Processing of missing data by neural networks

Marek Śmieja, Łukasz Struski, Jacek Tabor, Bartosz Zieliński, Przemysław Spurek
Jagiellonian University, Kraków, Poland

Problem

Learning from incomplete data is one of the fundamental challenges in machine learning as it appears naturally in many practical problems:

- in object detection, the system has to recognize partially hidden faces,
- in industry, some sensors may be unreliable or broken,
- in medical diagnosis, a doctor may be unable to complete the patient examination.

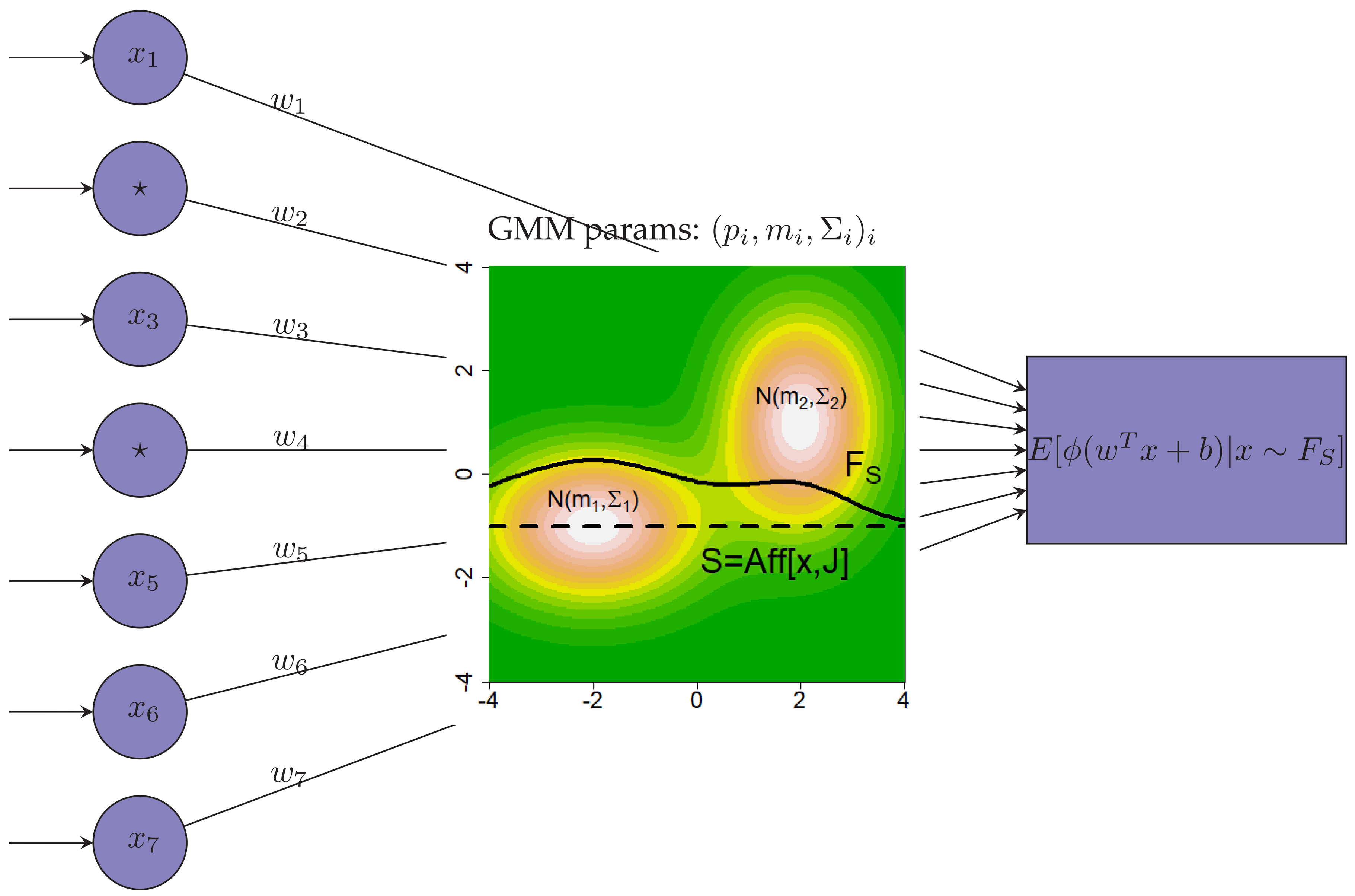
Due to the great interest in deep learning in the last decade, it is especially important to establish unified tools for practitioners to process missing data with arbitrary neural networks.

Main results

We introduce a general, theoretically justified methodology for feeding neural networks with missing data:

- The main advantage of the proposed approach is the ability to train neural network on data sets, which contain only incomplete samples (without a single fully observable data).
- It can be applied to various types of neural networks and requires only minimal modification in their architectures (i.e. modification of the first layer).
- Our main theoretical result shows that this generalization does not loss the information when processing the input.

Proposed model



Construction process:

- Assuming that the values at missing attributes come from a parametric probability density function F_Θ, we identify every missing data point (x, J), where $J \subset \{1, \dots, D\}$ denotes unobserved coordinates, by restricting F_Θ to the affine subspace $S = \mathrm{Aff}[x, J]$:

$$F_S(x) = \begin{cases} \frac{1}{\int_S F_\Theta(s)ds} F_\Theta(x), \text{ for } x \in S, \\ 0, \text{ otherwise.} \end{cases}$$

- Parameters Θ of density model F_Θ are trained together with remaining network weights.
- To process F_S by neural network, we compute the expected value of the neuron's response at the first hidden layer. Such generalized response (activation) of a neuron $\phi : \mathbb{R}^D \to \mathbb{R}$ on F_S is defined as the mean output:

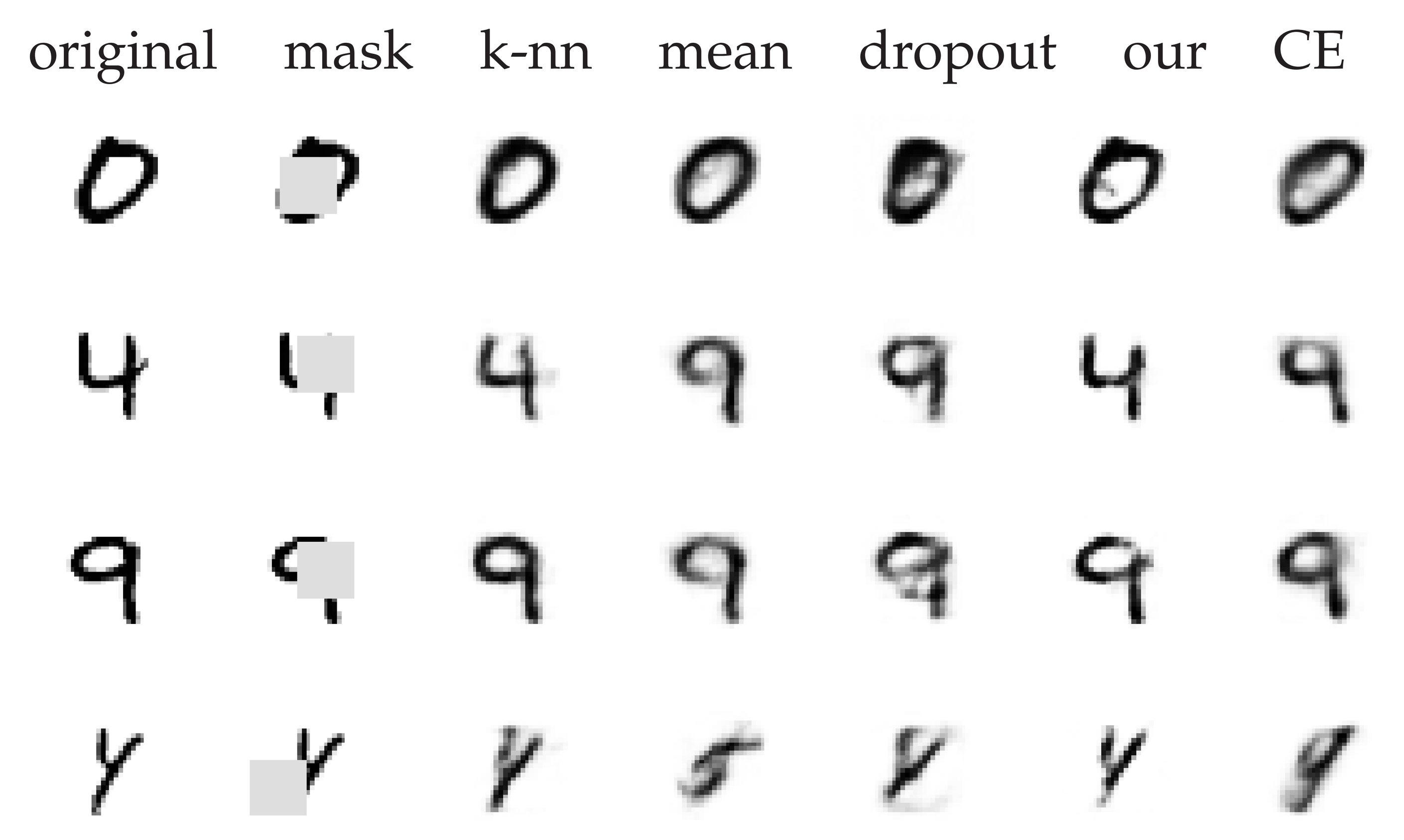
$$\phi(F_S) = E[\phi(x)|x \sim F_S] = \int \phi(x) F_S(x) dx.$$

  It can be interpreted as calculating the average neuron's activation over the imputations drawn from missing data density F_S.
- Formulas for generalized response of a neuron can be calculated analytically for typical activation functions such as ReLU and RBF when F_Θ is estimated with the mixture of Gaussians.

Experiments

Reconstructions of partially incomplete images using dense autoencoder

From left: (1) original image, (2) image with missing pixels passed to autoencoder; the output produced by autoencoder when unknown pixels were initially filled by (3) k-nn imputation and (4) mean imputation; (5) the results obtained by autoencoder with dropout, (6) our method and (7) context encoder.



All columns except the last one were obtained with MSE loss function limited to pixels outside the mask (i.e. without fully observable data in training phase). It can be noticed that our method gives much sharper images than baseline methods.

	only missing data				complete data
MSE error	**k-nn**	**mean**	**dropout**	**our**	**CE**
Total error	0.01189	0.01727	0.01379	**0.01056**	0.01326
Error inside the mask	0.00722	0.00898	0.00882	0.00810	**0.00710**
Error outside the mask	0.00468	0.00829	0.00498	**0.00246**	0.00617

Binary classification using shallow RBF network on UCI data sets with internally missing attributes

	only missing data in training								complete data
data	**karma**	**geom**	**k-nn**	**mice**	**mean**	**gmm**	**dropout**	**our**	**CE**
bands	0.580	0.571	0.520	0.544	0.545	0.577	**0.616**	0.598	0.621
kidney	**0.995**	0.986	0.992	0.992	0.985	0.980	0.983	0.993	0.996
hepatitis	0.665	0.817	0.825	0.792	0.825	0.820	0.780	**0.846**	0.843
horse	0.826	0.822	0.807	0.820	0.793	0.818	0.823	**0.864**	0.858
mammogr.	0.773	0.815	0.822	0.825	0.819	0.803	0.814	**0.831**	0.822
pima	0.768	0.766	0.767	**0.769**	0.760	0.742	0.754	0.747	0.743
winconsin	0.958	0.958	0.967	**0.970**	0.965	0.957	0.964	**0.970**	0.968

Web: ww2.ii.uj.edu.pl/~smieja
Email: marek.smieja@uj.edu.pl

Code: `https://github.com/lstruski/Processing-of-missing-data-by-neural-networks.`