Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- 0001000100010001000 1 . . . 0001000100010001000100010001
- 0111010011010010011 0 . . . 1010111010111011000101100010
- 0001100000101010000 0 . . . 0010001000010000001000110000

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- 01110100110100100110 . . . 10101110101110110001011000 10
- 00011000001010100000 . . . 00100010000100000010001100 00

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- print ”0111010011010000101010101...10101110101110110001011000010”; halt
- 0001100000101010000000 . . . 0010001000010000001000110000

MDL : Fundamental idea

Learning as Data Compression

We assume that each sequence is 10000 bits long :

- for i = 1 to 2500; print ”0001”; next; halt
- print ”0111010011010000101010101...10101110101110110001011000010”; halt
- can be compressed to some length αn, with 0 < α < 1

MDL and model selection

- The goal of statistical inference may be cast as trying to find regularity in the data.
- For a given set of hypotheses H and data set D, we should try to find the hypothesis or combination of hypotheses in H that compresses D most.

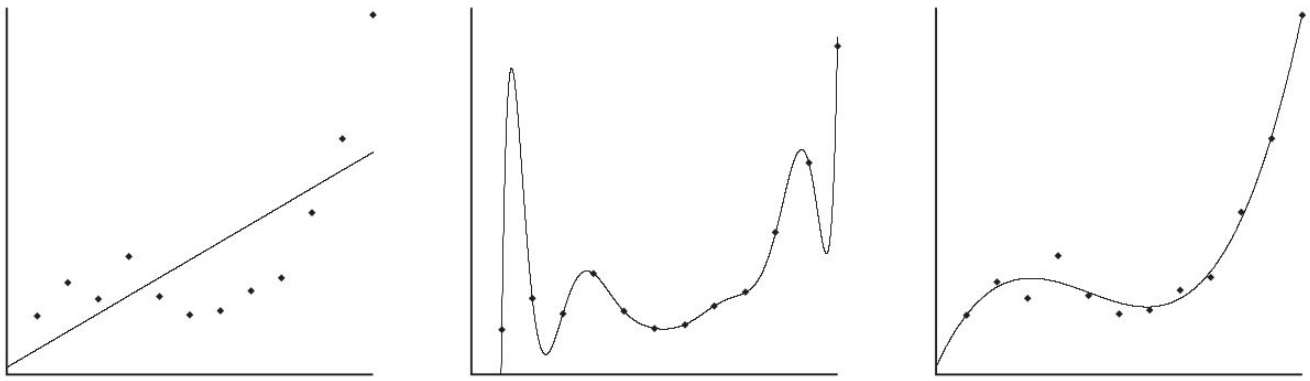
MDL and model selection

Crude , Two-Part Version of MDL principle (Informally Stated)
Let $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots$ be a list of candidate models (e.g., $\mathcal{H}^{(k)}$ is the set of kth-degree polynomials), each containing a set of point hypotheses (e.g., individual polynomials). The best point hypothesis $H \in \mathcal{H}^{(1)} \cup \mathcal{H}^{(2)} \cup \ldots$ to explain the data D is the one which minimizes the sum $L(H) + L(D|H)$, where

- $L(H)$ is the length, in bits, of the description of the hypothesis; and
- $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis.

The best *model* to explain D is the smallest model containing the selected H.

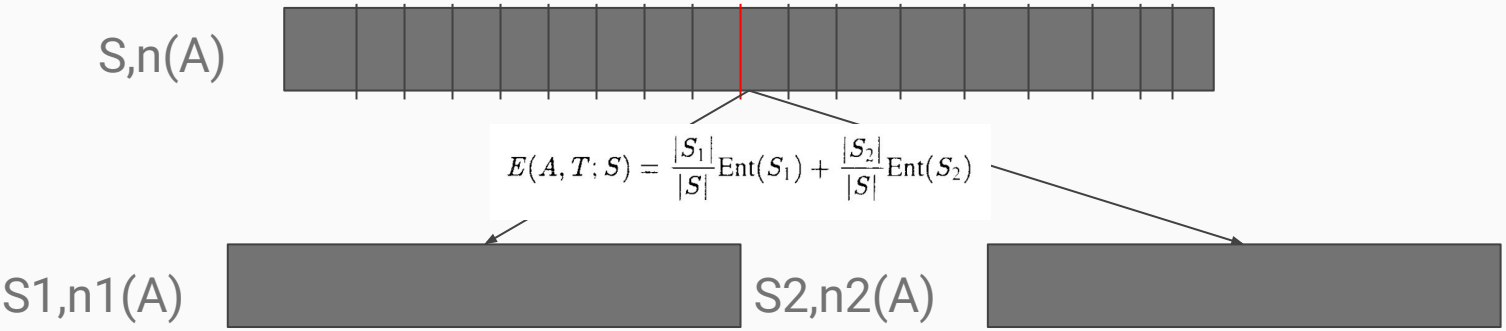
MDL and model selection



A simple, complex and tradeoff (third-degree) polynomial.

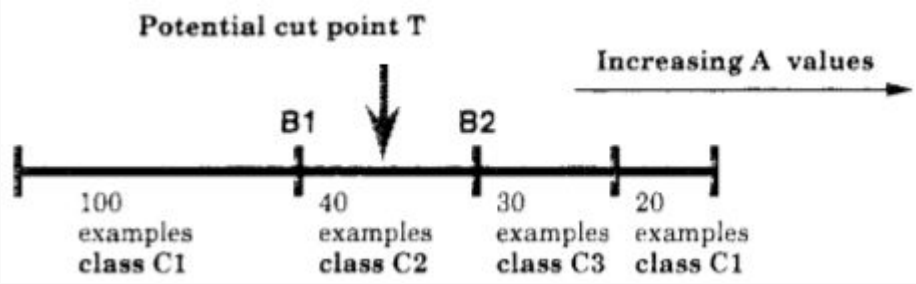
Fayyad & Irani

- Most real-world applications of classification learning involve continuous-valued attributes.
- The discretization process - crucial - often uses heuristics
- ID3, C4, CART use entropy minimization for binary discretization



Why multiple ranges rather than binary ranges ?

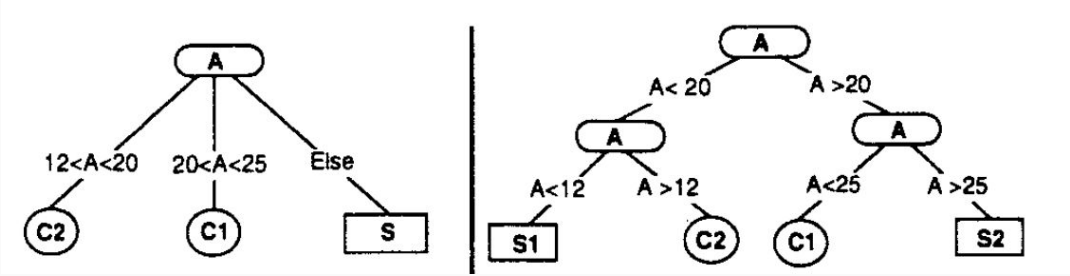
- Although polynomial in complexity, the selection criterion must be calculated N-1 times.
- Entropy minimizing cutpoints are always class boundaries



=>Motive for generalizing the algorithm to generate multiple intervals.

Why multiple ranges rather than binary ranges ?

- Unnecessary tree growth and irrelevant values problem.



- To cut or not to cut, that is the question !

Applying MDL, a coding problem.

- Coding the Null Theory:
  - Transmit the classes of the examples in S
  - Huffman(N messages) => total cost of transmitting S + overhead
- Coding the Partion Theory:
  - Specify the cut point, log2(N-1)
  - Sequence S1 then Sequence S2
- Cut if Cost(HT) < Cost(NT), the solution is readily available from Information theory.

```
>>> from my_module import MDLP
>>> from sklearn.datasets import
load_iris
>>> iris = load_iris()
>>> X = iris.data
>>> y = iris.target
>>> mdlp = MDLP()
>>> conv_X = mdlp.fit_transform(X, y)
```

What’s my micro-research ?
Implementation =)