1. For the following dataset:

| *ID* | *Outl* | *Temp* | *Humi* | *Wind* | PLAY |
|---|---|---|---|---|---|
| TRAINING INSTANCES | | | | | |
| A | s | h | h | F | N |
| B | s | h | h | T | N |
| C | o | h | h | F | Y |
| D | r | m | h | F | Y |
| E | r | c | n | F | Y |
| F | r | c | n | T | N |
| TEST INSTANCES | | | | | |
| G | o | c | n | T | ? |
| H | s | m | h | F | ? |

Classify the test instances using the **ID3 Decision Tree** method:

a) Using the **Information Gain** as a splitting criterion

b) Using the **Gain Ratio** as a splitting criterion

(a) IG: At each level of DT, choose attribute with largest IG

$$IG(A|R) = H(R) - \sum_{i \in A} P(A=i)\,H(A=i)$$

weight: fraction of instance at child node i

$H(R)$: entropy of parent node

$\sum_{i \in A} P(A=i)\,H(A=i)$: weighted average entropy across child nodes (Mean Information: MI)

$H$: entropy (impurity) of a node

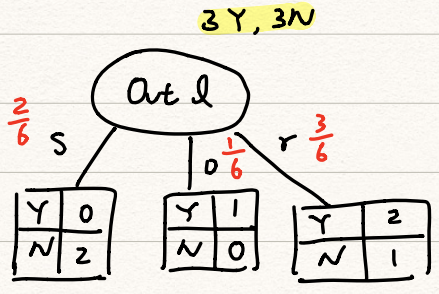
$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Root node: 3Y, 3N

$$H(R) = -\{\tfrac{1}{2}\log_2\tfrac{1}{2} + \tfrac{1}{2}\log_2\tfrac{1}{2}\} = 1$$

($\tfrac{1}{2}\log_2\tfrac{1}{2}$: Y; $\tfrac{1}{2}\log_2\tfrac{1}{2}$: N)

Example: For Outl:




$$H(Outl=s) = -\{0\log_2 0 + 1\log_2 1\} = 0$$

$$H(Outl=o) = -\{1\log_2 1 + 0\log_2 0\} = 0$$

$$H(Outl=r) = -\{\tfrac{2}{3}\log_2\tfrac{2}{3} + \tfrac{1}{3}\log_2\tfrac{1}{3}\} = 0.9183$$

$$MI(Outl) = \tfrac{2}{6}\times 0 + \tfrac{1}{6}\times 0 + \tfrac{3}{6}\times 0.9183 = 0.4592$$

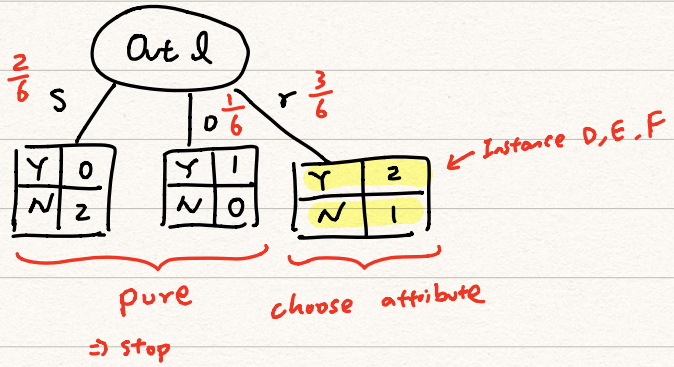
$$IG(Outl) = H(R) - MI(Outl) = 1 - 0.4592 = 0.5408$$

will get useless classifier

| | R | Outl | | | Temp | | | H | | Wind | | ID | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | s | o | r | h | m | c | h | n | T | F | A | B | C | D | E | F |
| Y | 3 | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 0 | 3 | 0 | 0 | 1 | 1 | 1 | 0 |
| N | 3 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| Total | 6 | 2 | 1 | 3 | 3 | 1 | 2 | 4 | 2 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| P(Y) | 1/2 | 0 | 1 | 2/3 | 1/3 | 1 | 1/2 | 1/2 | 1/2 | 0 | 3/4 | 0 | 0 | 1 | 1 | 1 | 0 |
| P(N) | 1/2 | 1 | 0 | 1/3 | 2/3 | 0 | 1/2 | 1/2 | 1/2 | 1 | 1/4 | 1 | 1 | 0 | 0 | 0 | 1 |
| H | 1 | 0 | 0 | 0.9183 | 0.9183 | 0 | 1 | 1 | 1 | 0 | 0.8112 | 0 | 0 | 0 | 0 | 0 | 0 |
| MI | | 0.4592 | | | 0.7924 | | | 1 | | 0.5408 | | 0 | | | | | |
| IG | | 0.5408 | | | 0.2076 | | | 0 | | 0.4592 | | 1 | | | | | |
| SI | | 1.459 | | | 1.459 | | | 0.9183 | | 0.9183 | | 2.585 | | | | | |
| GR | | 0.3707 | | | 0.1423 | | | 0 | | 0.5001 | | 0.3868 | | | | | |

largest

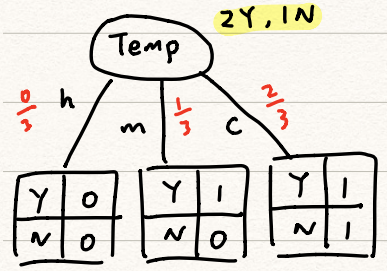
H : 3Y, 3N




⇒ stop

parent : H(Outl = r) = 0.9183

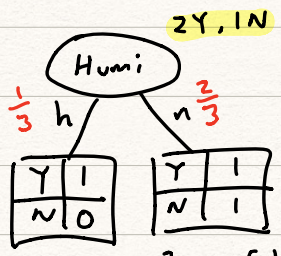
⇒ Find IG for Temp, Humi, Wind




$MI(T) = \frac{2}{3} \times -\{\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\}$

$= \frac{2}{3}$

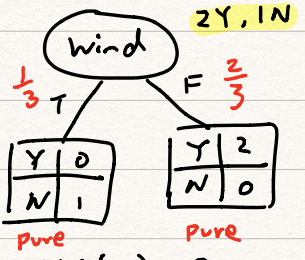
$IG(T) = 0.9183 - \frac{2}{3} = 0.2516$




$MI(w) = \frac{2}{3} \times -\{\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\}$

$= \frac{2}{3}$

$IG(w) = 0.2516$

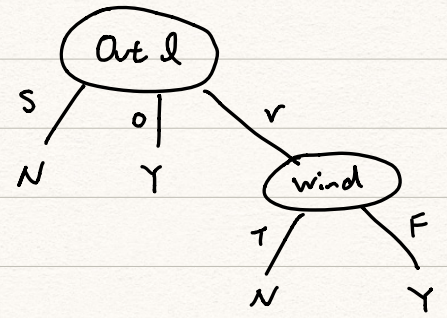



$MI(w) = 0$

(largest)

$IG(w) = 0.9183 - 0 = 0.9183$

DT :




Classify G: Outl = O => Y

Classify H: Outl = S => N

1. For the following dataset:

| *ID* | *Outl* | *Temp* | *Humi* | *Wind* | PLAY |
|---|---|---|---|---|---|
| TRAINING INSTANCES | | | | | |
| A | s | h | h | F | N |
| B | s | h | h | T | N |
| C | o | h | h | F | Y |
| D | r | m | h | F | Y |
| E | r | c | n | F | Y |
| F | r | c | n | T | N |
| TEST INSTANCES | | | | | |
| G | o | c | n | T | ? |
| H | s | m | h | F | ? |

Classify the test instances using the **ID3 Decision Tree** method:

a) Using the **Information Gain** as a splitting criterion

b) Using the **Gain Ratio** as a splitting criterion

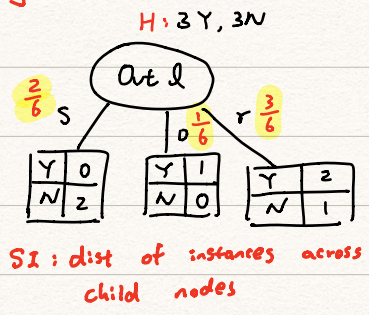
(b) Choose attribute with largest GR.

$$GR(A) = \frac{IG(A)}{SI(A)}$$ ← split info (entropy)

$$SI(Outl) = -\left\{\frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6} + \frac{3}{6}\log_2\frac{3}{6}\right\}$$

$$= 1.459$$

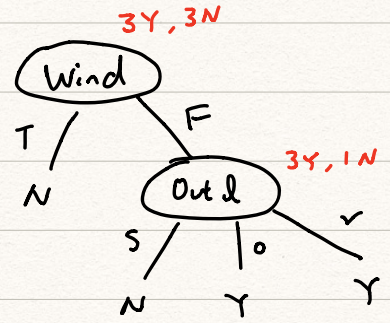
$$\Rightarrow GR(Outl) = \frac{0.5408}{1.459} = 0.3707$$




SI : dist of instances across child nodes

| | R | Outl | | | Temp | | | H | | Wind | | ID | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | s | o | r | h | m | c | h | n | T | F | A | B | C | D | E | F |
| Y | 3 | 0 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 0 | 3 | 0 | 0 | 1 | 1 | 1 | 0 |
| N | 3 | 2 | 0 | 1 | 2 | 0 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| Total | 6 | 2 | 1 | 3 | 3 | 1 | 2 | 4 | 2 | 2 | 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| P(Y) | 1/2 | 0 | 1 | 2/3 | 1/3 | 1 | 1/2 | 1/2 | 1/2 | 0 | 3/4 | 0 | 0 | 1 | 1 | 1 | 0 |
| P(N) | 1/2 | 1 | 0 | 1/3 | 2/3 | 0 | 1/2 | 1/2 | 1/2 | 1 | 1/4 | 1 | 1 | 0 | 0 | 0 | 1 |
| H | 1 | 0 | 0 | 0.9183 | 0.9183 | 0 | 1 | 1 | 1 | 0 | 0.8112 | 0 | 0 | 0 | 0 | 0 | 0 |
| MI | | 0.4592 | | | 0.7924 | | | 1 | | 0.5408 | | 0 | | | | | |
| IG | | 0.5408 | | | 0.2076 | | | 0 | | 0.4592 | | 1 | | | | | |
| SI | | 1.459 | | | 1.459 | | | 0.9183 | | 0.9183 | | 2.585 | | | | | |
| GR | | 0.3707 | | | 0.1423 | | | 0 | | 0.5001 | | 0.3868 | | | | | |

Skip to DT:




Classify G: Wind = T ⇒ N

Classify H: Wind = F → Outl = S ⇒ N

2. Imagine you are given a dataset from the university's library, and your job is to build a classification model that classify students based on the list of books that they borrowed.
The dataset includes the list of books available in the library (columns) and the students who borrowed them (rows), and the ranking for each item (ranking value is between 0–5, 0 if the book was not borrowed and 1–5 indicates the student's interest). The metadata for the books (e.g., titles) are not readily available to us, we just have the book IDs (e.g., Book #i). The dataset also includes the students' field of study (in total there are 10 fields), which can be used for the classification task. Answer the following questions, considering that there are 500,000 students and 100,000 books in this dataset.

predict

| Student ID | Book #1 | ⋯ | Book #100,000 | Label (Field of Study) |
|---|---|---|---|---|
| Student # 1 | 3 | ⋯ | 2 | Computer science |
| Student # 2 | 5 | ⋯ | 0 | Biology |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Student # 500,000 | 1 | ⋯ | 4 | Mathematic |

features | class

(i). Consider the following supervised machine learning methods, and for each one, explain why it would be appropriate or inappropriate to use for this problem:

i. Naïve Bayes
ii. k-NN
iii. Decision Tree

(i) NB : too many attributes ⇒ most probs based on numerical values will be essentially random.

Oversensitive to redundant / irrelevant attributes.

(ii) k-NN: too many dimensions ⇒ similarities are mostly meaningless & too many instances ⇒ computationally heavy → calculate similarity

(iii) DT: too many attributes ⇒ with too many nodes, DT could overfit

(ii). Would "feature selection" be useful here? Explain why, by referring to a single machine learning method.

Yes! Feature selection → improve any of the above approaches

k-NN: distances become meaningless in high-dim space

(everything is far away from everything)

NB: slightly more robust to irrelevant / many features.

BUT: no embedded feature weight mechanism
(e.g. logistic regression)

DT: prone to overfit, but you can prune the DT
(chop off branches with low IG)

↑ has embedded method for feature selection

(iii). Explain how you would evaluate the effectiveness of your system: you should briefly describe an evaluation strategy and an evaluation metric that are suitable for this data. What might be an example of a baseline?

Strategy: CV: partition data into M equal size portions
train for M times & take average performance

Metric: F1-Score / Accuracy

Baseline: O-R: majority class

Random-baseline: randomly assign a class