no | fair | no |
| 9 | youth | 30K | yes | fair | yes |
| 10 | senior | 88K | yes | fair | yes |
| 11 | youth | 25K | yes | excellent | yes |
| 12 | middleaged | 58K | no | excellent | yes |
| 13 | middleaged | 90K | yes | fair | yes |
| 14 | senior | 180K | yes | excellent | no |

Minimum: 20K
Maximum: 180K

Assumption

$$A < 50K$$
$$50K \leq B < 90K$$
$$C \geq 90K$$

# Sample Data: D after preprocessing

| ID | age | income | student | creditRating | buyComputer |
|---|---|---|---|---|---|
| 1 | youth | C | yes | fair | no |
| 2 | youth | C | no | excellent | no |
| 3 | middleaged | C | no | fair | yes |
| 4 | middleaged | B | no | fair | yes |
| 5 | senior | B | yes | fair | yes |
| 6 | senior | A | no | excellent | no |
| 7 | midleaged | A | yes | excellent | yes |
| 8 | youth | B | no | fair | no |
| 9 | youth | A | yes | fair | yes |
| 10 | senior | B | yes | fair | yes |
| 11 | youth | A | yes | excellent | yes |
| 12 | middleaged | B | no | excellent | yes |
| 13 | middleaged | C | yes | fair | yes |
| 14 | senior | C | yes | excellent | no |

# Necessary Equations

$$Info(D) = -\sum_{i=1}^{m} p_i \lg(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

# Sample Data: D

| ID | age | income | student | creditRating | buyComputer |
|---|---|---|---|---|---|
| 1 | youth | C | yes | fair | no |
| 2 | youth | C | no | excellent | no |
| 3 | middleaged | C | no | fair | yes |
| 4 | middleaged | B | no | fair | yes |
| 5 | senior | B | yes | fair | yes |
| 6 | senior | A | no | excellent | no |
| 7 | midleaged | A | yes | excellent | yes |
| 8 | youth | B | no | fair | no |
| 9 | youth | A | yes | fair | yes |
| 10 | senior | B | yes | fair | yes |
| 11 | youth | A | yes | excellent | yes |
| 12 | middleaged | B | no | excellent | yes |
| 13 | middleaged | C | yes | fair | yes |
| 14 | senior | C | yes | excellent | no |

$$Info(D) = -\frac{5}{14}\lg\left(\frac{5}{14}\right) - \frac{9}{14}\lg\left(\frac{9}{14}\right) = 0.9403$$

# Sample Data: Checking for “age”

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 1 | C | yes | fair | no |
| 2 | C | no | excellent | no |
| 8 | B | no | fair | no |
| 9 | A | yes | fair | yes |
| 11 | A | yes | excellent | yes |

$$Info(D_{age=youth}) = -\frac{2}{5}\lg\left(\frac{2}{5}\right) - \frac{3}{5}\lg\left(\frac{3}{5}\right) = 0.971$$

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 3 | C | no | fair | yes |
| 4 | B | no | fair | yes |
| 7 | A | yes | excellent | yes |
| 12 | B | no | excellent | yes |
| 13 | C | yes | fair | yes |

$$Info(D_{age=middleaged}) = 0$$

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 5 | B | yes | fair | yes |
| 6 | A | no | excellent | no |
| 10 | B | yes | fair | yes |
| 14 | C | yes | excellent | no |

$$Info(D_{age=senior}) = -\frac{2}{4}\lg\left(\frac{2}{4}\right) - \frac{2}{4}\lg\left(\frac{2}{4}\right) = 1$$

$$Info_{age}(D) = \frac{5}{14} \times 0.971 + \frac{5}{14} \times 0 + \frac{4}{14} \times 1 = 0.6325$$

$$Gain(age) = 0.9403 - 0.6325 = 0.3078$$

# Sample Data: Checking for “income”

| ID | age | student | creditRating | buyComputer |
|---|---|---|---|---|
| 6 | senior | no | excellent | no |
| 7 | midleaged | yes | excellent | yes |
| 9 | youth | yes | fair | yes |
| 11 | youth | yes | excellent | yes |

$$Info(D_{income=A}) = -\frac{1}{4}\lg\left(\frac{1}{4}\right) - \frac{3}{4}\lg\left(\frac{3}{4}\right) = 0.8113$$

| ID | age | student | creditRating | buyComputer |
|---|---|---|---|---|
| 1 | youth | yes | fair | no |
| 2 | youth | no | excellent | no |
| 3 | middleaged | no | fair | yes |
| 13 | middleaged | yes | fair | yes |
| 14 | senior | yes | excellent | no |

| ID | age | student | creditRating | buyComputer |
|---|---|---|---|---|
| 4 | middleaged | no | fair | yes |
| 5 | senior | yes | fair | yes |
| 8 | youth | no | fair | no |
| 10 | senior | yes | fair | yes |
| 12 | middleaged | no | excellent | yes |

$$Info(D_{income=B}) = -\frac{1}{5}\lg\left(\frac{1}{5}\right) - \frac{4}{5}\lg\left(\frac{4}{5}\right) = 0.7219$$

$$Info(D_{income=C}) = -\frac{2}{5}\lg\left(\frac{2}{5}\right) - \frac{3}{5}\lg\left(\frac{3}{5}\right) = 0.971$$

$$Info_{income}(D) = \frac{4}{14} \times 0.8113 + \frac{5}{14} \times 0.7219 + \frac{5}{14} \times 0.971 = 0.8364$$

$$Gain(income) = 0.9403 - 0.8364 = 0.1039$$

# Sample Data: Checking for “student”

| ID | age | income | creditRating | buyComputer |
|---|---|---|---|---|
| 1 | youth | C | fair | no |
| 5 | senior | B | fair | yes |
| 7 | midleaged | A | excellent | yes |
| 9 | youth | A | fair | yes |
| 10 | senior | B | fair | yes |
| 11 | youth | A | excellent | yes |
| 13 | middleaged | C | fair | yes |
| 14 | senior | C | excellent | no |

| ID | age | income | creditRating | buyComputer |
|---|---|---|---|---|
| 2 | youth | C | excellent | no |
| 3 | middleaged | C | fair | yes |
| 4 | middleaged | B | fair | yes |
| 6 | senior | A | excellent | no |
| 8 | youth | B | fair | no |
| 12 | middleaged | B | excellent | yes |

$$Info(D_{student=no}) = -\frac{3}{6}\lg\left(\frac{3}{6}\right) - \frac{3}{6}\lg\left(\frac{3}{6}\right) = 1$$

$$Info(D_{student=yes}) = -\frac{2}{8}\lg\left(\frac{2}{8}\right) - \frac{6}{8}\lg\left(\frac{6}{8}\right) = 0.8113$$

$$Info_{student}(D) = \frac{8}{14} \times 0.8113 + \frac{6}{14} \times 1 = 0.8922$$

$$Gain(student) = 0.9403 - 0.8922 = 0.0481$$

# Sample Data: Checking for “creditRating”

| ID | age | income | student | buyComputer |
|---|---|---|---|---|
| 1 | youth | C | yes | no |
| 3 | middleaged | C | no | yes |
| 4 | middleaged | B | no | yes |
| 5 | senior | B | yes | yes |
| 8 | youth | B | no | no |
| 9 | youth | A | yes | yes |
| 10 | senior | B | yes | yes |
| 13 | middleaged | C | yes | yes |

$$Info(D_{creditRating=fair}) = -\frac{2}{8}\lg\left(\frac{2}{8}\right) - \frac{6}{8}\lg\left(\frac{6}{8}\right) = 0.8113$$

| ID | age | income | student | buyComputer |
|---|---|---|---|---|
| 2 | youth | C | no | no |
| 6 | senior | A | no | no |
| 7 | midleaged | A | yes | yes |
| 11 | youth | A | yes | yes |
| 12 | middleaged | B | no | yes |
| 14 | senior | C | yes | no |

$$Info(D_{creditRating=excellent}) = -\frac{3}{6}\lg\left(\frac{3}{6}\right) - \frac{3}{6}\lg\left(\frac{3}{6}\right) = 1$$

$$Info_{creditRating}(D) = \frac{8}{14} \times 0.8113 + \frac{6}{14} \times 1 = 0.8922$$

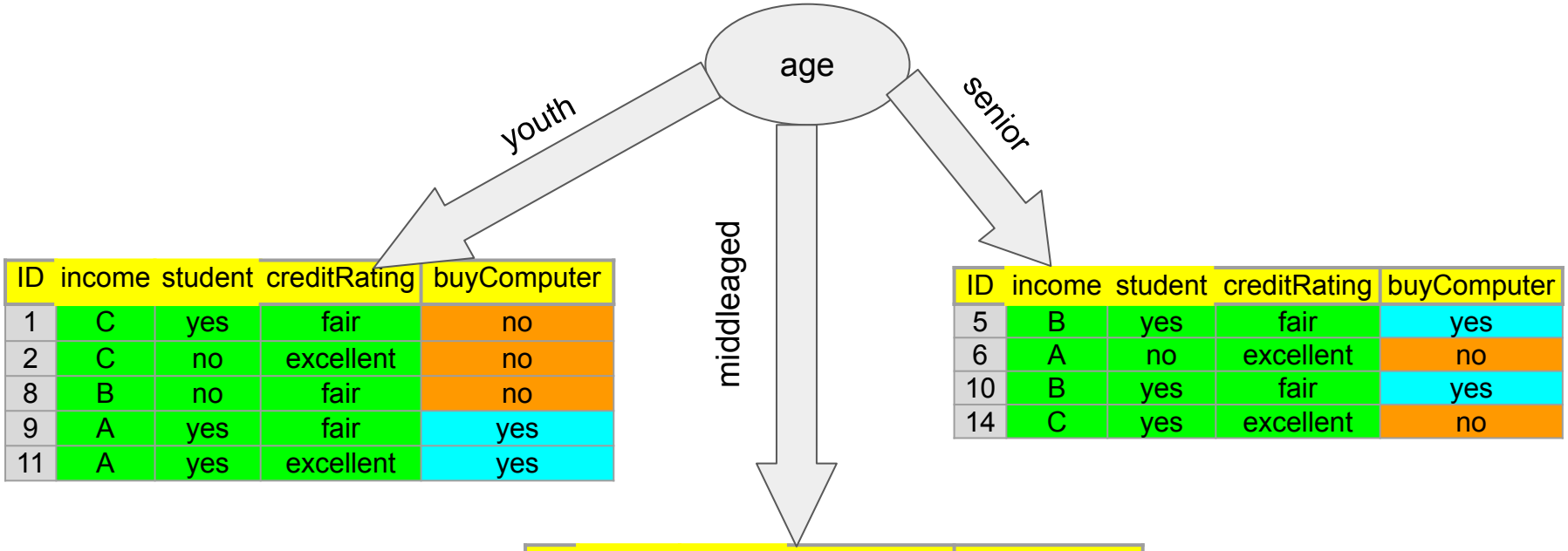
$$Gain(student) = 0.9403 - 0.8922 = 0.0481$$

# Attribute Selection for Splitting

| Available Attribute, X | Info_X(D) | Gain(X) |
|---|---|---|
| age | 0.6325 | 0.3078 |
| income | 0.8364 | 0.1039 |
| student | 0.8922 | 0.0481 |
| creditRating | 0.8922 | 0.0481 |

maximum gain

# Decision Tree After First Step

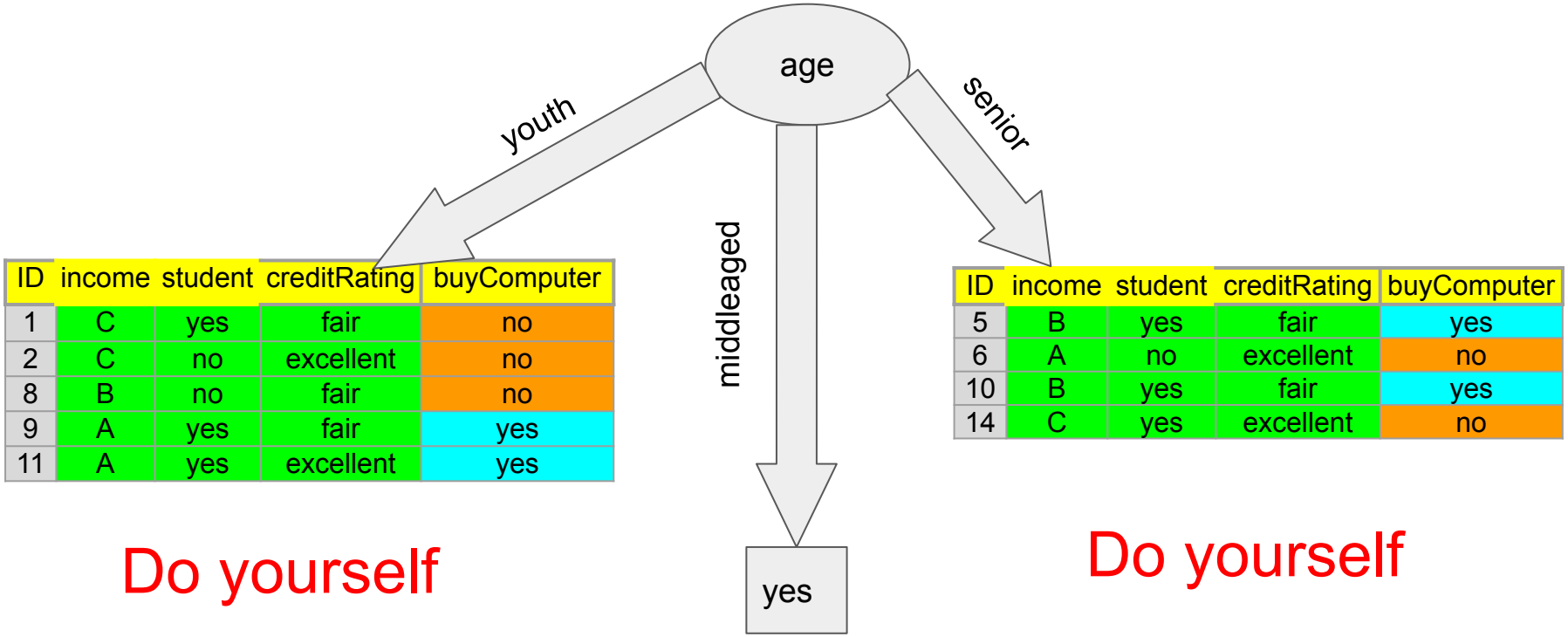



| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 1 | C | yes | fair | no |
| 2 | C | no | excellent | no |
| 8 | B | no | fair | no |
| 9 | A | yes | fair | yes |
| 11 | A | yes | excellent | yes |

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 5 | B | yes | fair | yes |
| 6 | A | no | excellent | no |
| 10 | B | yes | fair | yes |
| 14 | C | yes | excellent | no |

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 3 | C | no | fair | yes |
| 4 | B | no | fair | yes |
| 7 | A | yes | excellent | yes |
| 12 | B | no | excellent | yes |
| 13 | C | yes | fair | yes |

# Decision Tree After First Step




| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 1 | C | yes | fair | no |
| 2 | C | no | excellent | no |
| 8 | B | no | fair | no |
| 9 | A | yes | fair | yes |
| 11 | A | yes | excellent | yes |

Do yourself

| ID | income | student | creditRating | buyComputer |
|---|---|---|---|---|
| 5 | B | yes | fair | yes |
| 6 | A | no | excellent | no |
| 10 | B | yes | fair | yes |
| 14 | C | yes | excellent | no |

Do yourself

# Another example for practice

| ID | W | X | Y | Z | Class |
|---|---|---|---|---|---|
| 1 | yes | 2 | B | 1 | yes |
| 2 | yes | 2 | C | 2 | yes |
| 3 | yes | 2 | B | 3 | yes |
| 4 | no | 2 | B | 1 | yes |
| 5 | no | 3 | B | 2 | no |
| 6 | yes | 3 | B | 1 | yes |
| 7 | yes | 2 | C | 1 | no |
| 8 | no | 3 | A | 2 | yes |
| 9 | yes | 2 | B | 2 | yes |
| 10 | yes | 1 | C | 2 | yes |
| 11 | no | 1 | B | 2 | no |
| 12 | yes | 1 | B | 3 | yes |

*Thank You*