STATISTICAL LEARNING SYSTEMS

LECTURE 9: FINDING STRUCTURE IN DATA - contd.

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The project is co-financed by the European Union within the framework of European Social Fund

Independent component analysis (ICA)

We start with the model

$$\mathbf{x} = \mathbf{\Delta}\mathbf{z} + \mathbf{e},$$

where we not only assume that the data are centered, but also that the $x^{(i)}$, $i = 1, \ldots, p$, have unit variance and are uncorrelated. Moreover, for simplicity, we assume that $\mathbf{\Delta}$ is a $p \times p$ matrix, $\mathbf{e} = \mathbf{0}$ and none of the $x^{(i)}$ has normal distribution. In fact, therefore, our model is:

$$\mathbf{x} = \mathbf{\Delta}\mathbf{z}. \tag{1}$$

The task is to find such $\mathbf{\Delta}$ that the $z^{(i)}$ are mutually independent ($\mathbf{\Delta}$ is nonestimable if at least 2 of the $z^{(i)}$ are normal). Clearly, since we assume that the data are spherical, $\mathbf{\Delta}$ is orthogonal, and hence, once found, we have

$$\mathbf{z} = \mathbf{\Delta}'\mathbf{x}. \tag{2}$$

PhD STUDIES KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Independent component analysis (ICA)

Recall that the entropy of a random p-vector $\mathbf{z}$ with joint density $f(\mathbf{z})$ is (for convenience, we assume that $\mathbf{z}$ has continuous distribution):

$$H(\mathbf{z}) = -\int_{-\infty}^{\infty} f(\mathbf{z}) \, \log f(\mathbf{z}) d\mathbf{z}.$$

Mutual information between the $z^{(i)}$, $i = 1, \ldots, p$, and $\mathbf{z}$ is:

$$I(z^{(1)}, \ldots, z^{(p)}) = \sum_{i=1}^{p} H(z^{(i)}) - H(\mathbf{z}).$$

It is zero if and only if the $z^{(i)}$ are mutually independent. It is also equal to the Kullback-Leibler divergence (or Kullback-Leibler distance) of $f(\mathbf{z})$ from

$$f_1(z^{(1)}) \; f_2(z^{(2)}) \cdots f_p(z^{(p)}).$$

PhD STUDIES KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Independent component analysis (ICA)

This last fact readily follows from the definition of the Kullback-Leibler divergence of density $g_1(\mathbf{v})$ from density $g_2(\mathbf{v})$ on R^p:

$$\delta(g_1, g_2) = \int_{-\infty}^{\infty} g_1(\mathbf{v}) \log \frac{g_1(\mathbf{v})}{g_2(\mathbf{v})} d\mathbf{v}.$$

It is zero if and only if the two densities are equal. Moroever,

$$\int_{-\infty}^{\infty} |g_1(\mathbf{v}) - g_2(\mathbf{v})| d\mathbf{v} \leqslant \sqrt{2\delta(g_1, g_2)}.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Let us return to (2). Orthogonality of matrix $\mathbf{\Delta}'$ implies that the mutual information between the $z^{(i)}$ satisfies the following equality

$$I(z^{(1)}, \ldots, z^{(p)}) = \sum_{i=1}^{p} H(z^{(i)}) - H(\mathbf{z}) = \sum_{i=1}^{p} H(z^{(i)}) - H(\mathbf{x}) - \log|\det \Delta'|. \tag{3}$$

Minimizing (3) w.r.t. $\mathbf{\Delta}'$ is equivalent to finding such a transformation of the original data that the new features $z^{(i)}$ are as close to mutual independence as possible. Note also that minimizing (3) amounts to minimizing the sum of entropies $H(z^{(i)})$, i.e. to maximizing the distance (when measured by entropy) between the distributions of the $z^{(i)}$, $i = 1, \ldots, p$ and a normal distribution (with the same covariance).

PhD STUDIES
KAPITAŁ LUDZKI NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

This last fact is of crucial importance: Most of the algorithms designed to perform ICA, i.e., to find matrix $\mathbf{\Delta}$, are based on maximizing the distance (however understood) between the latent variables $z^{(i)}$ and a normal distribution; e.g., classical algorithms seek the maximum absolute value of kurtosis of the $z^{(i)}$. Generally speaking, such algorithms always rest on the ideas from nonlinear programming, in particular gradient or Newton-like algorithms.

An interesting non-classical algorithm, in which the problem of maximizing (3) is directly addressed, has been proposed in [HTF].

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Dissimilarity and similarity measures

Dissimiliarity measure does not need to be a metric (triangle inequality does not need to be satisfied).

For vectors in R^p: any metric on R^p may be considered; for quantitative features on different scales weighted Euclidean distance is appropriate

$$d(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{p} w_i^2 (x_i - y_i)^2)^{1/2}$$

where w_i
is either (standard deviation of i^{th} variable)$^{-1}$ or (range)$^{-1}$.

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
The project is co-financed by the European Union within the framework of European Social Fund

Dissimilarity and similarity measures

For vectors with binary (0 and 1) coordinates: $x = (x_1, x_2, \ldots, x_p)$, $y = (y_1, y_2, \ldots, y_p)$ we define

$$a = \#\{x_i = 1 \,\&\, y_i = 1\}, \qquad b = \#\{x_i = 0 \,\&\, y_i = 1\};$$
$$c = \#\{x_i = 1 \,\&\, y_i = 0\}, \qquad d = \#\{x_i = 0 \,\&\, y_i = 0\}.$$

Dissimilarity measures for binary data:

- Normalized Hamming distance: $\frac{b+c}{a+b+c+d}$
- Jacquard: $\frac{b+c}{a+b+c}$
  (lack of occurence of a feature does not make objects more similar)
- Czekanowski: $1 - \frac{2a}{2a+b+c}$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

For qualitative vectors having more than two levels:

$$1 - \frac{\#\text{coordinates having the same value}}{\#\text{coordinates}}$$

Gower coefficients for mixed variables:

We assume that the coefficient is normalized, i.e., its values are from the $[0,1]$ interval, and we start with (partial) similarities s_{ijk} between the i-th and j-th element in the sample calculated coordinatewise for each k-th feature (coordinate), $k = 1, \dots, p$. The s_{ijk} are assumed to be normalized too and they are related to the corresponding (partial) dissimilarities d_{ijk} between the i-th and j-th element along the k-th feature (coordinate) by equation

$$s_{ijk} = 1 - d_{ijk}.$$

The project is co-financed by the European Union within the framework of European Social Fund

We allow that the comparison between a pair of elements along one or more coordinates is impossible. Accordingly, we define coefficient δ_{ijk} and, if the comparison between the i-th and j-th elements along the k-th coordinate is impossible, we set $\delta_{ijk} = 0$ (s_{ijk} is then unknown, but for reasons that will prove obvious we set $s_{ijk} = 0$); otherwise, $\delta_{ijk} = 1$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

We define

$$s_{ij} = \sum_{k=1}^{p} s_{ijk} \Big/ \sum_{k=1}^{p} \delta_{ijk}, \qquad d_{ij} = 1 - s_{ij}, \tag{4}$$

where

$$s_{ijk} = 1 - \frac{|x_i^{(k)} - x_j^{(k)}|}{\text{range of } k\text{-th variable}},$$

for quantitative variables,

$$s_{ijk} = \begin{cases} 1, & \text{if } x_i^{(k)} = x_j^{(k)} \\ 0, & \text{otherwise} \end{cases}$$

for qualitative variables, and

	k-th variable's value			
i-th observation	1	1	0	0
j-th observation	1	0	1	0
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

for binary variables.

Multidimensional scaling (MDS)

Let d_{ij}, $i, j = 1, \ldots, n$, be Euclidean distances between observations $\mathbf{x}_i$ and $\mathbf{x}_j$ in R^p. Let our task consist in finding a subspace R^r of a fixed dimension r, $r < p$, such that the distances $\hat{d}_{ij}$ between the projections of $\mathbf{x}_i$ and $\mathbf{x}_j$ on this subspace make the following sum minimal

$$V = \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij}^2 - \hat{d}_{ij}^2). \tag{5}$$

Interestingly, the R^r sought is given by the first r principal components for $\mathbf{x}_i$, $i = 1, \ldots, n$. Actually the task described is the task of the so-called multidimensional scaling in the particular case when distances are Euclidean. In general, the task of (metric) multidimensional scaling is the same, albeit for any given dissimilarity matrix.

Remark: Note that, in fact, in all generality we even do not have to know the $\mathbf{x}_i$, but only the dissimilarities between them.

Multidimensional scaling (MDS)

In whatever way we have acquired dissimilarity matrix $[d_{ij}]$, $i, j = 1, \ldots, n$, a question worth an answer is the following:

given dissimilarity matrix $[d_{ij}]$, is it possible to find R^s of some dimension s and a set of n points in this space such that Euclidean distances between these points, $\tilde{d}_{ij}$, $i, j = 1, \ldots, n$, form matrix $[d_{ij}]$, i.e., $\tilde{d}_{ij} = d_{ij}$, $i, j = 1, \ldots, n$?

If yes, given the space R^s with the given property, is it possible, for any natural number u, $u < s$, to find a set of n points in R^u such that Euclidean distances between these points, $\hat{d}_{ij}$, minimize V defined by (5)?

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Multidimensional scaling (MDS)

During the lecture, we shall briefly discuss these issues (not forgetting about a discussion on how to verify results obtained [e.g., by properly using a minimum spanning tree of the original data]).

We shall also mention the problem of nonmetric MDS.

The project is co-financed by the European Union within the framework of European Social Fund