Unsupervised learning: beyond simple clustering and PCA

Liza Rebrova

Self organizing maps (SOM)

Goal: approximate data points in $\mathbb{R}^p$ by a low-dimensional manifold

Unlike PCA, the manifold does not have to be a subspace

Method: constrained K-means clustering, with prototypes (centers of clusters) are encouraged to lie on low (1 or 2) dimensional manifold in feature space.

This manifold is also called constrained topological map, since the original high-dimensional observations can be mapped down *onto* the two-dimensional coordinate system.

SOM general construction

- Start with two-dimensional rectangular grid of K prototypes $m_j \in \mathbb{R}^p$ (it's possible to use other grids, e.g. hexagonal) Each prototype is parametrized by integer coordinate pair $y_j = (y_j^1, y_j^2) \in \{1, \ldots, q_1\} \times \{1, \ldots, q_2\}$, $K = q_1 \cdot q_2$.
- Initialize m_j. Good idea is to assume them along the first principal component (maximize mutual distance) - "buttons" on the principal component plane in a regular pattern.
- Algorithm tries to bend the plane so that "buttons" approximate the data points as well as possible.

SOM algorithm update step

Algorithm processes observations X_i one at a time.

- Find the closest prototype m_j, such that

$$m_j = \text{argmin}_j \|x_i - m_j\|_2$$

  (L_2 distance in $\mathbb{R}^p$)
- For all grid neighbors $m_k \sim m_j$ update

$$m_k := m_k + \alpha(x_i - m_k)$$

Definition

Prototype m_k is a neighbor of m_j, if

$$\|y_k - y_j\|_2 \leq r$$

(L_2 distance in $\{1, \dots, q_1\} \times \{1, \dots, q_2\} \subset \mathbb{R}^2$) Also, r is a chosen threshold, m_j is always a neighbor to itself and will be updated.

SOM parameters

Parameters of the algorithm:

- α is a learning rate (typically decreases from 1.0 to 0.0 over a few 1000's iterations, one per iteration)
- r is a distance threshold (also decreases linearly from R to 1 over same iterations)

K-means and SOM

- If we take r small enough to contain exactly one point, then the spatial connection between prototypes is lost, and we get standard K-means.
- In general, SOM is constrained version of K-means.
- To check whether constraint is reasonable, we can compare the reconstruction error

  $$\sum_i \|X_i - m_j i\|^2$$

  for K-means and for SOM.
  SOM-error is always bigger K-mean-error, but for the reasonable constraint they are compatible.

SOM variations

1. Variation of the algorithm with more sophisticated update step:
$$m_k := m_k + \alpha h(\|y_j - y_k\|) \cdot (x_i - m_k),$$
where $h(.)$ is a neighborhood function, which gives more weight to the prototypes m_k with the indices y_k closer to y_j than to those further away.
2. The original SOM algorithm is online (observations processed one at a time), but we can do a "batch" variation:
$$m_j := \frac{\sum w_k X_k}{\sum w_k},$$
where X_k are the observation points, coming from (mapped from) the neighbors m_k of m_j. Weight function w might be rectangular (1 on neighbors of m_k and zero otherwise) or decrease smoothly with $\|y_j - y_k\|$.

SOM examples

Multidimensional scaling (MDS)

Goal: approximate data points in $\mathbb{R}^p$ by a low-dimensional manifold

Same goal as in SOM and in PCA

Method: start with observations $X_1, \dots, X_N \in \mathbb{R}^p$, define a dissimilarity measure

$$d_{ij} := \|X_i - X_j\|.$$

Usually it is L_2 distance, but not necessarily.
Optimization task: MDS seeks values $z_1, \dots, z_N \in \mathbb{R}^k$ $(k \ll p)$ to minimize stress function

$$S_M(z_1, \dots, z_N) := \sum_{i \neq i'} (d_{ii'} - \|z_i - z_{i'}\|)^2 \to \text{min}$$

MDS stress functions

Many variations of stress functions:

1. Least squares scaling:

$$S_M(z_1, \ldots, z_N) := \sum_{i \neq i'} (d_{ii'} - \|z_i - z_{i'}\|)^2$$

   Idea: find a lower-dimensional representation of the data that preserves the pairwise distances as well as possible. (Note that approximation is in terms of distances, not squares of the distances - this makes computations harder).

2. Variation of least squares (Summons mapping):

$$S_{S_m}(z_1, \ldots, z_N) := \sum_{i \neq i'} \frac{(d_{ii'} - \|z_i - z_{i'}\|)^2}{d_{ii'}}$$

   Gives more importance on preserving smaller pairwise distances.

MDS stress functions

1. Least squares scaling
2. Summons mapping
3. Classical scaling:

$$S_C(z_1, \ldots, z_N) := \sum_{i,i'} (s_{ii'} - \langle z_i - \bar{z}, z_{i'} - \bar{z} \rangle)^2$$

   Here $s_{ii'}$ are similarities between the data. Frequently,

$$s_{ii'} = \langle X_i - \bar{X}, X_{i'} - \bar{X} \rangle,$$

   then this is equivalent to principal components method.
4. Shephard-Kruskal nonmetrc scaling
   $S_{NM}(z_1, \ldots, z_N)$ uses only ranks

MDS optimization

- Usually, S_M is minimized by gradient descent
- In case of classical scaling (S_C) we can write an explicit solution in terms of eigenvectors

MDS and SOM

Advantages of SOM:

- manifold approximation is more flexible than subspace approximation
- provides a low-dimensional coordinate system for data

Advantages of MDS:

- various dissimilarity/similarity metrics can be used
- preserves distances (in case of SOM close points are kept close, but the points farther apart can also become close)

Independent component analysis (ICA)

What if our data comes as multiple indirect measurements from some underlying source, but the source itself cannot be directly measured?

Some examples:

- Sound recording from the noisy room, we want to separate music from people or two people speaking
- Educational and phycological test are supposed to use answers to questions to measure the underlying intelligence and other mental abilities of subjects
- EEG brain scans measure the neuronal activity in various parts of the brain indirectly via electromagnetic signals recorded at sensors located at various positions on the head

Goal: find these latent sources (components) producing data. (Note that it is different from PCA/SOM/MSD goals - we do not search for low-dim data approximation)

ICA problem formal statement

The model is

$$\bar{x} = A \cdot \bar{s},$$

where

- $\bar{x} \in \mathbb{R}^p$ - one p-dimensional observation (think about a vector with dependent coordinates, taken from some underlying probability space)
- $\bar{s} \in \mathbb{R}^p$ - a latent source p-vector, whose components are independently distributed random variables (on the same underlying probability space)
- A - $p \times p$ mixing matrix

Goals of ICA: given N observations (realizations $x_1, \ldots, x_N \in \mathbb{R}^p$),

- estimate A
- estimate the source distributions f_{s_j} (densities of s_j, $j \in [p]$).

ICA in terms of matrices

Equivalently, the model can be rewritten as

$$\bar{x} = \sum_{i=1}^{p} A_i s_i,$$

where A_i are the columns of the mixing matrix.

Also,

$$X = A \cdot S,$$

where

- X is $p \times N$ observation matrix (every observation is a column)
- S is $p \times N$ source matrix with independent rows
- A is $p \times p$ mixing matrix

ICA vs PCA: independent vs uncorrelated

From SVD (singular value decomposition) we can find such decomposiotion:

$$X^T = U\Sigma V^T = \sqrt{n}U \cdot \frac{1}{\sqrt{n}}\Sigma V^T =: S^T \cdot A^T$$
$$X = AS$$

Every observation x_i is a linear combination of latent variables s_i, which are uncorrelated (as S was orthogonal), mean 0 (assume X is centered), variance 1 (rescaling). Does this define mixing matrix and latent variables well?

No. Problem: for any orthogonal $p \times p$ matrix R

$$X = AS = AR^T RS = A^* S^*,$$

and S^* has the same properties, as S (mean 0, variance 1, no correlation).

This is why we require independence, not just zero correlation.

Ambiguities of ICA

Usually, both A and S are assumed unknown. Hence, it is impossible to determine them uniquely. In particular, we cannot determine

1. the variances of independent components s_j.
   Rescaling $A \to \alpha A$, $s \to s/\alpha$ does not change the result.
   Common assumption: $\mathbb{E}s_j^2 = 1$ (and $\mathbb{E}s_j = 0$, this follows if we centralize x)
2. the order of the independent components s_j
   For any permutation matrix P we have

$$x = A \cdot Id \cdot x = AP^{-1}Ps,$$

3. if distribution of s is rotationally invariant, we have a problem.
   Then matrix A is not identifiable, since for any orthonormal R

$$x = AR^T Rs = (AR^T)s.$$

Measure of "non-gaussianity"

Rotationally invariant = Gaussian. Also, recall, that for a gaussian random variables zero correlation is equivalent to independence. Hence, the assumption needed for ICA: underlying sources are NOT gaussian

Very informal explanation: sum of independent components (independent identically scaled random variables) tends to normal distribution by Central Limit Theorem, so any single s_i is "farther" from gaussian than any linear combination of s_i's (weights should satisfy condition on their size, this is very informal)

Method:

- measure "distance to gaussian distribution" in terms of entropy (next slide)
- $\bar{w}^T \bar{x} \to \max_w$ in sence of this measure ($2p$ local maxima in p-dim space, corresponding to $s_1, -s_1, s_2, -s_2, \ldots$)

ICA: entropy and negentropy

Definition

Entropy of a random variable Y with density $f(y)$ is

$$H(Y) = -\int f(y) \log f(y) dy$$

Entropy is maximized by Gaussian density $f(y)$.

Definition (Hyvarinen, Oja, 2000)

Negentropy measure is

$$J(Y_j) = H(Y_j) - H(Z_j),$$

where Z_j is a Gaussian random variable with same variance as Y_j.

It measures the departure from Gaussianity $\therefore$ ICA seeks to maximize negentropy.

ICA: mutual information

The notion of negentropy came from similarity to the mutual information, that measures the departure from independence.

Mutual Information is

$$I(Y) = \sum_{j=1}^{p} H(Y_j) - H(Y),$$

where

- Y is a random vector with components Y_j
- $I(Y)$ is also called Kullback-Leibler divergence between density $f_Y(.)$ and its independence version $\prod_1^j f_{Y_j}(.)$ (which is K-L closest of all independence densities to $f_Y(.)$)
- Hence $I(Y)$ is a measure of dependence between the components of a random vector Y.

Another approach to ICA: directly maximize mutual information (max likelihood principle)

ICA preprocessing

- Centering $\mathbb{E}X = 0$
- Whitening (unlike PCA!) $\mathbb{E}XX^T = \mathrm{Id}$

EXAMPLES

Literature

These slides are made as a complement to the lecture material slides:

- CME 250 Stanford course by Alexander Ioannidis and Karianne Bergen: https://sites.google.com/site/cme250winter2016/lecture-materials

Additional sources used (by topic, inside topic in order of helpfulness for me).

ICA:

1. A. Hyvarinen, E. Oja ICA: Algorithms and Applications (link in the course website near Lecture 3)
2. T. Hastie ICA by Product Density Estimation slides https://web.stanford.edu/ hastie/Papers/icatalk.pdf
3. T. Hastie et al The elements of statistical learning pp 557 - 570