Processing of missing data by neural networks

Reproduced by: Kai Lin, Lulu Meng, Yiwen Su, Zhanhong Tian

Dec. 05 2019

Śmieja, M., Struski, Ł., Tabor, J., Zieliński, B. and Spurek, P., 2018. Processing of missing data by neural networks. In Advances in Neural Information Processing Systems (pp. 2719-2729).

Claim / Target Task

- In this project, the authors want to introduce a general, theoretically justified methodology for feeding neural networks with missing data.

- The authors also want to model the uncertainty on missing attributes by probability density functions.

An Intuitive Figure Showing WHY Claim

- How to process missing data?

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel fi

B5 | *fx* | HORNBY Coach R4410A BR Hawksworth Corridor 3rd

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	uniq_id	product_nam	manufacture	price	number_ava	number_of_	number_of_	average_revi	amazon_cate	customers_v	description	product_info	product_desc	items
2	eac7efa5dbd	Hornby 2014	Hornby	¬£3.42	5¬†new	15	1	4.9 out of 5 s	Hobbies > M	http://www.	Product Desc	Technical De	Product Desc	http:/
3	b17540ef7e8	FunkyBuys¬A	FunkyBuys	¬£16.99		2	1			http://www.	Size Name:L	Technical De	Size Name:L	http:/
4	348f344247b	CLASSIC TOY	ccf	¬£9.99	2¬†new	17	2	3.9 out of 5 s	Hobbies > M	http://www.	BIG CLASSIC	Technical De	BIG CLASSIC	http:/
5	e12b92dbb8	HORNBY Coa	Hornby	¬£39.99		1	2				Hornby 00 Ga	Technical De	Hornby 00 Gauge E	

Motivation

- Collecting data is an important process for supervised learning. But sometimes data collected may have some missing fields. Discarding these data is a waste.
- Due to the great interest in deep learning in recent years, it is important to establish unified tools for practitioners to process missing data with arbitrary neural networks.
- This paper tries to figure out whether we can use data with missing value to train our model, how well the filling method performs, and whether it can be easily applied.

Background

- Neural Network
- Expectation–Maximization Algorithm
- GMM

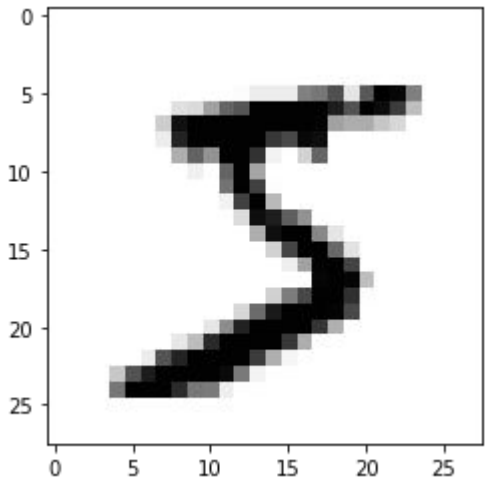
Related Work

Generating candidates for filling missing attributes

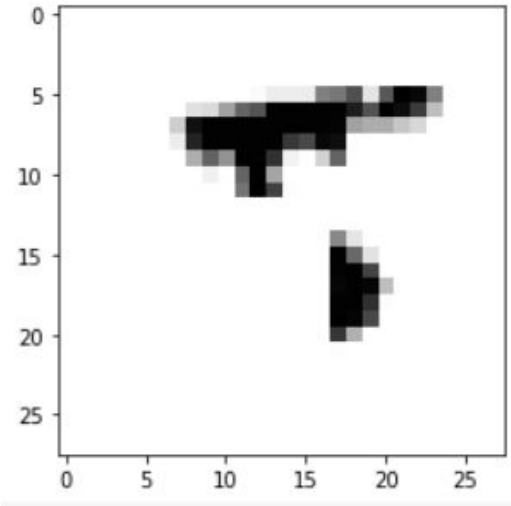
- Use mean/mode to fill the data
- Train another model to predict the data(Neural Network/ Extreme Learning Machine/ K-Nearest Neighbors)
- Generative Adversarial Net(GAN)

Building Probabilistic model of incomplete data(Make assumption about the data)

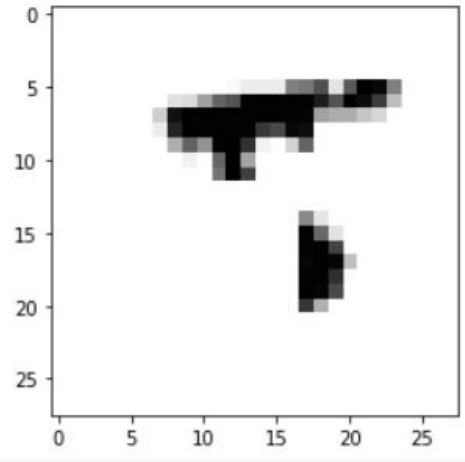
e.g. MNIST Dataset

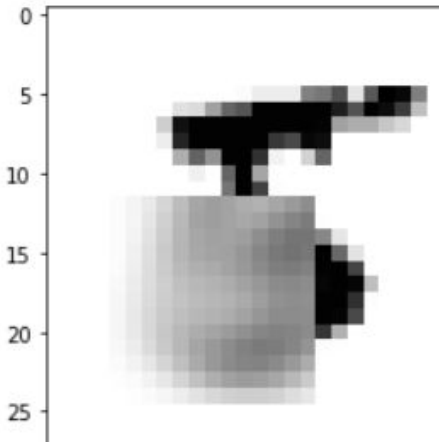


Adding Mask Randomly

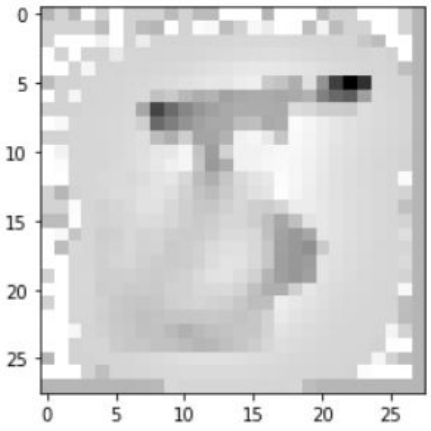


Simple Solution

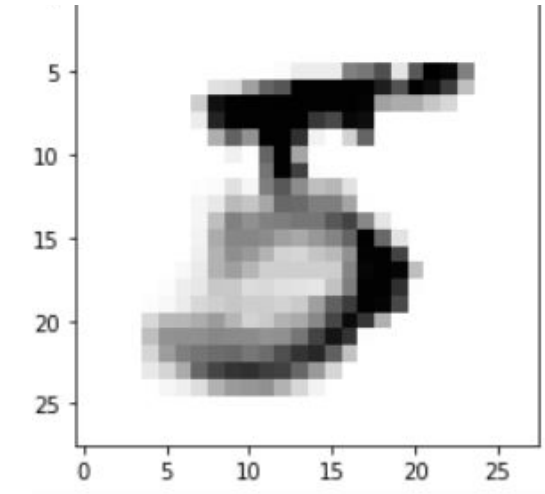




Mean



SoftImpute



KNN

- Neural Network
- Extreme learning machines
- ……

Proposed Approach

Paper's solution

Formula Derivative

Missing data representation:

Data point: (x,J)

S = Aff[x, J] = x + span(e J)

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

Generalized neuron's response:

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

NR (Auxiliary function) Derivation

$$\mathrm{NR}(w) = \mathrm{ReLU}[N(w, 1)],$$

$$\mathrm{ReLU}[N(m, \sigma^2)] = \sigma \mathrm{NR}(\frac{m}{\sigma}).$$

$$\mathrm{NR}(w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{w^2}{2}) + \frac{w}{2}(1 + \mathrm{erf}(\frac{w}{\sqrt{2}})),$$

$$\mathrm{erf}(z) = \frac{2}{\sqrt{pi}} \int_0^z \exp(-t^2) dt.$$

Key Formula for the program:

Mixture of nondegenerate gaussians in affine space

$$F_S^\gamma = \sum_i r_i N(m_S^i, \Sigma_S^i),$$

$$m_S^i = [x_{J'}, (m_i)_J], \Sigma_S^i = [0_{J'J'}, (\Sigma_i)_{JJ}],$$

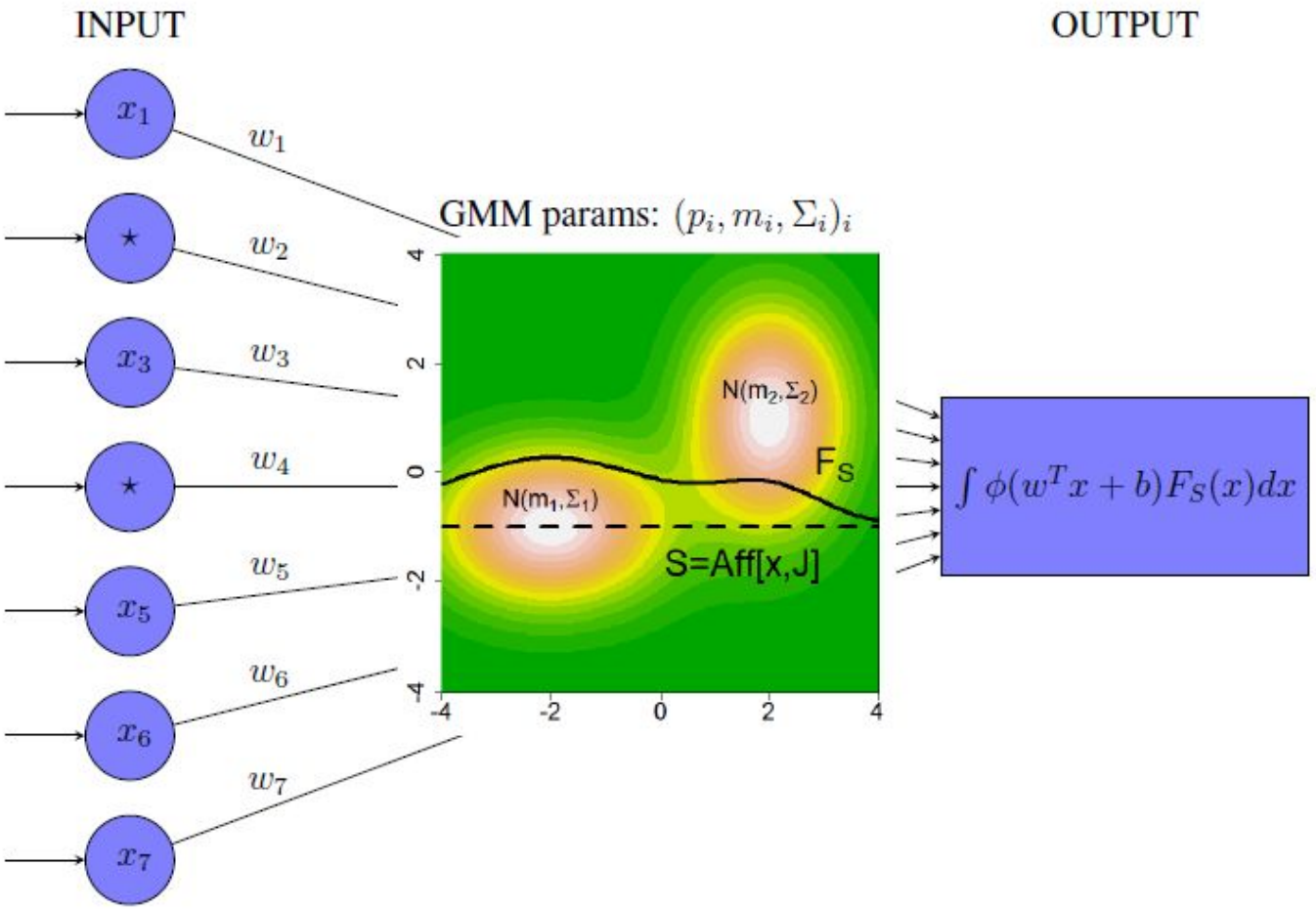
$$r_i = \frac{q_i}{\sum_j q_j}, q_i = C^\gamma_{m_i,\Sigma_i,S} \cdot p_i,$$

$$C^\gamma_{m,\Sigma,S} = \frac{1}{(2\pi)^{(D-|J|)/2} \prod_{l\in J'}(\gamma+\sigma_l)^{1/2}} \cdot \exp(-\tfrac{1}{2}\sum_{l\in J'} \tfrac{1}{\gamma+\sigma_l}(m_l - x_l^2).$$

Proposed Solution

Missing data representation: EM algorithm to estimate incomplete data density with the use of Gaussian Mixture Model. Then based on the data density, we can infer what the incomplete data should be. And the parameter of GMM is learnt by using neural network.

Generalized neuron's response: With changing of the first layer of the network structure, the method can be easily adapted to existing network structure to deal with the problem of incomplete data.


INPUT
OUTPUT
x_1
x_3
x_5
x_6
x_7
w_1
w_2
w_3
w_4
w_5
w_6
w_7
GMM params: $(p_i, m_i, \Sigma_i)_i$
N(m2,Σ2)
N(m1,Σ1)
FS
S=Aff[x,J]
4
2
0
-2
-4
-4
-2
0
2
4
$\int \phi(w^T x + b) F_S(x) dx$

The Output Parameters of GMM

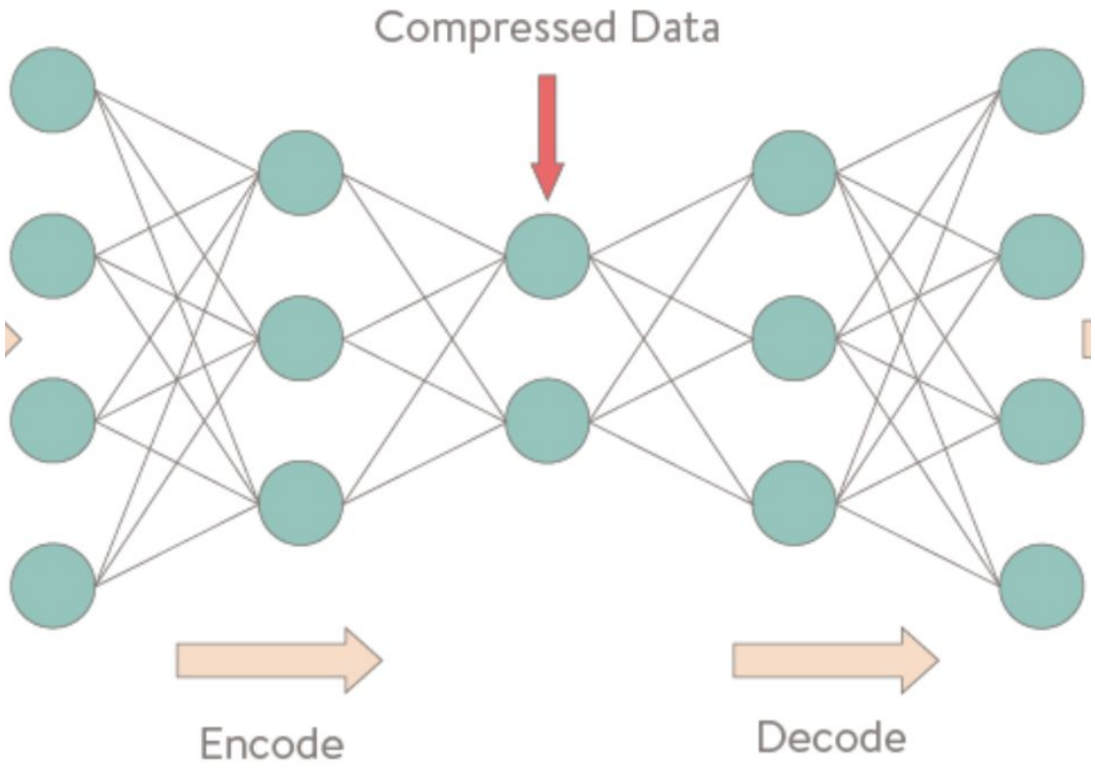
We get these parameters from GMM as input to the convolutional layer:

1. The weights of each mixture components.
2. The mean of each mixture component.
3. The covariance of each mixture component.
4. A tensor of the specified shape filled with random normal values.

These parameters will be updated within the network structure(encoder-decoder network) to better model the data.

Network architecture

The author proposed to use encoder and decoder to solve the missing data of MNIST dataset.



Loss Function and Optimization Method

Loss Function: **mean-square error**

The loss inside the mask, outside the mask and total area can all be calculated. Since the input of model has no complete data, the loss function only based on outside the mask area.

Optimization Method: **RMSProp**

Implementation

Adaptation of a given neural network to incomplete data relies on the 2 steps:

- **Estimation of missing data density with the use of mixture of diagonal Gaussians.**

  If data satisfies missing at random assumption, EM algorithm should be used to estimate incomplete data density with the use of GMM.

  general case:

  The network should learn optimal parameters of GMM with respect to its cost function.

Implementation

Adaptation of a given neural network to incomplete data relies on the 2 steps:

- **Generalization of neuron's response.**

  Generalizing the activation functions of all neurons in the first hidden layer of the network to process probability measures. The modification only presents on the first hidden layer.

The paper performs several implementations:

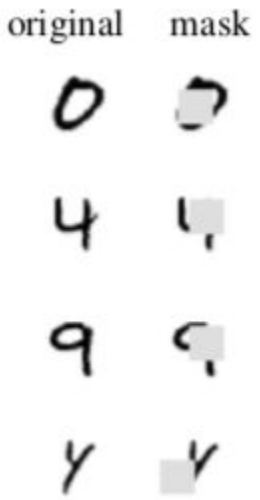
- Reconstruction of incomplete MNIST images
- Running RBFN on 8 examples retrieved from UCI repository

Data Summary

2 kinds of data set:

1. MNIST with a removed square patch on each image(Uniformly sample the location)

2. UCI repo two-class data sets with internally missing attributes



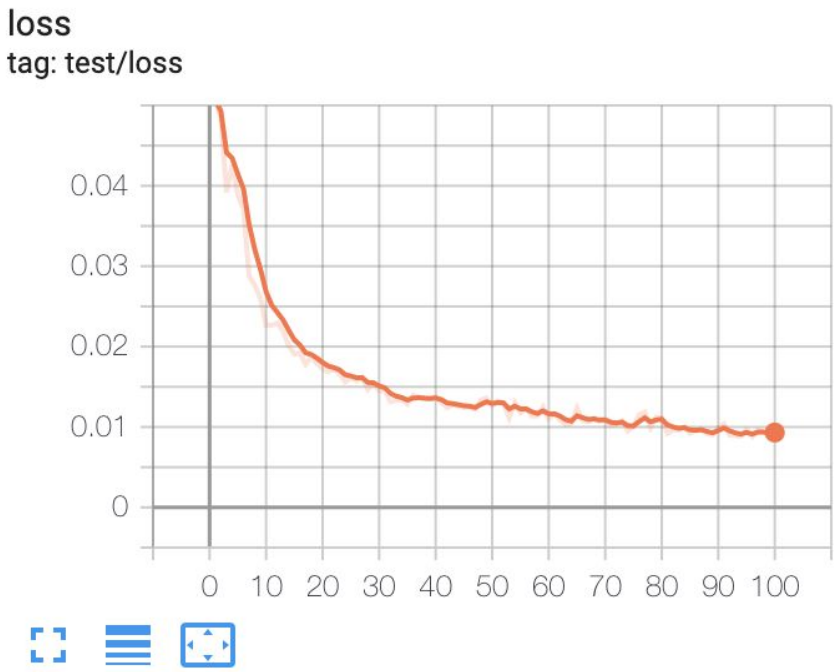
Experimental Results & Analysis

Evaluated in 2 types of architectures:

1. Autoencoder(AE)

Structure: 5 hidden layers with 256, 128, 64, 128, 256 neurons (first use ReLU, rests use sigmoids) in encoder.

Results: After running for 100 epochs, it gives sharper images and lower error.



Experimental Results & Analysis

Evaluated in 2 types of architectures:

2. Radial basis function network(RBFN)

Structure: 1 hidden layer and softmax in the output layer applied with cross-entropy function

	only missing data								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

Our experiment's result:

(training epochs 50; learning rate: 0.25; batch size: 75; hidden layer: 25; distribution: 3)

bands: Train: 0.6089 test: 0.5056

kidney: Train: 0.8539 test: 0.8625

hepatitis: Train: 0.7135 test: 0.6978

horse: Train: 0.7291 test: 0.7555

Conclusion and Future Work

- The paper presented a general approach for adapting neural networks to process incomplete data.
- The paper's approach can be used for a wide range of networks.
- The paper gives comparable results to the other methods, some of which even require complete data in training.

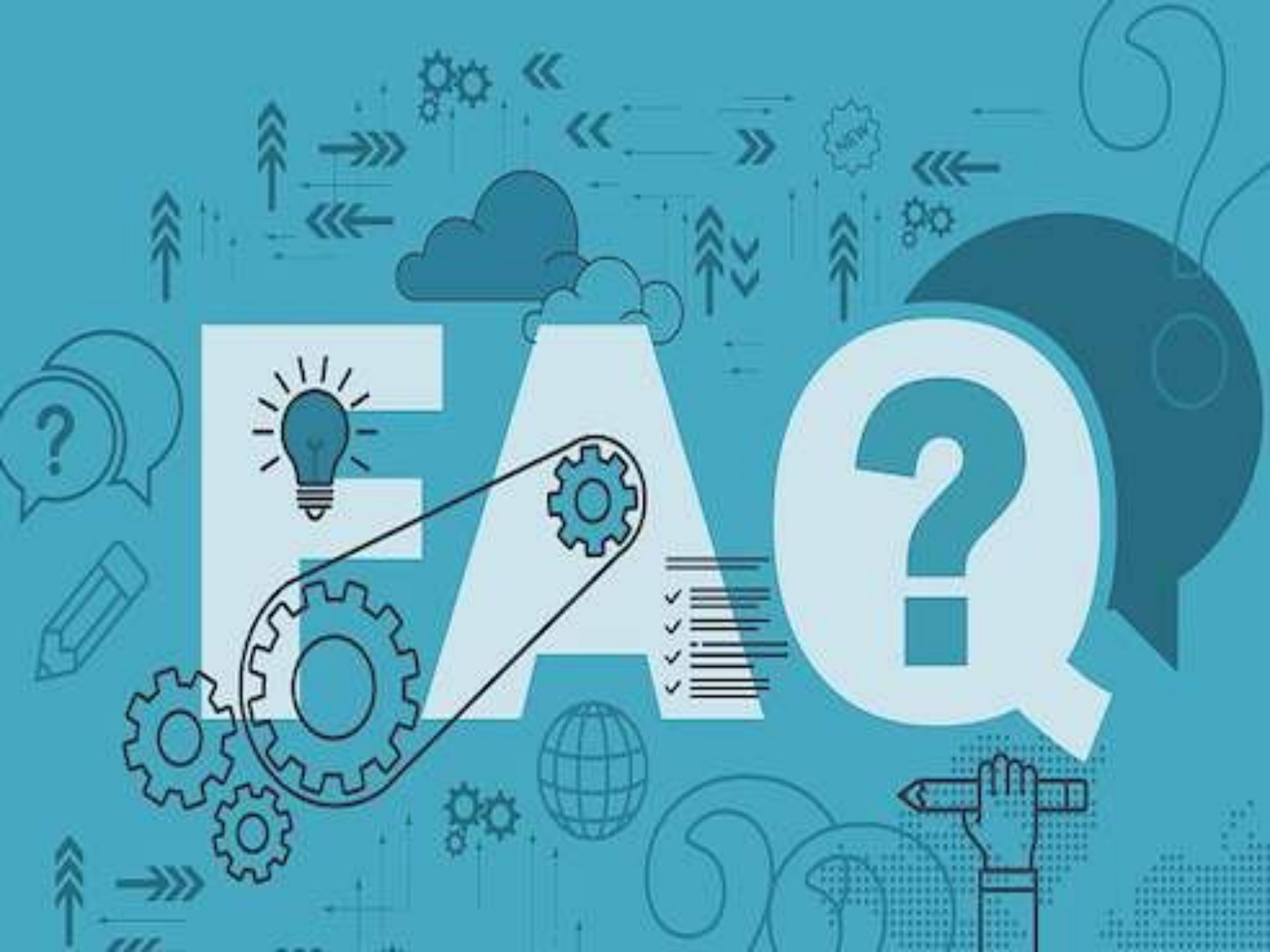
Task Allocation

Yiwen Su: Revise the presentation slides, analyze the source code structure and reproduce the result of the paper(AE). Try different methods to fill in the blank in the data, including KNN, softmulte and mean.

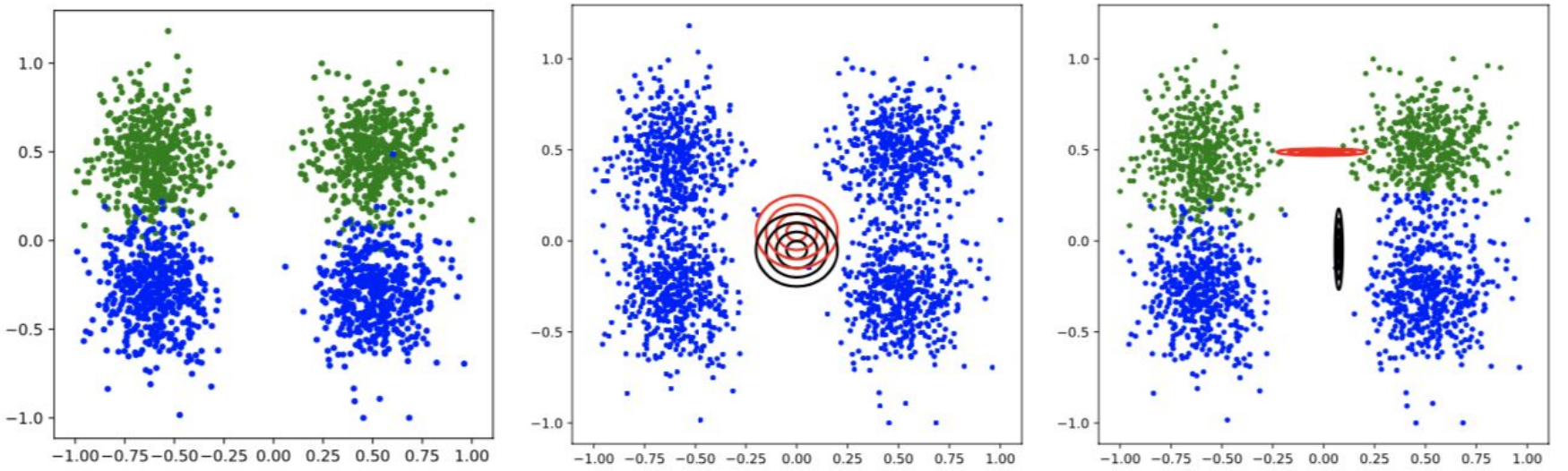
Zhanhong Tian: Revise the presentation slides, reproduce some of the result of the paper. Finding out one wrong description of formula in the paper. Visualize the data in MNIST.

Lulu Meng: Revise the presentation slides, reproduce RBFN model binary classification experiment of the paper.

Kai Lin: Revise the presentation slides, derive the formula in the paper.


FAQ

Why GMM parameters are tuned together within the network



(a) Reference classification (b) Initial Gaussians (c) Final Gaussians and resulted classification

After training with neural network, we can get a GMM, where its first component estimates a density of class 1, while the second component matches class 2, which means it can help perform better classification then estimating GMM directly by EM algorithm.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[2] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2536–2544, 2016.

[3] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, page 3, 2017.

[4] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In Advances in neural information processing systems, pages 341–349, 2012.

[5] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5485–5493, 2017.

[6] Patrick E McKnight, Katherine M McKnight, Souraya Sidani, and Aurelio Jose Figueredo. Missing data: A gentle introduction. Guilford Press, 2007.

[7] Peter K Sharpe and RJ Solly. Dealing with missing values in neural network-based diagnostic systems. Neural Computing & Applications, 3(2):73–77, 1995.

References

[8] Dušan Sovilj, Emil Eirola, Yoan Miche, Kaj-Mikael Björk, Rui Nian, Anton Akusok, and Amaury Lendasse. Extreme learning machine for missing data using multiple imputations. Neurocomputing, 174:220–231, 2016.

[9] Gustavo EAPA Batista, Maria Carolina Monard, et al. A study of k-nearest neighbour as an imputation method. HIS, 87(251-260):48, 2002.

[10] Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. Journal of statistical software, 45(3), 2011.

[11] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputa- tion by chained equations: what is it and how does it work? International journal of methods in psychiatric research, 20(1):40–49, 2011.

[12] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. pages 5689–5698, 2018.

[13] Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck. Monotonic calibrated interpolated look-up tables. The Journal of Machine Learning Research, 17(1):3790–3836, 2016.

[14] Zoubin Ghahramani and Michael I Jordan. Supervised learning from incomplete data via an EM approach. In Advances in Neural Information Processing Systems, pages 120–127. Citeseer, 1994.

References

[15] Volker Tresp, Subutai Ahmad, and Ralph Neuneier. Training neural networks with deficient data. In Advances in neural information processing systems, pages 128–135, 1994.

[16]MarekS ´mieja,ŁukaszStruski,andJacekTabor.Generalizedrbfkernelforincompletedata. arXiv preprint arXiv:1612.01480, 2016.

[17] David Williams, Xuejun Liao, Ya Xue, and Lawrence Carin. Incomplete-data classification using logistic regression. In Proceedings of the International Conference on Machine Learning, pages 972–979. ACM, 2005.

[18] Alexander J Smola, SVN Vishwanathan, and Thomas Hofmann. Kernel methods for missing variables. In Proceedings of the International Conference on Artificial Intelligence and Statistics. Citeseer, 2005.

[19] David Williams and Lawrence Carin. Analytical kernel matrix completion with incomplete multi-view data. In Proceedings of the ICML Workshop on Learning With Multiple Views, 2005.

[20] Pannagadatta K Shivaswamy, Chiranjib Bhattacharyya, and Alexander J Smola. Second order cone programming approaches for handling missing and uncertain data. Journal of Machine Learning Research, 7:1283–1314, 2006.

[21] Diego PP Mesquita, João PP Gomes, and Leonardo R Rodrigues. Extreme learning machines for datasets with missing values using the unscented transform. In Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on, pages 85–90. IEEE, 2016.

References

[22]XuejunLiao,HuiLi,andLawrenceCarin.Quadraticallygatedmixtureofexpertsforincomplete data classification. In Proceedings of the International Conference on Machine Learning, pages 553–560. ACM, 2007.

[23] Uwe Dick, Peter Haider, and Tobias Scheffer. Learning from incomplete data with infinite imputations. In Proceedings of the International Conference on Machine Learning, pages 232–239. ACM, 2008.

[24] Ofer Dekel, Ohad Shamir, and Lin Xiao. Learning to classify with missing and corrupted features. Machine Learning, 81(2):149–178, 2010.

[25] Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In Proceedings of the International Conference on Machine Learning, pages 353–360. ACM, 2006.

[26] Gal Chechik, Geremy Heitz, Gal Elidan, Pieter Abbeel, and Daphne Koller. Max-margin classification of data with absent features. Journal of Machine Learning Research, 9:1–21, 2008.

[27] JingXia,ShengyuZhang,GuolongCai,LiLi,QingPan,JingYan,andGangminNing.Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recognition, 69:52–60, 2017.

[28] Kristiaan Pelckmans, Jos De Brabanter, Johan AK Suykens, and Bart De Moor. Handling missing values in support vector machine classifiers. Neural Networks, 18(5):684–692, 2005.

References

[29] Elad Hazan, Roi Livni, and Yishay Mansour. Classification with low rank and missing data. In Proceedings of The 32nd International Conference on Machine Learning, pages 257–266, 2015.

[30] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. Transduction with matrix completion: Three birds with one stone. In Advances in neural information processing systems, pages 757–765, 2010.

[31] Yoshua Bengio and Francois Gingras. Recurrent neural networks for missing or asynchronous data. In Advances in neural information processing systems, pages 395–401, 1996.

[32] Robert K Nowicki, Rafal Scherer, and Leszek Rutkowski. Novel rough neural network for classification with missing data. In Methods and Models in Automation and Robotics (MMAR), 2016 21st International Conference on, pages 820–825. IEEE, 2016.

[33] Ian Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Multi-prediction deep boltzmann machines. In Advances in Neural Information Processing Systems, pages 548–556, 2013.

[34] Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. Linear statistical inference and its applications, volume 2. Wiley New York, 1973.

References

[35] Ralph G Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E Elger. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E, 64(6):061907, 2001.

[36] Arthur Asuncion and David J. Newman. UCI Machine Learning Repository, 2007.