# Computer Science 477

# Using Frequency Tables GINI Index and $\chi^2$ for Attribute Selection

Lecture 7

# Optimizing Entropy Calculations

- Calculating entropy laborious
- At each node a table of values such as needs to be calculated for every possible value of every categorical attribute.
- More efficient method: single table to be constructed for each categorical attribute at each node.

| Value of attribute | | | | Class |
|---|---|---|---|---|
| age | specRx | astig | tears | |
| 1 | 1 | 1 | 1 | 3 |
| 1 | 1 | 1 | 2 | 2 |
| 1 | 1 | 2 | 1 | 3 |
| 1 | 1 | 2 | 2 | 1 |
| 1 | 2 | 1 | 1 | 3 |
| 1 | 2 | 1 | 2 | 2 |
| 1 | 2 | 2 | 1 | 3 |
| 1 | 2 | 2 | 2 | 1 |

Lens24 training data for age = 1

| | age = 1 | age = 2 | age = 3 |
|---|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

Frequency table for age.

Number of occurrences for each class and each value of the attribute age.

# Frequency Table

- Entire lens24 data set.

| Value of attribute | | | | Class |
|---|---|---|---|---|
| age | specRx | astig | tears | |
| 1 | 1 | 1 | 1 | 3 |
| 1 | 1 | 1 | 2 | 2 |
| 1 | 1 | 2 | 1 | 3 |
| 1 | 1 | 2 | 2 | 1 |
| 1 | 2 | 1 | 1 | 3 |
| 1 | 2 | 1 | 2 | 2 |
| 1 | 2 | 2 | 1 | 3 |
| 1 | 2 | 2 | 2 | 1 |
| 2 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 2 | 2 |
| 2 | 1 | 2 | 1 | 3 |
| 2 | 1 | 2 | 2 | 1 |
| 2 | 2 | 1 | 1 | 3 |
| 2 | 2 | 1 | 2 | 2 |
| 2 | 2 | 2 | 1 | 3 |
| 2 | 2 | 2 | 2 | 3 |
| 3 | 1 | 1 | 1 | 3 |
| 3 | 1 | 1 | 2 | 3 |
| 3 | 1 | 2 | 1 | 3 |
| 3 | 1 | 2 | 2 | 1 |
| 3 | 2 | 1 | 1 | 3 |
| 3 | 2 | 1 | 2 | 2 |
| 3 | 2 | 2 | 1 | 3 |
| 3 | 2 | 2 | 2 | 3 |

f

| | age = 1 | age = 2 | age = 3 |
|---|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

Frequency table for age.

# Calculation of Entropy

- Denote the total number of instances by $N$, so $N = 24$.
- $E_{new}$, average entropy of the training sets resulting from splitting on a specified attribute, calculated by forming a new sum.
- (1) For every non-zero value $V$ in the main body of the table (part above the 'column sum' row),subtract $V \times \log_2 V$.
- (2) For every non-zero value $S$ in the column sum row, add $S \times \log_2 S$.
- Divide total by $N$

| | age = 1 | age = 2 | age = 3 |
|---|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

$$-2 \cdot \log_2 2 - 1 \cdot \log_2 1 - 1 \cdot \log_2 1$$
$$-2 \cdot \log_2 2 - 2 \cdot \log_2 2 - 1 \cdot \log_2 1$$
$$-4 \cdot \log_2 4 - 5 \cdot \log_2 5 - 6 \cdot \log_2 6$$
$$+ 8 \cdot \log_2 8 + 8 \cdot \log_2 8 + 8 \cdot \log_2 8$$

# Calculating Entropy

- Using table of logs:
- $-2 \cdot \log_2 2 - 1 \cdot \log_2 1 - 1 \cdot \log_2 1 - 2 \cdot \log_2 2 - 2 \cdot \log_2 2 - 1 \cdot \log_2 1 - 4 \cdot \log_2 4 - 5 \cdot \log_2 5 - 6 \cdot \log_2 6 + 8 \cdot \log_2 8 + 8 \cdot \log_2 8 + 8 \cdot \log_2 8$
- Collecting terms, rearranging and dividing by 24:
- $(-3 \times 2 \cdot \log_2 2 - 3 \cdot \log_2 1 - 4 \cdot \log_2 4 - 5 \cdot \log_2 5 - 6 \cdot \log_2 6 + 3 \times 8 \cdot \log_2 8)/24$
- Giving: 1.2867 bits
  - Agrees with previous calculation

| $x$ | $\log_2 x$ |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 1.5850 |
| 4 | 2 |
| 5 | 2.3219 |
| 6 | 2.5850 |
| 7 | 2.8074 |
| 8 | 3 |
| 9 | 3.1699 |
| 10 | 3.3219 |
| 11 | 3.4594 |
| 12 | 3.5850 |

Useful table of logs.

## Observation about Zero

- New method of computing entropy excludes empty classes from the summation.
- They correspond to zero entries in the body of the frequency table
- If a complete column of the frequency table is zero it means that the categorical attribute never takes one of its possible values at the node under consideration.

# Gini Index of Diversity

- Another measure of node coherence
- Given $K$ classes, with the probability of the *i*th class being $p_i$, the Gini Index is defined as $1 - \sum_{i=1}^{n} p_i^2$
- Its smallest value is zero
  - When all the classifications are the same.
- Largest value $1 - \frac{1}{K}$
  - Classes are evenly distributed between the instances
  - The frequency of each class is $1/K$.

## Calculating the GINI index

- For each non-empty column, form the sum of the squares of the values in the body of the table and divide by the column sum.
- Add the values obtained for all the columns and divide by *N*
  - (the number of instances).
- Subtract the total from 1.

# GINI Example Calculation

| | age = 1 | age = 2 | age = 3 |
|---|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

**age = 1:** $(2^2 + 2^2 + 4^2)/8 = 3$
**age = 2:** $(1^2 + 2^2 + 5^2)/8 = 3.75$
**age = 3:** $(1^2 + 1^2 + 6^2)/8 = 4.75$

- Giving $GINI_{new} = 1 - \frac{3+3.27+4.75}{24} = 0.5208$
- Reduction by splitting on **age** is $0.5382 - 0.5208 = 0.0174$

# Various GINI Calculations

- specRx: $G_{new}$ = 0.5278, so the reduction is 0.5382 − 0.5278 = 0.0104
- astig: $G_{new}$ = 0.4653, so the reduction is 0.5382 − 0.4653 = 0.0729
- tears: $G_{new}$ = 0.3264, so the reduction is 0.5382 − 0.3264 = 0.2118
- The attribute selected - the one which gives the largest *reduction* in the value of the Gini Index, i.e. tears.
- This is the same attribute that was selected using entropy.

# Implicit Bias

- Entropy has bias towards selecting attributes with a large number of values
- Example: a dataset about people that includes an attribute 'place of birth'
  - Classifies them (as responding to some medical treatment) 'well' 'badly' or 'not at all'.
- Do not expect place of birth to have significant effect on the classification.
- Information gain selection method will almost certainly choose it as the first attribute to split.
  - Generating one branch for each possible place of birth
  - Large branching factor at top of tree.
- The decision tree will be very large, with many branches (rules) with very low value for classification.

# Gain Ratio for Attribute Selection

- The the average entropy of the training sets resulting from splitting on attribute *age*, $1.2867$
- Entropy of the original training set $E_{start} = 1.3261.$
- Information Gain $= E_{start} - E_{new} = 1.3261 - 1.2867 = 0.0394$
- Gain Ratio = Information Gain/Split Information
  - Split Information is a value based on the column sums
- Each non-zero column sum *s* contributes $-(s/N)\log_2(s/N)$ to the Split Information.
- Value of Split Information is
  $-(8/24)\log_2(8/24) - (8/24)\log_2(8/24) - (8/24)\log_2(8/24) = 1.5850$
- Gain Ratio = 0.0394/1.5850 = 0.0249

| | age = 1 | age = 2 | age = 3 |
|---|---|---|---|
| Class 1 | 2 | 1 | 1 |
| Class 2 | 2 | 2 | 1 |
| Class 3 | 4 | 5 | 6 |
| Column sum | 8 | 8 | 8 |

## Properties of Split Information

- Split Information - denominator in the Gain Ratio formula.
  - Higher the value of Split Information, the lower the Gain Ratio.
- Split Information depends on
  - The number of values a categorical attribute has
  - How uniformly those values are distributed.

# Split Information Examples

- 32 instances
- Consider splitting on an attribute ***a***
  - Values 1, 2, 3 and 4.
- 'Frequency' row in the tables below is the same as the column sum row tables
- Possibility 1 - Single Attribute Value

| | $a = 1$ | $a = 2$ | $a = 3$ | $a = 4$ |
|---|---|---|---|---|
| Frequency | 32 | 0 | 0 | 0 |

- Split Information = $-(32/32) \times \log_2(32/32) = -\log_2 1 = 0$

# Split Information Examples

| | $a=1$ | $a=2$ | $a=3$ | $a=4$ |
|---|---|---|---|---|
| Frequency | 16 | 16 | 0 | 0 |

- Split Information = $-(16/32) \times \log 2(16/32) - (16/32) \times \log_2(16/32) = -\log_2(1/2) = 1$

| | $a=1$ | $a=2$ | $a=3$ | $a=4$ |
|---|---|---|---|---|
| Frequency | 16 | 8 | 8 | 0 |

- Split Information = $-(16/32) \times \log 2(16/32) - 2 \times (8/32) \times \log_2(8/32) = -(1/2)\log_2(1/2) - (1/2)\log 2(1/4) = 0.5 + 1 = 1.5$

# Split Information Examples

| | $a = 1$ | $a = 2$ | $a = 3$ | $a = 4$ |
|---|---|---|---|---|
| Frequency | 16 | 8 | 4 | 4 |

- Split Information = $-(16/32) \times \log2(16/32) - (8/32) \times \log2(8/32) - 2 \times (4/32) \times \log2(4/32) = 0.5 + 0.5 + 0.75 = 1.75$

| | $a = 1$ | $a = 2$ | $a = 3$ | $a = 4$ |
|---|---|---|---|---|
| Frequency | 8 | 8 | 8 | 8 |

- Split Information = $-4 \times (8/32) \times \log_2(8/32) = -\log_2(1/4) = \log_2 4 = 2$
  - With $M$ attribute values, each equally frequent, the Split Information is $\log_2$ (irrespective of the frequency value).

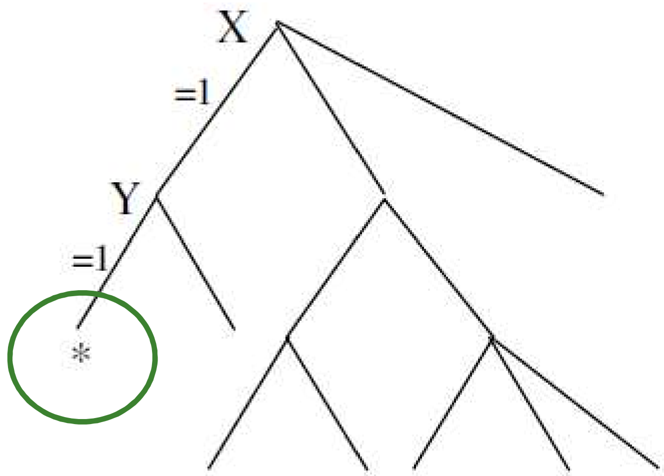
# Gain Ratio and Branching

- Number of Rules Generated by Different Attribute Selection Criteria

| Dataset | Excluding Entropy and Gain Ratio | | Entropy | Gain Ratio |
|---|---|---|---|---|
| | most | least | | |
| contact_lenses | 42 | 26 | <u>16</u> | 17 |
| lens24 | 21 | <u>9</u> | <u>9</u> | <u>9</u> |
| chess | 155 | 52 | <u>20</u> | <u>20</u> |
| vote | 116 | 40 | 34 | <u>33</u> |
| monk1 | 89 | 53 | <u>52</u> | <u>52</u> |
| monk2 | 142 | 109 | <u>95</u> | 96 |
| monk3 | 77 | 43 | 28 | <u>25</u> |

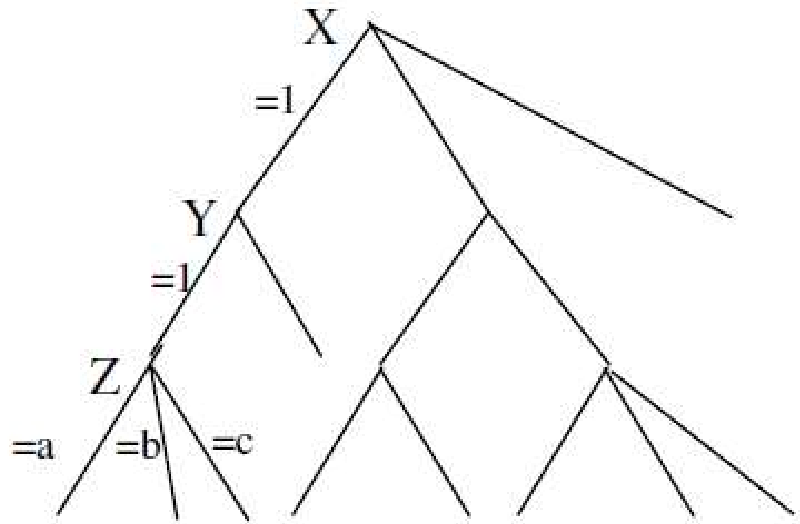
- Gain ratio branches fewer
  - With exceptions
- In practice Information Gain more common than Gain Ratio
  - But C4.5 popular

# Missing Branches




- Splitting next on $Z$ may result in an attribute value unrepresented
- If attribute $Z$ has four possible values, but the branch at $*$ offers three possibilities

# Missing Branches




- If Z has four values, a, b, c, d new instance with $X = 1, Y = 1, Z = d$ will be unclassified
- It may be considered preferable to leave an unseen instance unclassified rather than to classify it wrongly.
- Easy to provide a facility for any unclassified instances to be given a default classification
  - The largest class.
  - Largest class such that $X = 1, Y = 1$ and $Z = d$