Updated 10/25/2025

Part 3: Applications
L13: MDL, Occam, Kolmogorov (1/2)

[A small MDL example for decision trees]

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

https://northeastern-datalab.github.io/cs7840/fa25/

10/23/2025

MDL
(Minimum Description Length)

Preference bias: Occam's Razor

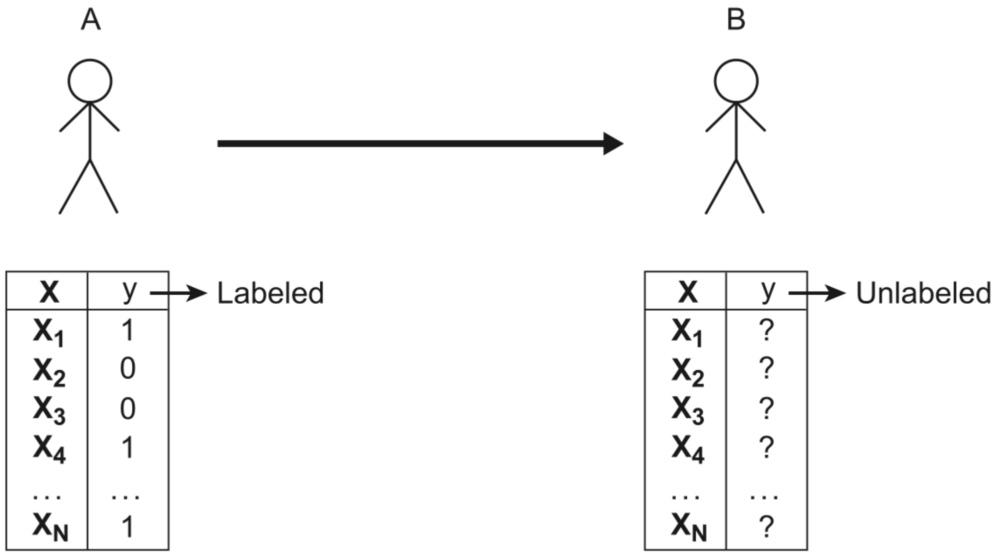
- Idea: The simplest consistent explanation is usually the best
- Principle attributed to William of Ockham (1285-1347)
  - "Entia non sunt multiplicanda praeter necessitatem"
    = "Entities must not be multiplied beyond necessity"
  - also known as "Ockham's Razor" and "principle of parsimony"

- For Decision Tree (DT) learning:
  - Given two DT's with the same generalization errors, the simpler one is preferred
  - Idea: adding some penalty for model complexity

See excellent discussion on Occam's Razor in Ch 3.6.2 of [Mitchell'97]. Introduction to Machine Learning, 1997. https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



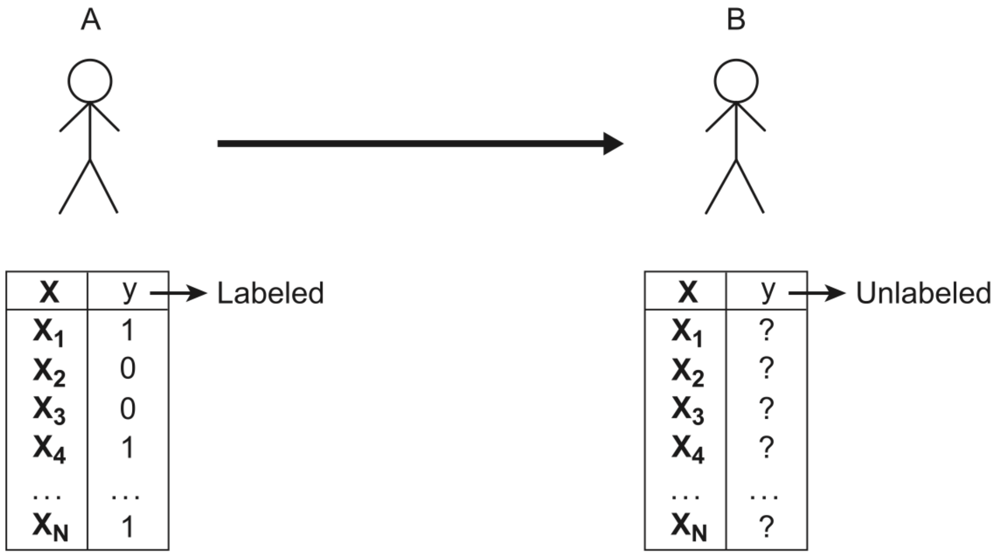
- Assume A and B are both given a set of instances with known attribute values x.
- Assume only person A also knows the class label y for every instance,
- A would like to share the class information with B by sending a message containing the labels.
- How many bits of information would such a message require?

?

Figure from [Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/fa25/

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



- Assume A and B are both given a set of instances with known attribute values x.
- Assume only person A also knows the class label y for every instance,
- A would like to share the class information with B by sending a message containing the labels.
- How many bits of information would such a message require?

$\Theta(n)$, where n is the total number of instances

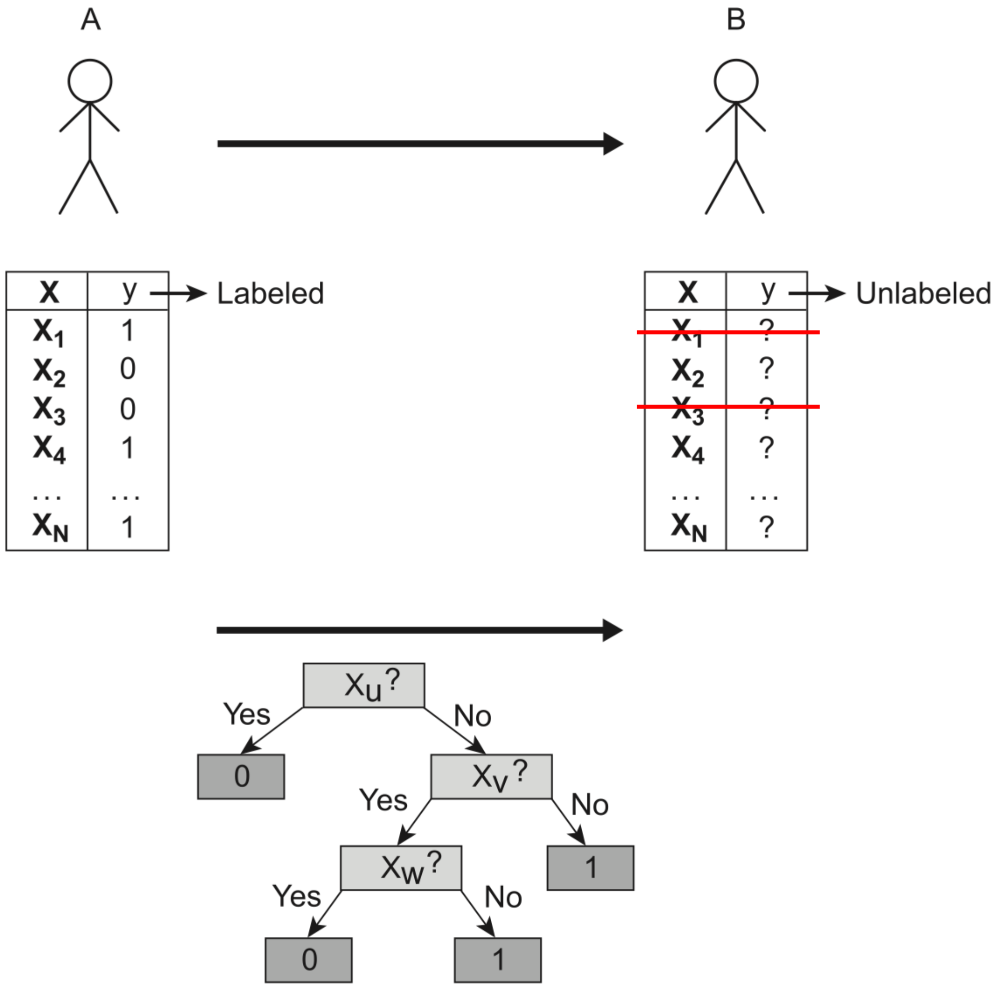
(big) theta: bounded from above and below:

- an upper bound (you can't need more than proportional to n bits), and
- a lower bound (you can't get away with fewer bits if labels are arbitrary).

Figure from [Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



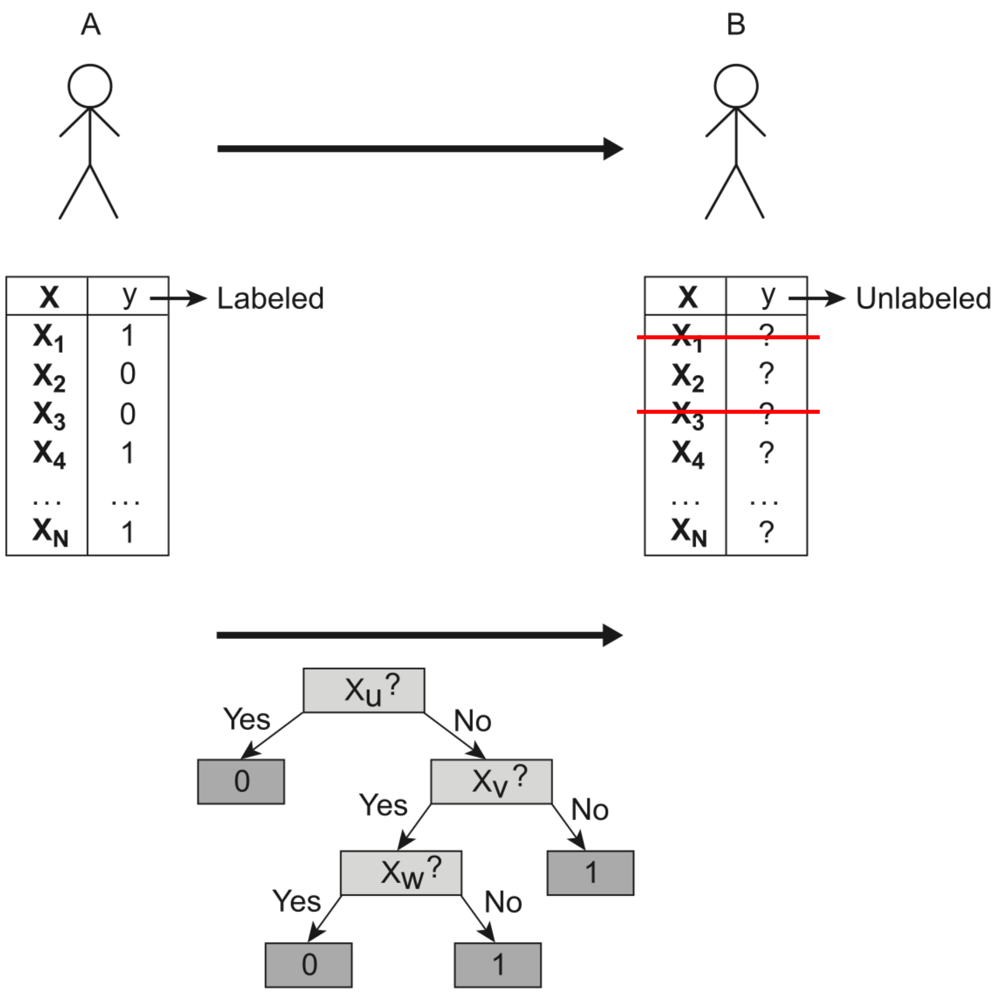
- Alternatively, A builds a DT from the instances and labels
- A transmits the DT to B
- B applies the DT to determine the class labels
- If the model is 100% accurate, then the transmission cost is just the number of bits required to encode the model.
- Otherwise, A must also transmit information about which instances are misclassified
- How big is the extra information needed assuming a fraction f of misclassified instances?

?

Figure from [Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/fa25/

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



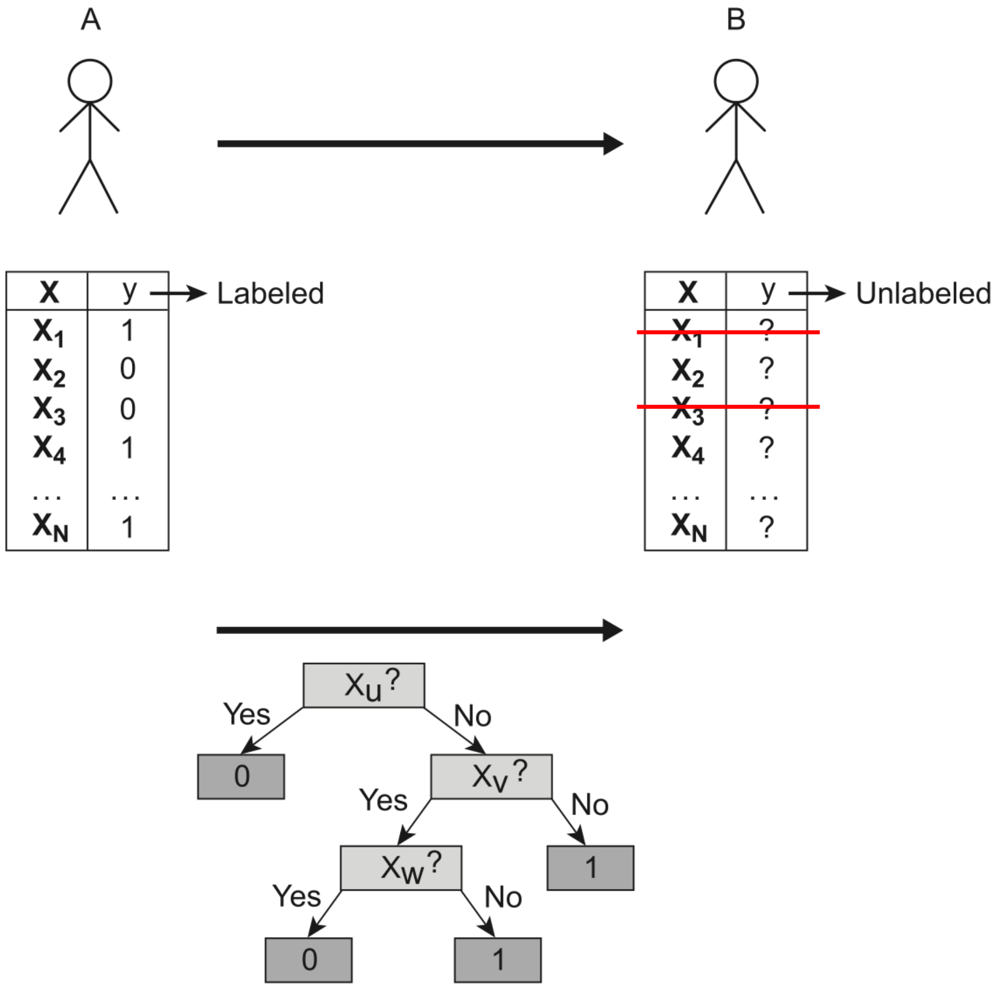
- Alternatively, A builds a DT from the instances and labels
- A transmits the DT to B
- B applies the DT to determine the class labels
- If the model is 100% accurate, then the transmission cost is just the number of bits required to encode the model.
- Otherwise, A must also transmit information about which instances are misclassified
- How big is the extra information needed assuming a fraction f of misclassified instances?

$$O(f \cdot n \cdot \lg n)$$

- Each misclassified instance must be identified among n (so its index needs $\log n$ bits).
- And there are $f \cdot n$ of those

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



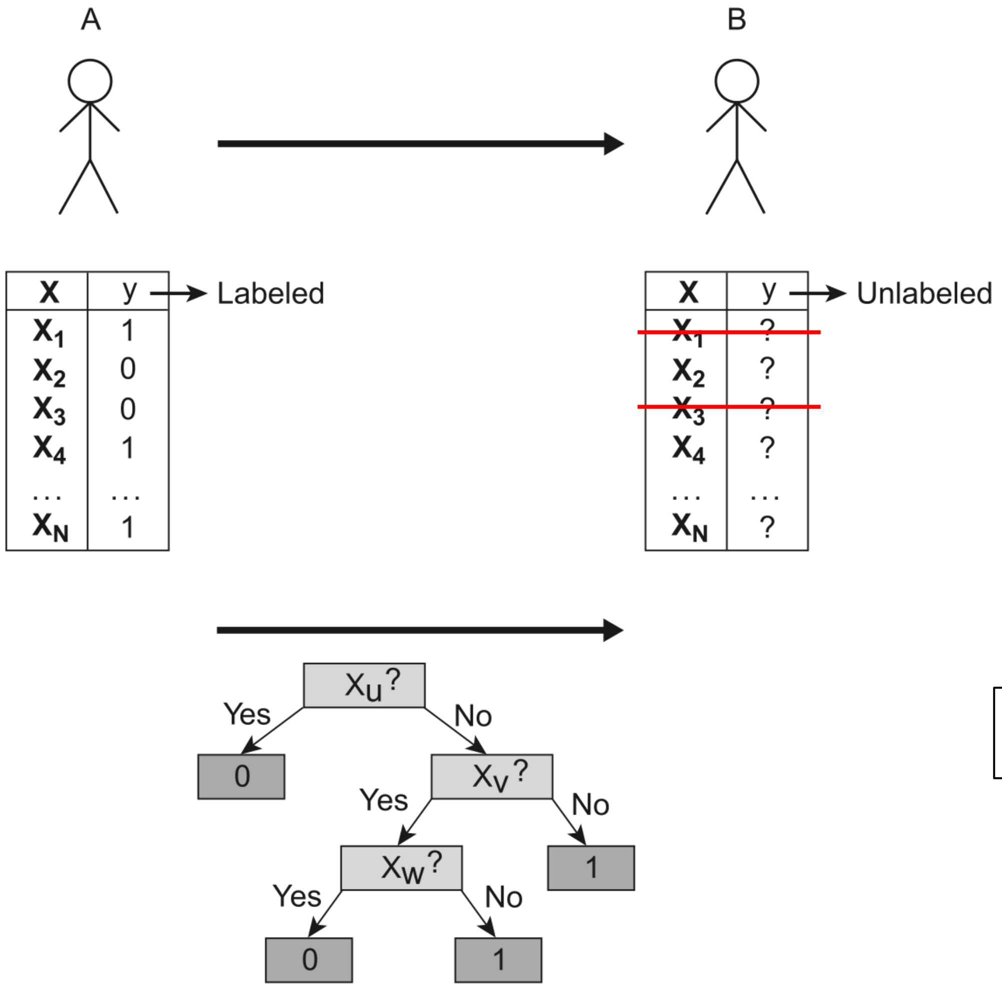
- Alternatively, A builds a DT from the instances and labels
- A transmits the DT to B
- B applies the DT to determine the class labels
- If the model is 100% accurate, then the transmission cost is just the number of bits required to encode the model.
- Otherwise, A must also transmit information about which instances are misclassified
- How big is the total **description length (DL)** of the message (= overall transmission cost)?

?

Figure from [Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/fa25/

Minimum Description Length (MDL)

MDL: an information-theoretic approach to incorporate model complexity



- Alternatively, A builds a DT from the instances and labels
- A transmits the DT to B
- B applies the DT to determine the class labels
- If the model is 100% accurate, then the transmission cost is just the number of bits required to encode the model.
- Otherwise, A must also transmit information about which instances are misclassified
- How big is the total **description length** (**DL**) of the message (= overall transmission cost)?

$$\boxed{cost(\text{DT}, \text{data}) = \alpha \cdot cost(\text{DT}) + cost(\text{data}|\text{DT})}$$

hyper-parameter for trade-off

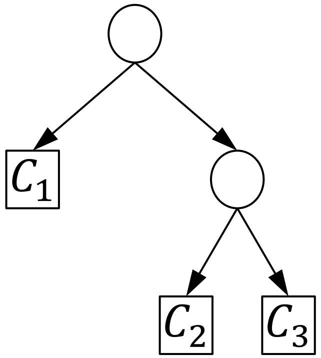
encoding of the model

encoding of misclassified instances

MDL Example

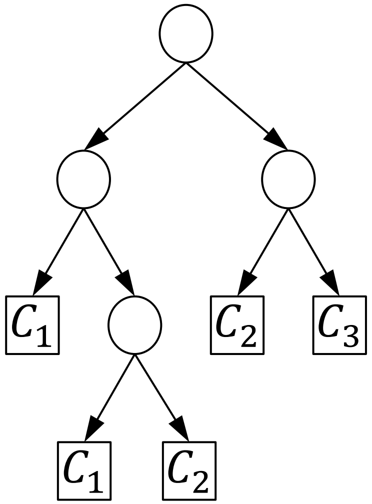
EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

DT 1



7 errors

DT 2



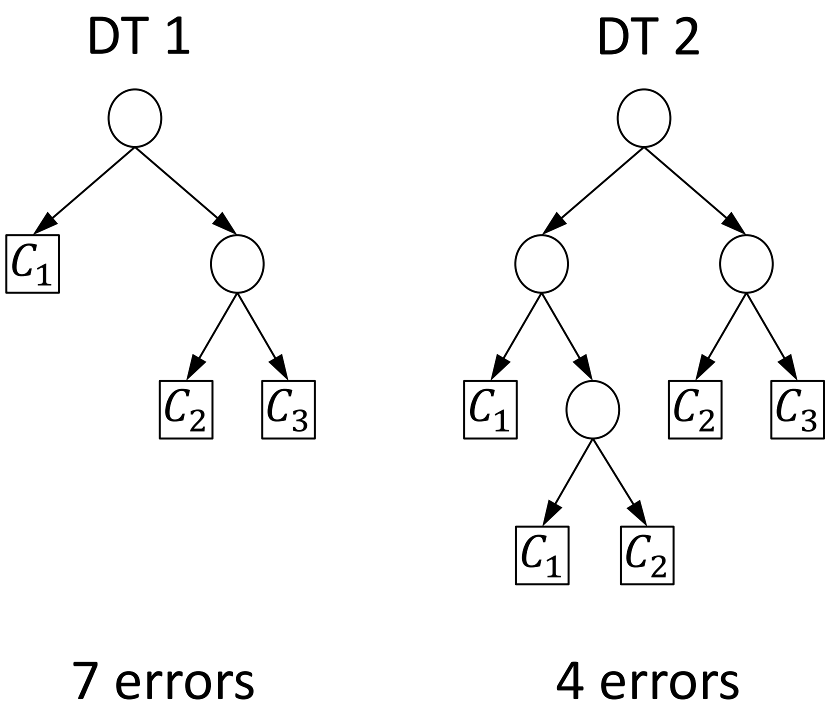
4 errors

?

[Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Example 10 in Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

MDL Example

EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.



- Total DL: $cost(\text{DT}, \text{data}) = cost(\text{DT}) + cost(\text{data}|\text{DT})$
- $\underline{cost(\text{DT})}$:

?

- $\underline{cost(\text{data}|\text{DT})}$:

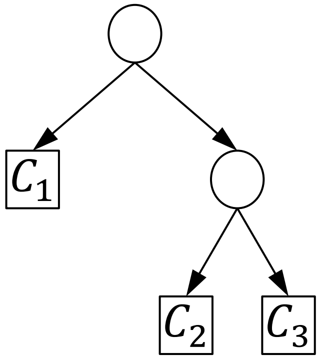
?

[Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Example 10 in Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

MDL Example

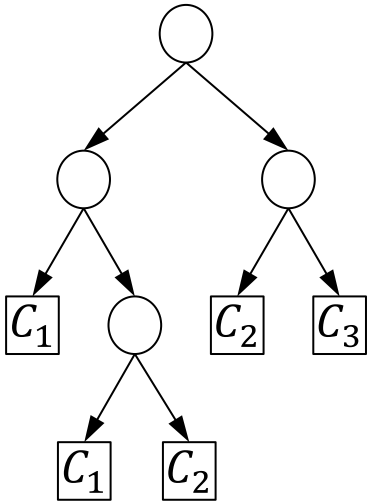
EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

DT 1



7 errors

DT 2



4 errors

- Total DL: $cost(\mathrm{DT}, \mathrm{data}) = cost(\mathrm{DT}) + cost(\mathrm{data}|\mathrm{DT})$
- $\underline{cost(\mathrm{DT})}$: cost of encoding all nodes and edges of DT

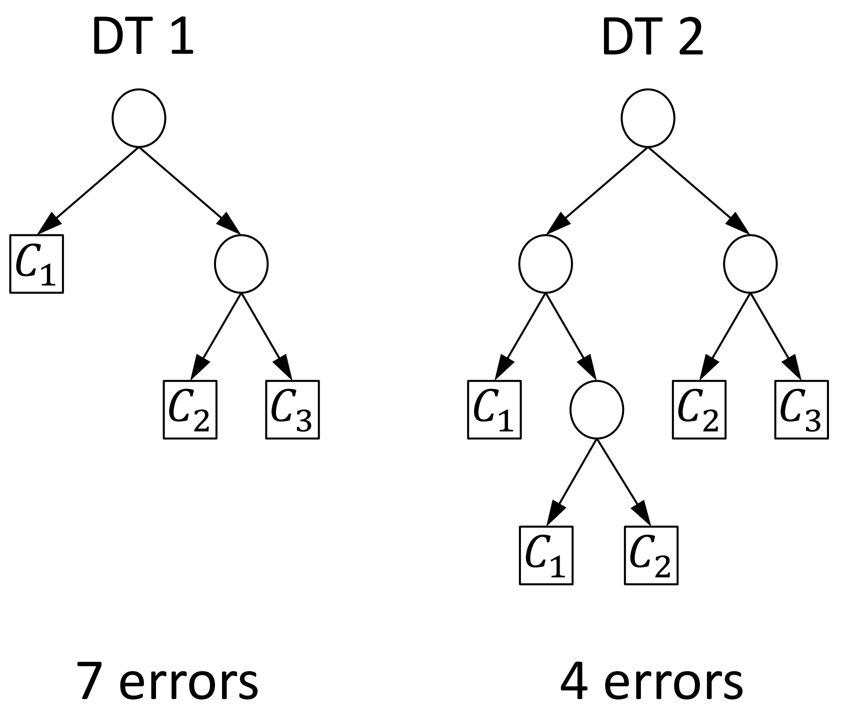
  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: ?
  - Encoding of a leaf node: ?
- $\underline{cost(\mathrm{data}|\mathrm{DT})}$:

?

MDL Example

EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.



- Total DL: $cost(\mathrm{DT}, \mathrm{data}) = cost(\mathrm{DT}) + cost(\mathrm{data}|\mathrm{DT})$
- $\underline{cost(\mathrm{DT})}$: cost of encoding all nodes and edges of DT

  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: by ID of splitting attribute

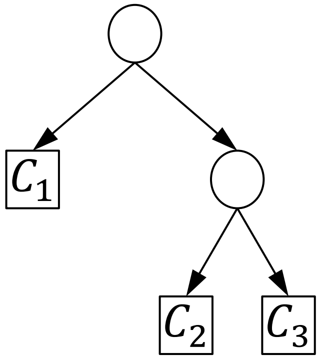
    cost per internal node: ?
  - Encoding of a leaf node: by ID of class

    cost per leaf node: ?
- $\underline{cost(\mathrm{data}|\mathrm{DT})}$: ?

MDL Example

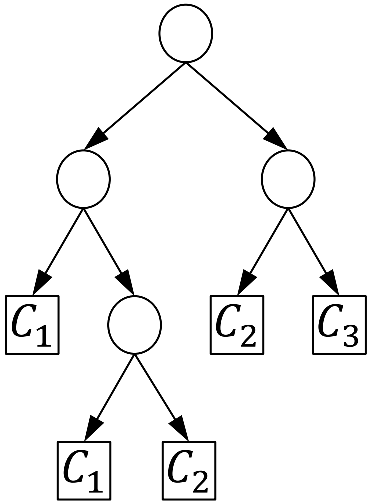
EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

DT 1



7 errors

DT 2



4 errors

- Total DL: $cost(\mathrm{DT}, \mathrm{data}) = cost(\mathrm{DT}) + cost(\mathrm{data}|\mathrm{DT})$
- $\underline{cost(\mathrm{DT})}$: cost of encoding all nodes and edges of DT

  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: by ID of splitting attribute

    cost per internal node: $\lg(m) = \lg(16) = 4$
  - Encoding of a leaf node: by ID of class

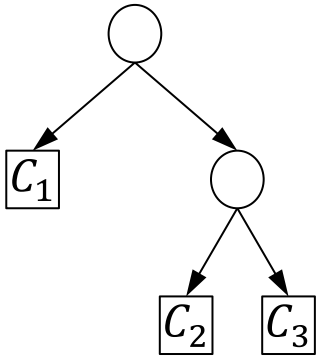
    cost per leaf node: $\lg(k) = \lceil \lg(3) \rceil = 2$
- $\underline{cost(\mathrm{data}|\mathrm{DT})}$:

?

MDL Example

EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

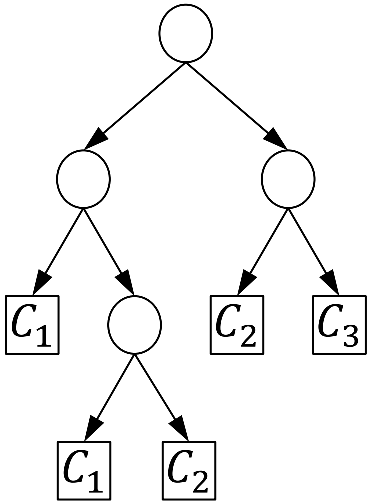
DT 1



7 errors

?

DT 2



4 errors

?

- Total DL: $cost(\mathrm{DT}, \mathrm{data}) = cost(\mathrm{DT}) + cost(\mathrm{data}|\mathrm{DT})$
- $\underline{cost(\mathrm{DT})}$: cost of encoding all nodes and edges of DT

  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: by ID of splitting attribute

    cost per internal node: $\lg(m) = \lg(16) = 4$
  - Encoding of a leaf node: by ID of class

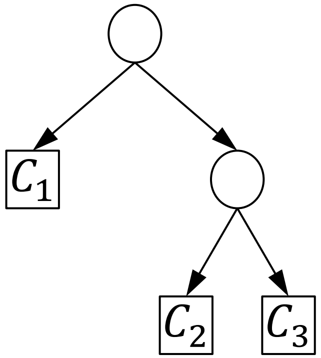
    cost per leaf node: $\lg(k) = \lceil \lg(3) \rceil = 2$
- $\underline{cost(\mathrm{data}|\mathrm{DT})}$: cost of encoding all erroneous data points

  cost per error: $\lg(n)$

MDL Example

EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

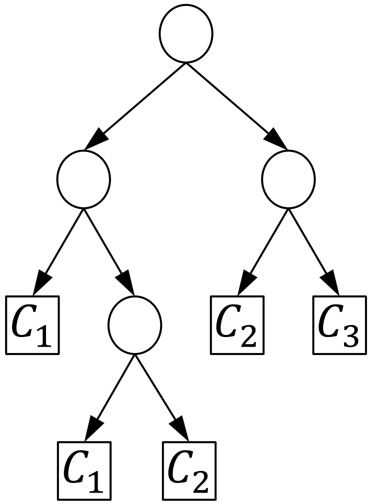
DT 1



7 errors

$$\underbrace{14}_{2\cdot 4+3\cdot 2} + 7 \cdot \lg(n)$$

DT 2



4 errors

$$\underbrace{26}_{4\cdot 4+5\cdot 2} + 4 \cdot \lg(n)$$

- Total DL: $cost(\text{DT}, \text{data}) = cost(\text{DT}) + cost(\text{data}|\text{DT})$
- $\underline{cost(\text{DT})}$: cost of encoding all nodes and edges of DT

  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: by ID of splitting attribute

    cost per internal node: $\lg(m) \quad = \lg(16) = 4$
  - Encoding of a leaf node: by ID of class

    cost per leaf node: $\lg(k) \quad = \lceil \lg(3) \rceil = 2$
- $\underline{cost(\text{data}|\text{DT})}$: cost of encoding all erroneous data points

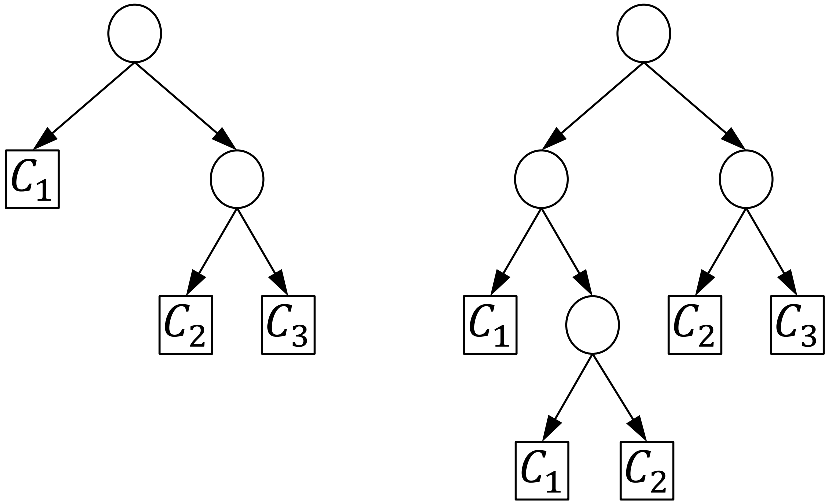
  cost per error: $\lg(n)$

[Tan+'18] Tan, Steinbach, Kumar. Introduction to Data Mining, 2nd ed, 2018. Example 10 in Chapter 3. https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/fa25/

MDL Example

EXAMPLE: Assume a dataset with $m = 16$ binary attributes, $k = 3$ classes $\{C_1, C_2, C_3\}$, and n tuples. Consider the following two DTs with their respective number of classification errors. Compare the total description length (DL) for the two DTs according to the MDL principle.

DT 1 DT 2



7 errors 4 errors

- Total DL: $cost(\mathrm{DT}, \mathrm{data}) = cost(\mathrm{DT}) + cost(\mathrm{data}|\mathrm{DT})$
- $\underline{cost(\mathrm{DT})}$: cost of encoding all nodes and edges of DT

  Simplification: we only add up the encoding costs for nodes

  - Encoding of an internal node: by ID of splitting attribute

    cost per internal node: $\lg(m) \quad = \lg(16) = 4$
  - Encoding of a leaf node: by ID of class

    cost per leaf node: $\lg(k) \quad = \lceil \lg(3) \rceil = 2$
- $\underline{cost(\mathrm{data}|\mathrm{DT})}$: cost of encoding all erroneous data points

  cost per error: $\lg(n)$

$$\underbrace{14}_{2 \cdot 4 + 3 \cdot 2} + 7 \cdot \lg(n) \quad > \quad \underbrace{26}_{4 \cdot 4 + 5 \cdot 2} + 4 \cdot \lg(n) \quad \text{for } n > 16$$