Computational Intelligence: Methods and Applications

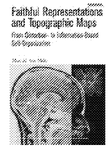
Lecture 11

Multi-Dimensional Scaling & SOM

Włodzisław Duch

Google: Duch

Problems with SOM



- SOM and its variants become so popular because they allow for visualization of data clusters, unsupervised classification.
- SOM does not guarantee faithful approximation to probability density of data samples (too many nodes are placed in areas with a few vectors only), but there are ways to correct it:
  Ex: M. Van Hulle, Faithful Representations and Topographic Maps From Distortion to Information-based Self-organization. Wiley 2000
- Quantization error is useful as an estimate of the quality of approximation by SOM, but it is not useful to measure distortions in visualization.
- How to visualize clusters in topographically correct way?
  Multi-Dimensional Scaling, MDS, is our best answer.

MDS: idea

- MDS, Multi-Dimensional Scaling, Sammon mapping (Thorton 1954, Kruskal 1964, Sammon 1964)
  Inspirations mostly from psychology, or psychometrics!
  Problem: visual representation of dissimilarity data.
- Information about the clusters in data is in relations between maxima of probability densities of finding data vectors.
- Visualization of multidimensional data leads to some topographical distortions, so a measure of these distortions is required.
- How different are the distances in the original, *d*-dimensional space, and in the target space, usually 2-dimensional?

MDS algorithm

- Data (feature) space $X \in R^d$, vectors $X^{(i)}$, i=1.. *n* represented by points Y in the target space, usually $Y \in R^2$.
- Distances $R_{ij} = ||X^{(i)}-X^{(j)}||$ between $X^{(i)}$ and $X^{(j)}$ in feature space R^d;

  distances $r_{ij} = ||Y^{(i)}-Y^{(j)}||$ in target space R^2.
- Find mapping $X \rightarrow Y=M(X)$ that minimizes some global index of topographical distortion, based on differences between R_{ij} and r_{ij}.

$$E(\mathbf{r}) = \sum_{i>j}^{n} \left(R_{ij} - r_{ij}\right)^2 = \sum_{i>j}^{n} \left(R_{ij} - \sqrt{\left(Y_1^{(i)} - Y_1^{(j)}\right)^2 + \left(Y_2^{(i)} - Y_2^{(j)}\right)^2}\right)^2$$

$\mathbf{r}=\mathbf{r}(\mathbf{Y})$, so $E(\mathbf{r})$ depends on $2n$ adaptive parameters **Y**;
3 are redundant: $Y^{(1)}=(0,0)^T$, choosing coordinate center,
and $Y_1^{(2)}=0$ choosing orientation of the coordinate center.

Initialize **Y** randomly or using PCA, minimize $E(\mathbf{r}(\mathbf{Y}))$

General distortion measures

Any non-negative function $f(R_{ij}, r_{ij})$ may be used.

$$E(\mathbf{r};\mathbf{W}) = \sum_{i>j}^{n} W_{ij}\left(R_{ij} - r_{ij}\right)^2 \geq 0$$

Weights may depend on the distance, for example decreasing exponentially for large R_{ij} distances.
Various normalization factors may be introduced, but those that do not depend on the r_{ij} distances have no influence on the MDS map.

$$A(\mathbf{r}) = \frac{\sum_{i>j}^{n}\left(R_{ij} - r_{ij}\right)^2}{\sum_{i>j}^{n} R_{ij}^2 + \sum_{i>j}^{n} r_{ij}^2} \in [0,1]$$

Information loss;
0 = no loss, perfect representation,
1 = all lost, no info, ex. r_{ij}=0

Topographical distortion measures

Stress (Kruskal), or absolute error, large distances may dominate, preserves overall cluster structure.

$$S_1(\mathbf{r}) = \sum_{i>j}^{n}\left(R_{ij} - r_{ij}\right)^2 \geq 0$$

Sammon measure, or intermediate error, contributions from large distances is reduced.

$$S_2(\mathbf{r}) = \sum_{i>j}^{n}\frac{\left(R_{ij} - r_{ij}\right)^2}{R_{ij}} \geq 0$$

Relative error, all distance scales treated in the same way.

$$S_3(\mathbf{r}) = \sum_{i>j}^{n}\left(1 - r_{ij}/R_{ij}\right)^2 \geq 0$$

Alienation coefficient (Guttman-Lingoes) – similar to relative error, but more difficult to minimize.

$$E_a(\mathbf{r}) = \sum_{i>j}^{n}\left(1 - R_{ij}/r_{ij}\right)^2 \geq 0$$

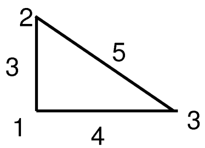
Metric and non-metric MDS

MDS does not require original X vectors, only R_{ij} distances, that may be interpreted as dissimilarities of $X^{(i)}$ and $X^{(j)}$ objects.

Gradient iterative minimization procedures for MDS may be formulated, but the problem is difficult: $2n$-3 variables for minimization, stress and other functions have multiple minima.

Sometimes its easy

$$R_{ij} = \begin{bmatrix} 0 & 3 & 4 \\ 3 & 0 & 5 \\ 4 & 5 & 0 \end{bmatrix}$$



Some measurements (especially in psychology) are very inaccurate.
Instead of precise distances ordinal, or non-metric, information may be used, if ranking is established:

X_1 is more similar to X_2 than to X_3.

Measures of distortions use $D_{ij} = F(R_{ij})$, where $F()$ may be any function preserving the ranking.

MDS vs. SOM

Questions:

1) How good are SOM maps (in the sense of MDS measures)? (not quite fair, because SOM does not optimize these!)

2) How optimal maps look like ?

Problem: there is no function mapping X → Y !
Adding new point on the map requires new minimization;
mapping function may be written only for a fixed number of points.

Maps with large stress: zoom on smaller areas, avoid large topographical distortions.

SOM does classification and visualization, but separating these to functions may improve accuracy of both.
MDS and SOM – sensitive to noise in data (irrelevant features).

MDS vs. PCA

MDS may start from random positions of points, moving them to minimize the measure of topographical distortions, results may strongly depend on the starting point.

Multi-starts are recommended, followed by selection of the best configuration (with lowest stress value).

PCA may be a good starting point for MDS, although PCA is only a linear method and does not preserve distances between points.

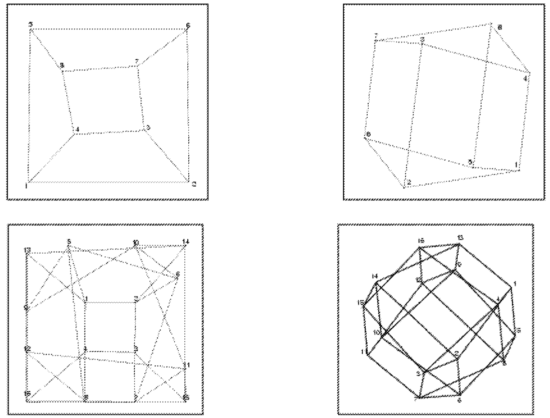
For n data vectors and d dimensions:

PCA requires building the dxd covariance matrix, with complexity $O(nd^2)$ and diagonalization of this matrix, complexity $O(d^3)$.

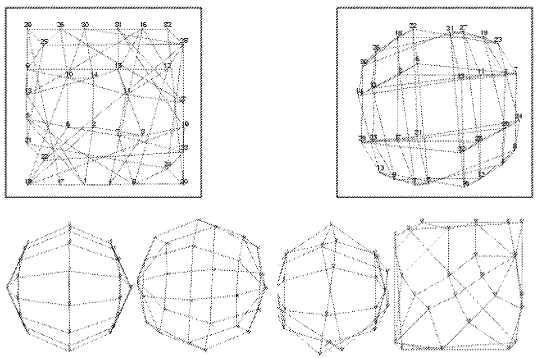
MDS requires calculation of a distance matrix $O(n^2d)$ and minimization procedure of $2n$-3 parameters.

Some data has d=10^4, and n=100;
other data has d=10, and n=10^4 ... which method to use?

Hipercubes

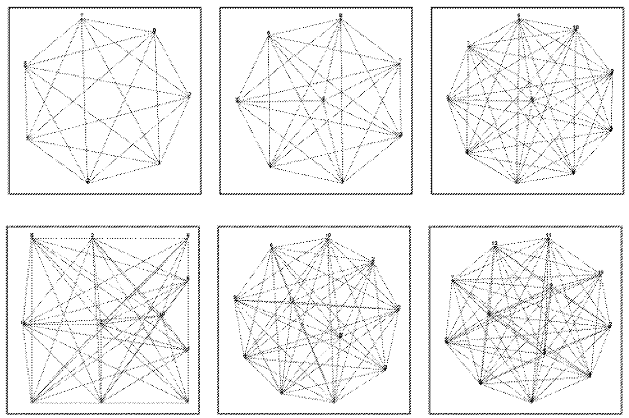


Hipercubes 5D + sphere in 3D

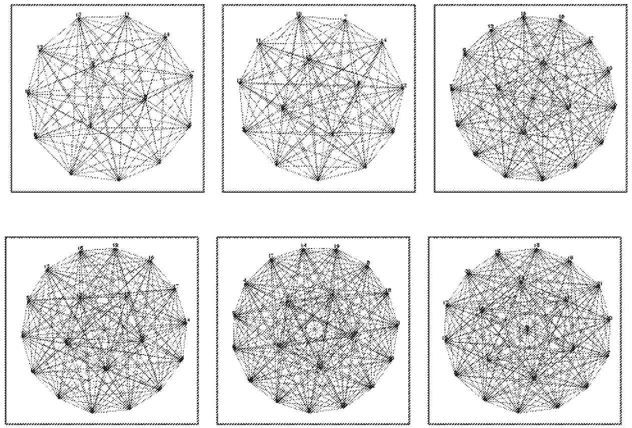


The two-dimensional representations of the 26 points on the sphere obtained by minimization of S, E, A, and by SOFM (left to right) with a 20 x 20 neurons map.

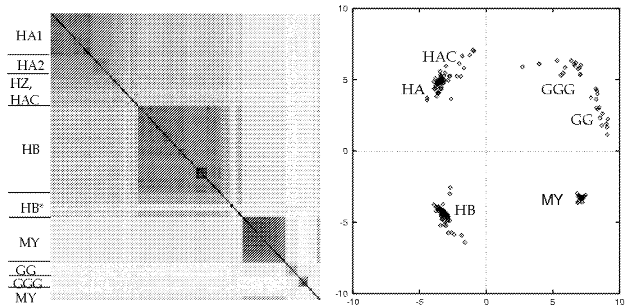
Simplexes 6-11



Simplexes 15-20D

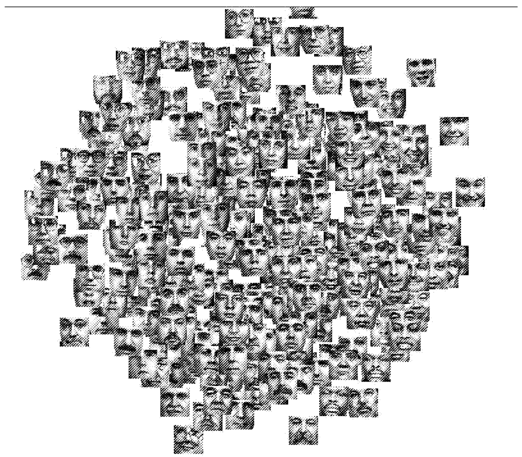


Sequences of the Globin family



226 protein sequences of the Globin family; similarity matrix shows high similarity values (dark spots) within subgroups, MDS shows cluster structure of the data (from Klock & Buhmann 1997).

Similarity of faces



300 faces, similarity matrix evaluated and Sammon mapping applied (from Klock & Buhmann 1997).

Semantic maps

How to capture the meaning of words/concepts?
It is contained in properties of concepts, their semantic relations.

Simplest example - take 8 birds and 8 mammals:
dove, hen, duck, goose, owl, hawk, eagle,
fox, dog, wolf, cat, tiger, lion, horse, zebra, cow.

Create some concept descriptions, for example use 13 binary features:

- size is: small, medium large;
- has 2 legs, 4 legs,
- has hair, hoofs, mane, feathers;
- likes to: hunt, run, fly, swim.

Form sentences that describe these 16 animals using 13 features.
More complex description should give more relations and features.

Sentences about animals

Dove is small, has 2 legs, has feathers, likes to fly.
Hen is small, has 2 legs, has feathers.
Duck is small, has 2 legs, has feathers, likes to swim.

Horse is big, has 4 legs, mane, hair, hoofs, likes to run.

This is a common sense knowledge that may be represented in a vector space and collected in data table.

Assign a vector of properties to each animal:

$V^{(horse)}$= [small=0,medium=0,large=1,has 2 legs=0, 4 legs=1, ...]
= [0,0,1,0,1,1,1,1,0,0,1,0,0]

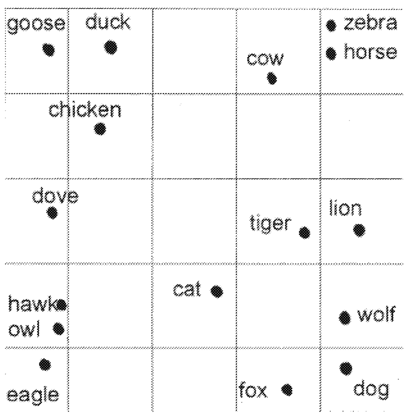
Map these 13D vectors in 2D to see relations between them.

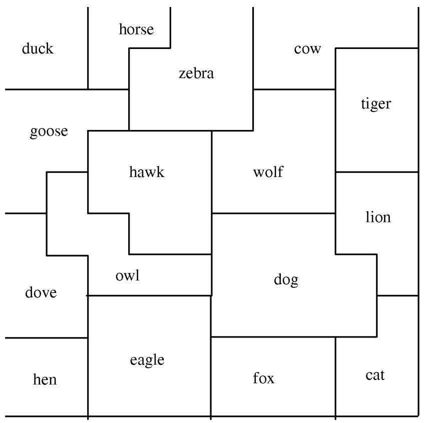
Semantic map example

	animal	dove	hen	duck	goose	owl	hawk	eagle	fox	dog	wolf	cat	tiger	lion	horse	zebra	cow
is	small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
has	2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
	4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
	hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
	feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
likes to	hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
	run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
	fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
	swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

SOM was used on such data by Ritter and Kohonen 1989,
MDS by Naud & Duch (1996).

Semantic maps: MDS & SOM





MDS and SOM was used on data vectors from the previous page.

Semantic maps - comment

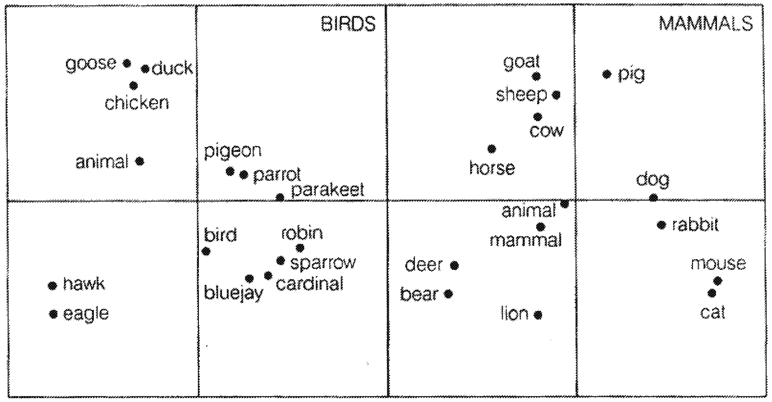
Natural classification of animals, preserving their similarity:
- mammals are separated from birds,
- hunters from grass-eaters,
- small animals separated from large.

Similarity of animals is related to the distance on toposemantic map.

Psychological experiments provide a matrix of perceived similarities: this is derived either from frequencies of answers to questions like: what other animal is similar to horse?
Sometimes response times to questions are also measured.

Information Retrieval: use concepts spaces, properties derived automatically from context information, to set up vectors representing words, phrases, concepts.
This allows for semantic retrieval of information.

MDS on similarity data



From Rips, Shoben, and Smith (1973)

MDS was used on data from psychological experiments with perceived similarity between animals: note that it is very similar to previous maps.

Some MDS examples

Demonstrations:

some examples of MDS maps made with the GhostMiner software

http://www.fqspl.com.pl/ghostminer/

Few lessons from visualization of:

- parity data – shows structure that is not visible with other methods, but very sensitive to starting point;
- medical data (from Lancet) – shows subgroups of different types of cancer, that can be labeled and identified;
- liver diseases 4-class data – show that same patients were used several times, leading to excellent results from the nearest neighbor methods;
- DNA promoters data – shows the importance of feature transformations.

MDS books and resources

Schiffman S.M, Reynolds L, Young F.W.
Introduction to multidimensional scaling. Academic Press 1981

Cox T.F, Cox M.A.A,
Multidimensional Scaling, 2nd ed, Chapman 2001

Borg, I. and Groenen, P.J.F.
Modern multidimensional scaling. 2nd ed, Springer 2005.

MDS links to papers and software:

http://www.cs.technion.ac.il/~mbron/research_mds.html

Elementary introduction to MDS.

MDS algorithms described in details.

Assignment !

Now is the time for the first assignment!

Many other visualization methods exist; some will be mentioned later when we are ready to understand what they do.

See links on the assignment page + search for interesting methods using as keywords names of methods already covered, and in addition:

- isometric mapping;
- principal curves and principal graphs;
- stochastic proximity embedding;
- non-linear dimensionality reduction;
- reducing the dimensionality of data with neural networks
- etc.