# MDLP approach to clustering

Jacek Tabor

Jagiellonian University

Theoretical Foundations of Machine Learning 2015

# Ockham's razor

*Among competing hypotheses, the one with the fewest assumptions should be selected. Other, more complicated solutions may ultimately prove correct, but—in the absence of certainty—the fewer assumptions that are made, the better.*

Figure: William of Ocham 1287–1347.

# Morse code

Morse Code/telegraph [1836 S. Morse, A. Vail]: Vail determined the frequency of use of letters in the English language by counting the movable type he found in the type-cases of a local newspaper in Morristown: *the code of typical text should be as short as possible*, i.e.: the more common the letter, the shorter should be its code.

# Morse code

Morse Code/telegraph [1836 S. Morse, A. Vail]: Vail determined the frequency of use of letters in the English language by counting the movable type he found in the type-cases of a local newspaper in Morristown: *the code of typical text should be as short as possible*, i.e.: the more common the letter, the shorter should be its code.

ANECDOTE: Ernest Hemingway [newspaper notes from war in Spain]

# Morse code

Morse Code/telegraph [1836 S. Morse, A. Vail]: Vail determined the frequency of use of letters in the English language by counting the movable type he found in the type-cases of a local newspaper in Morristown: *the code of typical text should be as short as possible*, i.e.: the more common the letter, the shorter should be its code.

ANECDOTE: Ernest Hemingway [newspaper notes from war in Spain] READER (with admiration): how is it possible to learn to write the notes so short, but so wonderfully informative?

# Morse code

Morse Code/telegraph [1836 S. Morse, A. Vail]: Vail determined the frequency of use of letters in the English language by counting the movable type he found in the type-cases of a local newspaper in Morristown: *the code of typical text should be as short as possible*, i.e.: the more common the letter, the shorter should be its code.

ANECDOTE: Ernest Hemingway [newspaper notes from war in Spain]
READER (with admiration): how is it possible to learn to write the notes so short, but so wonderfully informative?
HEMINGWAY: easy – you just have to pay one dollar for each word send by the telegraph

# Morse code

Morse Code/telegraph [1836 S. Morse, A. Vail]: Vail determined the frequency of use of letters in the English language by counting the movable type he found in the type-cases of a local newspaper in Morristown: *the code of typical text should be as short as possible*, i.e.: the more common the letter, the shorter should be its code.

ANECDOTE: Ernest Hemingway [newspaper notes from war in Spain]
READER (with admiration): how is it possible to learn to write the notes so short, but so wonderfully informative?
HEMINGWAY: easy – you just have to pay one dollar for each word send by the telegraph

Information=money

# Entropy

C. Shannon [1948. "A Mathematical Theory of Communication". Bell System Technical Journal 27] Precise formulation of the idea of Morse leads to the formal definition of Shannon's entropy: *in the optimal coding (that is those with the shortest code-length), if the signal appears with probability p, its code-length should equal to* $-\log_2 p$.

# Entropy

C. Shannon [1948. "A Mathematical Theory of Communication". Bell System Technical Journal 27] Precise formulation of the idea of Morse leads to the formal definition of Shannon's entropy: *in the optimal coding (that is those with the shortest code-length), if the signal appears with probability p, its code-length should equal to* $-\log_2 p$.

Thus if the symbols $x_1, \ldots, x_n$ appear with probabilities $p_1, \ldots, p_n$:

$$\text{entropy} = \text{minimal code length} = p_1 \cdot -\log_2 p_1 + \ldots + p_n \cdot -\log_2 p_n.$$

In practice leads to Huffman's coding (used for example in jpg).

# Minimum description length (MDL) principle

Formulation of the MDL principle Jorma Rissanen [1978. "Modeling by shortest data description". Automatica 14]: the best hypothesis for a given set of data is the one that leads to the best compression of the data.

# Minimum description length (MDL) principle

Formulation of the MDL principle Jorma Rissanen [1978. "Modeling by shortest data description". Automatica 14]: the best hypothesis for a given set of data is the one that leads to the best compression of the data.

P. Grünwald, [1998]: "any regularity in a given set of data can be used to compress the data, i.e. to describe it using fewer symbols than needed to describe the data literally."

# Minimum description length (MDL) principle

Formulation of the MDL principle Jorma Rissanen [1978. "Modeling by shortest data description". Automatica 14]: the best hypothesis for a given set of data is the one that leads to the best compression of the data.

P. Grünwald, [1998]: "any regularity in a given set of data can be used to compress the data, i.e. to describe it using fewer symbols than needed to describe the data literally."

Connected to the notion of Kolmogorov complexity [1963. "On Tables of Random Numbers". Sankhya Ser. A. 2]: the complexity of a string is the length of the shortest possible description of the string in some fixed universal description language.

# How many clusters?

In most clustering methods, one has to specify the number of clusters. This implies, that the procedure does not return the right number of clusters (or reduce unnecessary clusters on-line during the clustering procedure).

# How many clusters?

In most clustering methods, one has to specify the number of clusters. This implies, that the procedure does not return the right number of clusters (or reduce unnecessary clusters on-line during the clustering procedure).

Advantages of the use of MDL principle in clustering:

- automatically reduces the complexity of the model (number of clusters)
- has high adaptability
- (often) small requirements on the the data: (we do not require vector or metric structures, but only the existence of encoding or compression methods)
- easily allows potential modifications

# MDL principle in clustering

***Requirements***

Data type which we want to cluster, available compression methods $\mathcal{W}$.

# MDL principle in clustering

***Requirements***

Data type which we want to cluster, available compression methods $\mathcal{W}$.

***Determination of the overall cost of the memory***

Let us assume that message $X = (x_1, \ldots, x_N)$ and the compression methods $w_1, \ldots, w_K \in \mathcal{W}$ are given. We denote the algorithm that encodes point $x_l$ (defines the cluster it belongs to) with $w_{k_l}$, where $k_l \in \{1, \ldots, K\}$.
Then the memory cost of coding the message $X$ equals

$$\sum_{i=1}^{K} \text{memory cost of } w_k + \sum_{l=1}^{N} (\text{cost of identification of } k_l +$$
the amount of memory algorithm $w_{k_l}$ uses to code $x_l$).

# MDL principle in clustering

***Requirements***

Data type which we want to cluster, available compression methods $\mathcal{W}$.

***Determination of the overall cost of the memory***

Let us assume that message $X = (x_1, \ldots, x_N)$ and the compression methods $w_1, \ldots, w_K \in \mathcal{W}$ are given. We denote the algorithm that encodes point $x_l$ (defines the cluster it belongs to) with $w_{k_l}$, where $k_l \in \{1, \ldots, K\}$.
Then the memory cost of coding the message $X$ equals

$$\sum_{i=1}^{K} \text{memory cost of } w_k + \sum_{l=1}^{N} (\text{cost of identification of } k_l +$$
the amount of memory algorithm $w_{k_l}$ uses to code $x_l$).

***Memory minimization step***

We seek $K$, compression methods $w_1, \ldots, w_K$ and indices $k_l$, such that the total cost needed to encode the message $X$ is minimal.

# MDL principle in clustering

***Requirements***

Data type which we want to cluster, available compression methods $\mathcal{W}$.

***Determination of the overall cost of the memory***

Let us assume that message $X = (x_1, \ldots, x_N)$ and the compression methods $w_1, \ldots, w_K \in \mathcal{W}$ are given. We denote the algorithm that encodes point $x_l$ (defines the cluster it belongs to) with $w_{k_l}$, where $k_l \in \{1, \ldots, K\}$.
Then the memory cost of coding the message $X$ equals

$$\sum_{i=1}^{K} \text{memory cost of } w_k + \sum_{l=1}^{N} (\text{cost of identification of } k_l +$$
the amount of memory algorithm $w_{k_l}$ uses to code $x_l$).

***Memory minimization step***

We seek $K$, compression methods $w_1, \ldots, w_K$ and indices $k_l$, such that the total cost needed to encode the message $X$ is minimal.

***Construction of the clustering***

Points which are coded by the same algorithm are assigned to the same cluster.

# MDL principle in clustering

Observe that in the above approach the use of any encoding method takes memory (needed for its determination), as a consequence we obtain the upper limit for the amount of possible clusters. Moreover, in practice – when the amount of elements encoded by a given algorithm is small – it will be worthwhile to give up the use of this algorithm, which leads to a reduction of the complexity of the constructed model (understood as a number of clusters).

# Differential entropy

The coder adapted to the data generated by the continuous random variable with the density $f$.

Differential entropy is the limiting case of entropy (smaller and smaller bins):

$$h(f) := \int f(x) \cdot -\log_2(f(x))dx.$$

# Differential entropy

The coder adapted to the data generated by the continuous random variable with the density $f$.

Differential entropy is the limiting case of entropy (smaller and smaller bins):

$$h(f) := \int f(x) \cdot -\log_2(f(x))dx.$$

Cross-entropy

$$h(f) := \int g(x) \cdot -\log_2(f(x))dx.$$

Gives the memory cost of compressing the data with the density $g$ by the coder optimally adapted to the density $f$.

# Gaussian models

In practice, we can compute the above for the Gaussian models!

# Gaussian models

In practice, we can compute the above for the Gaussian models!

There is only the need for the knowledge of the mean of the data and the covariance matrix.

# Papers

TABOR, SPUREK:
*Cross-entropy clustering,* Pattern Recognition 2014
TABOR, MISZTAL:
*Detection of elliptical shapes via cross-entropy clustering (IbPRIA 2013)*
SPUREK, TABOR, ZAJĄC:
*Detection of disk-like particles in electron microscopy images (CORES 2013)*
ŚMIEJA, TABOR:
*Image segmentation with use of cross-entropy clustering (CORES 2013)*

# EM vs CEC

[Geoffrey McLachlan,Thriyambakam Krishnan *The EM Algorithm and Extensions*]

Fit the data $X$ by

$$X \sim p_1 f_1 + \ldots + p_k f_k,$$

where $f_i$ belong to the fixed density family.

CEC:

$$X \sim \max(p_1 f_1, \ldots, p_k f_k),$$

# Cluster reduction

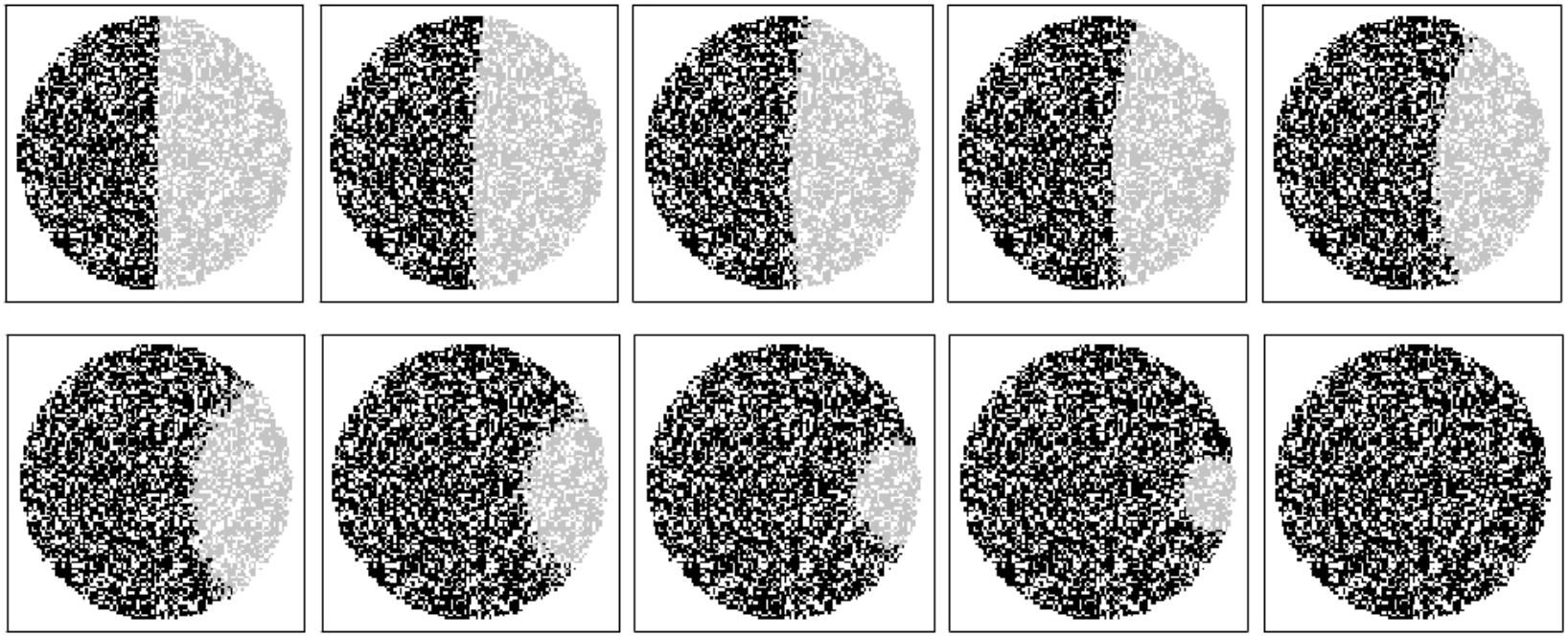

Figure: Cluster reduction.