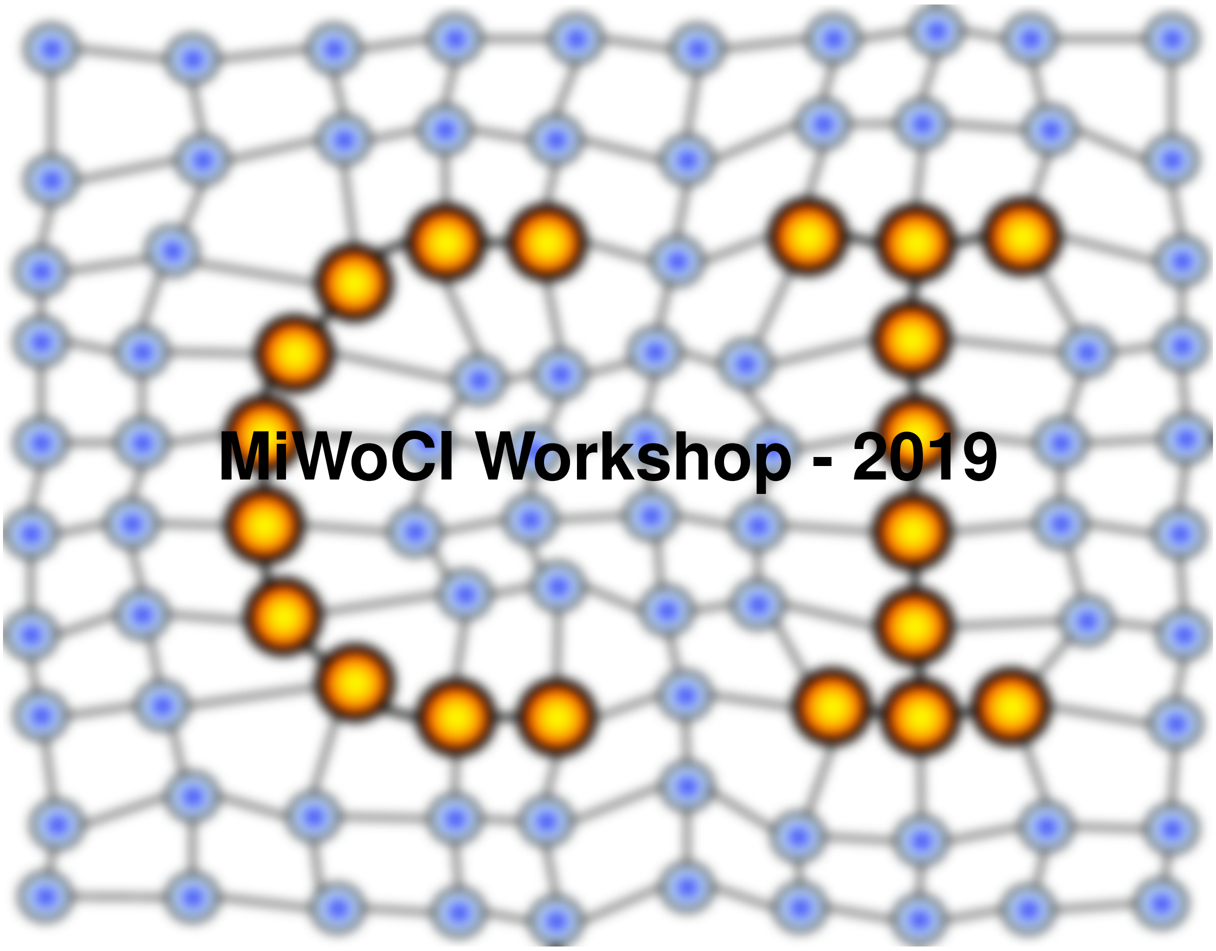
MACHINE LEARNING REPORTS



MiWoCI Workshop - 2019

Report 02/2019
Submitted: 01.11.2019
Published: 23.11.2019

Frank-Michael Schleif[1,2*,3*], Thomas Villmann[2] (Eds.)
(1) University of Applied Sciences Wuerzburg-Schweifurt, Sanderheinrichsleitenweg 20, 97074 Wuerzburg, Germany (2) University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany (3) University of Birmingham, School of Computer Science, Edgbaston, B15 2TT Birmingham, UK

Impressum

Publisher:	University of Applied Sciences Mittweida Technikumplatz 17, 09648 Mittweida, Germany
Editor:	Prof. Dr. Thomas Villmann Dr. Frank-Michael Schleif
Techical-Editor:	Dr. Frank-Michael Schleif
Contact:	fschleif@techfak.uni-bielefeld.de
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

Abstracts of the 11^{th} Mittweida Workshop on Computational Intelligence - MiWoCI 2019 -

Frank-Michael Schleif, Marika Kaden, and Thomas Villmann

Machine Learning Report 02/2019

Preface

The 11 th international *Mittweida Workshop on Computational Intelligence* (MiWoCI) gathering together 35 scientists from different universities including Bielefeld, Groningen, UAS Mittweida, UAS Würzburg-Schweinfurt, UAS Zwickau, research facilities including Porsche AG in Weissach, and IFF Fraunhofer in Magdeburg. The workshop took place in Mittweida, Germany, from 11.9. - 13.9.2019 and continued the tradition of scientific presentations, vivid discussions, and exchange of novel ideas at the cutting edge of research connected to diverse topics in computer science, automotive industry, and machine learning.

This report is a collection of abstracts and short contributions about the given presentations and discussions, which cover theoretical aspects, applications, as well as strategic developments in the fields.

Contents

Reject Options for Particle Filter in Medical Tracking Applications

Johannes Kummert

University of Bielefeld, CITEC, Germany

Abstract

A new approach in jaw reconstruction surgeries utilizes bone cut from the patients pelvis to create a dental implant that is readily accepted by the body [2]. For the cutting procedure a custom 3d printed model is created to serve as a template for the surgeon. In our project, a robotic arm is outfitted with a depth camera and a projector to track a representation of the model in the depth image using a particle filter [1] and project cutting lines onto the body. Detection of whether the tracking is still accurate would enable a safer procedure for example by automatically disabling the cutting tool. In this work we study how to measure certainty for particle filter and explore how this is dependent on the parameterization and choice of prediction model.

References

[1] Christian Gentner, Siwei Zhang, and Thomas Jost. "Log-PF: Particle Filtering in Logarithm Domain". en. In: Journal of Electrical and Computer Engineering 2018 (2018), pp. 1-11. issn: 2090-0147, 2090-0155.

[2] Ali Modabber et al. "Computer-assisted zygoma reconstruction with vascularized iliac crest bone graft". In: The International Journal of Medical Robotics and Computer Assisted Surgery 9.4 (2013), pp. 497-502.

Developing a latent space for hyperspectral camera standardization

Patrick Menz* and Udo Seiffert

Biosystems Engineering, Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany

Abstract

The hyperspectral camera standardization is a desirable goal to achieve nearly the same classification performance from a already generated machine learning model under changing sensor hardware. An obvious idea would be an interpolation of the spectra to the wavelength of a new sensor, but it has been shown in [1], that this leads to inappropriate results, but there are still other ways to get such a standardization. One could be a method of transfer learning as already shown in [1], in cost of the need of new samples in the problem do- main. Furthermore, a calibration model transfer can be used to get rid of offsets in reflectance, by using a standard calibration material [2]. Another new way for a standardization is to perform a latent space transformation of the spectra. We want to introduce one way of a latent space transformation via Chebyshev polynomial approximation of the spectra and compare them with results from [1] and [2]. In addition, we will reveal some pitfalls and important aspects to achieve a good performance during the way of developing a latent space via Chebyshev polynomial approximation.

References

[1] Yan Liu, Wensheng Cai, and Xueguang Shao. Standardization of near infrared spectra measured on multi-instrument. Analytica chimica acta, 836:18-23, 2014.

[2] Patrick Menz, Andreas Backhaus, Udo Seiffert. Transfer learning for transferring machine-learning based models among hyperspectral sensors. In ESANN 2019 - Proceedings.

[3] Patrick Menz, Andreas Backhaus, Udo Seiffert. Transferring machine learning models within a soft sensor system to achieve constant task performance under changing sensor hardware. In Machine Learning Reports 2016.

[*] presenter

f-Divergence and its Application to Feature Selection

Fabian Hinder

Bielefeld University, CITEC, Germany

Abstract

Nowadays most of machine learning is based on probability theory or at least justified using probabilistic arguments. It is therefore essential to examine the main subject in probability theory: probability measures. One way to do so is by considering quantitative similarity or dissimilarity measures or, more geometrically speaking distances (in this context also known as divergences), which, considering their extensive usage not only in mathematics and science but also everyday live, seems particularly promising. In the following we will shortly recap the main definitions of probability theory and the theoretical framework that is needed to describe and understand a class of divergences called f-divergences, that not only covers many important special cases (like Kullback-Leibler divergence, Jensen-Shannon divergence and total variation norm) but also allows a particularly simple, and general, kind of estimations. We will derive the notion of f-mutual information, for which we will derive two estimators (presented in [1] and [2]) and apply it to feature selection, comparing it with other methods.

References

[1] T. Sakai and M. Sugiyama. Computationally Efficient Estimation of Squared- Loss Mutual Information with Multiplicative Kernel Models. In: IEICE Trans- actions on Information and Systems E97.D.4 (2014), pp. 968971. doi: 10.1587/transinf.E97.D.968.

[2] T. Suzuki, M. Sugiyama, and T. Tanaka. Mutual information approximation via maximum likelihood estimation of density ratio. In: 2009 IEEE International Symposium on Information Theory. June 2009, pp. 463467. doi: 10.1109/ISIT.2009.5205712.

Probabilistic angle LVQ (because our medical counterparts want to know: What are the chances?)

Sreejita Ghosh

Bernoulli Inst. for Mathematics, Computer Science and Artificial Intelligence
University of Groningen

Abstract

In the medical domain there are mostly cyans, purples, oranges, and shades of greys, i.e., overlapping conditions and diagnosis are more common than a crisp single condition. In such scenarios doctors are more interested in knowing which condition(s) are often confused or appear together than just knowing the crisp labels. Thus motivated by the needs of our medical collaborators we recently developed a parameterized probabilistic version of angleLVQ, which in addition to being able to decently classify in the presence of systematic missingness, heterogeneous measurements, and imbalanced classes, can also provide the probabilities of a novel subject of belonging to each of the classes of the training set. We have used Kullback Liebler divergence as the cost function. The parameter varies with datasets. The results from the preliminary experiments are promising.

Counterfactuals for explaining machine learning models

André Artelt

Bielefeld University, CITEC, Germany

Abstract

Machine learning (ML) models are being more and more used in practice and applied to real-world scenarios. In order to be accepted by the user and because of legal regulations like the EU regulation "General Data Protection Right" (GDPR) [1], that contains a "right to an explanation", it is nowadays indispensable to explain the output and behavior of artificial intelligence (AI) in a comprehensible way. An example of easy to understand explanations of AI/ML models are counterfactual explanations [2]. A counterfactual explanation is a change of the original input that leads to a different (specific) prediction/behavior of the ML model - what has to be different in order to change the prediction of the model? In this contribution we:

- Review counterfactual explanations of ML models [2]
- Present a Python toolbox for computing counterfactual explanations [3]
- Present a mathematical modeling that lead to efficient algorithms for computing counterfactual explanations of LVQ models [4]

References

[1] European parliament and council. General data protection regulation.https://eur-lex.europa.eu/ eli/reg/2016/679/oj, 2016.

[2] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR, abs/1711.00399, 2017.

[3] André Artelt. Ceml: Counterfactuals for explaining machine learning models - a python toolbox. https://www.github.com/andreArtelt/ceml, 2019.

[4] André Artelt and Barbara Hammer. Efficient computation of counterfactual explanations of LVQ models. CoRR, abs/1908.00735, 2019.

Classification-by-Components: Current and Future Work

Lars Holdijk*[1], Sascha Saralajew[2], and Thomas Villmann[3]

[1]Radboud University, Nijmegen, Netherlands
[2] Dr. Ing. h.c. F. Porsche AG, Weissach, Germany
[3]University of Applied Sciences, Mittweida, Germany

Abstract

In the newly introduced Classification-by-Components (CBC) method input samples are classified by decomposing them into components and consecutively matching the extracted decomposition plan against the decomposition plans of possible classes. The components used in CBC models function as an extension of the prototype concept in classifiers, such as Learning Vector Quantization. Opposed to prototypes, components are not required to be class specific and can have a smaller size than the input samples. The matching of the decomposition plan is realized in CBCs by learning which components are important to be detected and which components are important to not be detected for an object to belong to a specific class. Additionally, there is also the possibility for components to be not important to the classification process at all. In this contribution we will provide a high-level overview of how the components and the class specific decomposition plans are modelled and learned in CBCs, shortly discuss the results presented in the introductory paper on CBCs and discuss the current projects and future work.

*presenter

Feature Cost Sensitive Learning Vector

Johannes Brinkrolf

University of Bielefeld, CITEC, Germany

Abstract

In many applications, it is necessary to consider not only the predictive power of the machine learning model, but also its cost at prediction time of new samples. This can be the case if many predictions should be done in real-time, running on a microchip, or the extraction of some features are more expansive (time-consuming or financial) than others. Selecting a subset of relevant features which still enable an equally well classification, can for example be done by adding a regularization term of the model parameter like Lasso. Other feature selection tools like filter methods, which consider statistical or intrinsic properties of the data, do not generally guarantee a good classification. Wrapper based algorithms have to evaluate many subsets by training the model for each one. However, these methods cannot handle costs which vary for different features. In this contribution, a new feature selection scheme is proposed for the Generalized Learning Vector Quantization (GLVQ) with adaptive metric learning. The ability to take correlation between different features and their importance for the classification into account is still retained by full relevance matrices in the learned metrics. The feature selection is done by adapting the cost function and adding a regularization term based on Lasso. Simultaneously, weights of wasteful features are pushed towards zero and can be removed for prediction without changing their outcomes. Results based on theoretical and real world data sets are demonstrated.

Phase Transitions in Layered Neural Networks: Rectified Linear Units vs. Sigmoidal Activation

Elisa Oostwal, Michiel Straat, and Michael Biehl*

Bernoulli Inst. for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen, Nehterlands

Abstract

We study layered neural networks of rectified linear units (ReLU) in a modelling framework for stochastic training processes. Here, the comparison with the case of conventional, sigmoidal activation functions is in the center of interest. Matching student-teacher scenarios are studied by applying concepts from the statistical physics of learning. In particular, we compute typical learning curves for shallow networks with K hidden units. We show that the trained networks exhibit sudden changes of their generalization behavior via the process of hidden unit specialization at a critical size of the training set. Surprisingly, our the- oretical results indicate that networks of ReLU and classical sigmoidal units display significantly different generalization and training behavior. The transition is found to be discontinuous in large networks of sigmoidal units ($K \geq 3$): Specialized hidden unit configurations compete with un- specialized ones which display poor performance. The latter persist as metastable states even for very large training sets. On the contrary, the use of ReLU activations results in continuous transitions for all K: Specialized weight configurations compete with partially specialized states of sub-optimal performance. In contrast to the case of sigmoidal units, fully unspecialized configurations become unstable above the transition.

*presenter

Better Spectral Code - Towards a Functional Autoencoder for Optimal Encoding of Hyperspectral Imaging Data

Andreas Backhaus* and Udo Seiffertt

Biosystems Engineering, Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany

Abstract

Material reflectance data, measured by hyperspectral camera systems, are a common base for the development of soft-sensors for monitoring applications in agriculture, breeding and plant research. For that purpose, vegetation is measured for a certain biochemical effect related to the task at hand which typically happens against the background of a multitude of confounding factors from environmental growth conditions and treatment as well as plant varieties. These changes can lead to decreasing soft-sensor performance while at the same time measurement campaigns are costly. Producing a fully representative problem dataset is often not an option within a competitive industrial research and development environment. In classical RGB image classification, this problem was solved by using a large amount of unlabeled image scenes to train an optimal ?encoder network? and then attach a "problem network", trained with the subset of problem data for classification or regression [1, 2] We want to transfer this idea into the domain of processing reflectance data. A unique property of hyperspectral data is its functional character. An autoencoder network should take this functional character into consideration on all levels, in the performance evaluation, the loss function as well as the autoencoder network itself. We start with the scientific question on how to assess the encoder network quality independent of a problem application, which might also be not known. We tested a number of autoencoder approaches for their ability to encode and reconstruct functional data taken from a number of measurement campaigns spanning multiple years, plant varieties, and field locations followed by extensive cross condition validation. We will compare a number of reconstruction performance measures for their usefulness to asses the quality of encoding and reconstruction.

References

*presenter

[1] Jonathan Masci, Ueli Meier, Dan Cirffsan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I, ICANN11, pages 52-59, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 2352-2360. Curran Associates, Inc., 2016.

Towards a statistical physics analysis of on-line learning in deep ReLU neural networks

Michiel Straat

University of Groningen, Netherlands

Abstract

Techniques from statistical physics can be used in the analysis of machine learning algorithms. Machine learning models, and in particular neural net- works, consist of a large number of adaptive weights. Under special assumptions, it becomes possible to model the macroscopic learning behavior of these systems by a set of deterministic differential equations. Examples of the approach for the analysis of on-line learning in two-layer sigmoidal neural networks can be found in [1, 2]. Recently, a

first statistical physics analysis of on-line gradient descent learning in two-layer ReLU neural networks has been done [3]. Now, the aim is to analyze within the framework the learning behavior of more ex- tended architectures: First, the previously studied two-layer ReLU network will be augmented with biases and second layer weights. This gives rise to a machine that is capable of representing any real-valued continuous function on compact subsets of RN, a so-called universal approximator, see [4, 5], and proved specifically for ReLU activation in [6]. Secondly, we will revisit tree-like architectures, in which the neurons' receptive

fields are non-overlapping. The consideration of these tree-like networks may prove as an important step in a potential extension of the theory towards deep neural networks.

References

[1] David Saad and Sara A. Solla. "On-Line Learning in Soft Committee Machines". In: Physical Review E 52.4 (Oct. 1, 1995), pp. 4225–4243. doi: 10.1103/PhysRevE.52.4225.

[2] Michael Biehl, Peter Riegler, and Christian Wöhler. "Transient Dynamics of On- Line Learning in Two-Layered Neural Networks". In: Journal of Physics A: Mathematical and General 29.16 (Aug. 1996), pp. 4769-4780. issn: 0305-4470. doi: 10.1088/0305-4470/29/16/005.

[3] Michiel Straat and Michael Biehl. "On-Line Learning Dynamics of ReLU Neural Networks Using Statistical Physics Techniques". In: Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Apr. 24, 2019, pp. 517-522. isbn: 978-2-87587-065-0. arXiv: 1903.07378. url: http://arxiv.org/abs/1903.07378.

[4] George Cybenko. "Approximation by Superpositions of a Sigmoidal Function". In: Mathematics of Control, Signals and Systems 2 (1989), pp. 303-314. doi: 10.1007/bf02551274.

[5] Kurt Hornik. "Approximation Capabilities of Multilayer Feedforward Networks". In: Neural Networks 4.2 (Jan. 1, 1991), pp. 251-257. issn: 0893-6080. doi: 10.1016/0893-6080(91)90009-T.

[6] Sho Sonoda and Noboru Murata. "Neural Network with Unbounded Activation Functions Is Universal Approximator". In: Applied and Computational Harmonic Analysis 43.2 (Sept. 1, 2017), pp. 233-268. issn: 1063-5203. doi: 10.1016/j.acha.2015.12.005.

Complex Maps of Behavior in Parameter Space of Neural Networks

Andreas Herzog* and Udo Seiffert

Biosystems Engineering, Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany

Abstract

One of the astonishing skills of complex biological systems is the evolving of novel and useful phenotypes by changing their genotype. But how does it work? The high dimensional parameter space of genotypes is huge and useful parameter combinations are rare. Long distance random jumps have not really a chance to reach a new genotype that produces a new useful phenotype. The key element in understanding how evolution in biological systems works is the existence of genotype networks with the same phenotype [2]. By a series of phenotypes preserving small single point changes, populations can explore different regions in genotyping space [1]. Genotype networks of different phenotypes enlace each other and come so close together that a very few changes are sufficient to switch the networks and thus reach a new useful phenotype [3]. These special properties of parameter space can be found in several biological systems (proteins, RNA molecules, gene regulation circuits), simulated in computational studies [1], and programmable hardware [4]. It seems to be an intrinsic property in complex systems with high dimensional redundant parameter space. We use the potential of genotype networks for the training of artificial neural networks. The phenotype is here a specific input-output behavior determined by the genotype, the high dimensional parameter space of network architecture and the weights. Starting from classical trained networks, we explore the parameter space by parameter shift under preserving the validation error. In result, we get a complex map of useful solutions in parameter space. The training of new tasks starts from this map and will be faster than starting from random initialization.

References

[1]Miguel A. Fortuna, Luis Zaman, Charles Ofria, and Andreas Wagner. The genotype-phenotype map of an evolving digital organism. PLOS Computational Biology, 13(2):1-20, 02 2017.

[2] Martijn A. Huynen. Exploring phenotype space through neutral evolution. Journal of Molecular Evolution, 43(3):165-169, Sep 1996.

*presenter

[3] Joshua L Payne and Andreas Wagner. The causes of evolvability and their evolution. Nature Reviews Genetics, 20:24-38, 2018.

[4] Karthik Raman and Andreas Wagner. The evolvability of programmable hardware. Journal of the Royal Society, Interface, 8(55):269-281, February 2011.

Evaluating Model Complexity for Structural Decomposition of Galaxy Images

Mohammad Mohammadi

Bernoulli Inst. for Mathematics, Computer Science and Artificial Intelligence
University of Groningen

Abstract

Galaxies are building blocks of the universe, and our understanding of their formation and evolution helps to have a better view about the universe. However, a galaxy is a complex entity, and it may contain different types of structures. These structures have many stories to tell about the formation of a galaxy. Structural decomposition of galaxy images is a topic in astronomy to discover the existing structures inside a galaxy. It is a challenging task, and it happens that for a given galaxy different people have different ideas about the type of structures. Therefore, it is necessary to have a qualitative way to evaluate the type of existing structures in a galaxy. Here, I will give a description about this problem. Then I will present an overview of the existing techniques for model selection in Machine Learning.

Neurobiological correlates of individual differences in mathematical development

Ulrike Kuhl*,[1, 2], Angela D. Friederic[2], the LEGASCREEN consortium[2, 3], and Michael A. Skeide[2]

[1]University of Bielefeld, CITEC, Germany
[2]Max Planck Institute for Human Cognitive and Brain Sciences, Germany
[3]Fraunhofer Institute for Cell Therapy and Immunology, Germany

Abstract

By acquiring core mathematical abilities in the first school years, children lay the foundation for later academic achievement. However, neural plasticity and reorganization processes associated with individual differences in early mathematical learning are still poorly understood. To fill this research gap, we followed a sample of 5-6-year-old children longitudinally to the end of second grade in school combining magnetic resonance imaging and comprehensive behavioral assessments. Our analysis revealed significant links between neuroplastic changes of cortical surface anatomy and children?s early mathematical skills. In particular, our findings suggest that distinct subregions of the parietal lobe support distinct processes contributing to mathematical cognition. Specifically, children?s visuospatial magnitude processing was related by the change in cortical thickness in the right superior parietal cortex. Moreover, children?s early arithmetic abilities were associated with the change in cortical folding in the right intraparietal sulcus. Additional associations were found for arithmetic abilities and cortical thickness change of the right temporal lobe, and visuospatial abilities and right precentral thickness as well as right medial frontal gyrus folding plasticity. Importantly, these effects were independent of other individual differences in IQ, literacy and maternal education. Our findings highlight the critical role of cortical plasticity during the acquisition of fundamental mathematical abilities.

*presenter

Approximation Domain Adaptation via Low-Rank Basis

Christoph Raab

UAS Würzburg-Schweinfurt, Germany

Abstract

Transfer learning focuses on the reuse of supervised learning models in a new context. Prominent applications can be found in robotics, image processing or web mining. In these areas, learning scenarios change by nature, but often remain related and motivate the reuse of existing supervised models. While the majority of symmetric and asymmetric domain adaptation algorithms utilize all available source and target domain data, we show that domain adaptation requires only a substantial smaller subset. This makes it more suitable for real-world scenarios where target domain data is rare. The presented approach

finds a target subspace representation for source and target data to address domain differences by orthogonal basis transfer. We employ Nystrom techniques and show the reliability of this approximation without a particular landmark matrix by applying post-transfer normalization. It is evaluated on typical domain adaptation tasks with standard benchmark data.

Topological Data Analysis, a thriving field

Abolfazl Taghribi

Bernoulli Inst. for Mathematics, Computer Science and Artificial Intelligence
University of Groningen

Abstract

From genetics to the economy, from computer science to quantum physics, Topological seems to solve many distinct problems. In recent years, it found several new applications in machine learning and for studying high dimensional data. Topological Data Analysis can be used as an unsupervised or semisupervised method for computing features inside a large amount of data which is unique for that data type. Here we use This technique to count the number of supernovas and measure their size inside a galaxy. Moreover, we propose a new method for decreasing the amount of computation which is needed for approximating the same properties. We will show on some example that our method is faster and its results provide a closer approximation of the data with respect to other sampling methods. Using this tool, one can follow the behavior of supernovas inside a simulation and examine them in more details.

Comprehensive study on Random Projection in non-stationary environments

Moritz Heusinger[*] and Frank-Michael Schleif

UAS Würzburg-Schweinfurt, Germany

Abstract

Random Projection (RP) is a popular and efficient technique to preprocess high-dimensional data and to reduce its dimensionality. While RP has been widely used and evaluated in stationary data analysis scenarios this is not the case for non-stationary environments. In this paper we provide a comprehensive evaluation of RP on streaming data. We discuss why RP can be used on Streaming Data w.r.t. the Johnson-Lindenstrauss bound and also how it can deal with Stream specific situations, i. e. Concept Drift and Feature Drift. We also provide experiments with RP on Streaming Data, using state-of-the-art Streaming Classifiers like Adaptive Hoeffding Tree and Concept Drift Detectors (e. g. Adaptive Windowing and Kolmogorov-Smirnov Windowing), to evaluate its efficiency.

keywords Streaming Data, High Dimensional Data, Random Projection

[*] presenter

Knowledge representations for robotic systems

Lydia Fischer

Honda Research Institute Europe GmbH

Abstract

The current trend in research is to come up with intelligent systems that are able to act and interact properly in the real world, doing useful tasks. In contrast to humans one can inject knowledge of various kind in such systems before their operation. This is handy if systems have a very specific task to solve. It is getting more complicated if different tasks need to be solved by the same system because different information and skills are required. Nevertheless it is impractical to provide all potential relevant information and skills a-priori, since no designer can foresee all potential use-cases and situations such systems can encounter. A more promising approach is to equip systems with basic useful informations as well as a basic skill set, and a mechanism to dynamically learn new information and skills. One possible way to encode information and skills for intelligent robotic systems is to use a knowledge representation, e. g. knowledge graphs. Such representations can be modeled in a machine interpretable from. A pilot project for having a robot with a hand-designed knowledge representation doing a specific task is described in [1]. There the range of understood tasks was increased by using common sense data bases. To overcome limitations of this early approach, a more general way of describing knowledge is formulated in [2]. A key challenge is to find ways to translate knowledge from common sense databases, mostly very differently structured, in a form compliant with the knowledge representation of intelligent robotic systems.

References

[1] L. Fischer, S. Hasler, J. Deigmoller, T. Schnurer, M. Redert, U. Pluntke, K. Nagel, C. Senzel, J. Ploennigs, A. Richter, and J. Eggert Which tool to use? Grounded reasoning in everyday environments with assistant robots. CogRob@KR, 3-10, 2018.

[2] J. Eggert, J. Deigmoller, L. Fischer, and A. Richter Memory Nets: A Knowledge Representation for Autonomous Entities. International Conference on Knowledge Engineering and Ontology Development, accepted, 2019.

Visualizing the decision function of deep networks learning models

Alexander Schulz[*], Fabian Hinder, and Barbara Hammer

Bielefeld University, CITEC, Germany

Abstract

Recent progress in the field of deep neural networks produces increasingly powerful models which are able to achieve human level and partially even super human performance [4, 3]. However, these networks are growing in complexity making them increasingly difficult to comprehend and more vulnerable to adversarial attacks [5]. Most of the present literature focusses on explaining the investigated model with respect to individual data samples [1]. Following ideas from [2], we propose a novel methodology to visualize the decision function of a deep neural network in two dimensions. For this purpose, we propose to compute a discriminative dimensionality reduction based on the Fisher information using the neural network without the necessity to calculate gradients for each distance calculation. This visualization then allows to directly observe important properties of the trained model and the used data, such as adversarial examples, multimodality or biased data. Additionally, this method is complementary to interpretation methods present in existing literature as mentioned above and might be even more useful when combined with those.

References

[1] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73:1 – 15, 2018.

[2] A. Schulz, A. Gisbrecht, and B. Hammer. Using discriminative dimensionality reduction to visualize classifiers. Neural Processing Letters, 42(1):27–54, Aug 2015.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrit- twieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484, 2016.

[4] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks, 32:323 – 332, 2012. Selected Papers from IJCNN 2011.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. preprint arXiv:1312.6199, 2013.

[*] presenter

First Steps on Incident Detection for Field Bus Systems

Dmitrii Lekomtsev*[1], Anne Satzke[1], Philip Karopka[1,2], Horst-Michael Gross[3], Tina Geweniger[2], and Sven Hellbach*, [2]

[1]Indu-Sol GmbH, Schmölln
[2]Fakultät Physikalische Technik/Informatik, Westsächsische Hochschule Zwickau
[3]Neuroinformatics and Cognitive Robotics Lab, Technical University Ilmenau

Abstract

Diagnostic network information consists of different parameters which are difficult to understand without the knowledge of the scope. In this work, we evaluated the network condition in Process Field Network (PROFINET) using Artificial Neural Networks (ANNs) and collected diagnostic information. ANN decides whether the network is functioning normally or not. An essential part of the work is data preprocessing. It is done using quantization, data aligning, reducing the number of inputs and other preprocessing techniques to create a new version of the dataset to improve the accuracy. The obtained data makes possible to do a number of experiments and to find out what approach of data preprocessing shows the best results. The results were evaluated on two datasets. The first one contains diagnostic data of a well-functioning network, and the second one consists of data in which network problems were detected. The highest accuracy obtained in this work is 98.91when the network is working fine. These first results hint, that the posed problem provides possibilities for further investigations. Hence, our goal is to establish an ongoing research cooperation.

keywords PROFINET, Network Diagnostics, Artificial Neural Networks, Machine Learning, Industrial Ethernet.

*presenter

Mathematical Implications of the GLVQ Margin Analysis for the Non-Eulidean Case

Sascha Saralajew[1] and Thomas Villmann*[2]

[1]Dr. Ing. h.c. F. Porsche AG Weissach, Germany
[2]University of Applied Sciences Mittweida – SICIM, Germany

Summery of the Talk

This contribution re-investigates the paper 'Margin Analysis of LVQ' by Crammer et al. [1]. In particular we address the problem of margin analysis and discuss the resulting consequences with respect to robustness of GLVQ networks as well as adversarial attacks [2] as an follow-up of [3].

Additionally, we explain in detail the mathematical implications, if GLVQ is not based on the Euclidean distance, which requires to consider of perceptron networks with so-called Banach-like-perceptrons in the proofs given by Crammer et al. . Banach-like-perceptrons are based on semi-inner products instead of the standard (Euclidean) inner product [5].

Further, we shortly discuss in the light of margin analysis the case of non-standard transfer (activation) functions in GLVQ-networks as suggested in [4] for better convergence and performance.

References

[1] K. Crammer, R. Gilad-Bachrach, A. Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

*presenter

[2] S. Gu and L. Rigazio. Towards deep neural network architectures robust to adversarial examples. In *Proceedings of the 3rd International Conference of Learning Representation (ICLR), San Diego (USA)*, 2015.

[3] S. Saralajew, L. Holdijk, M. Rees, and T. Villmann. Robustness of generalized learning vector quantization models against adversarial attacks. In A. Vellido, K. Gibert, C. Angulo, and J. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 189–199. Springer Berlin-Heidelberg, 2019.

[4] T. Villmann, J. Ravichandran, A. Villmann, D. Nebel, and M. Kaden. Investigation of activation functions for Generalized Learning Vector Quantization. In A. Vellido, K. Gibert, C. Angulo, and J. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 179–188. Springer Berlin-Heidelberg, 2019.

[5] T. Villmann, A. Villmann, and M. Kaden. Generalized multilayer perceptrons using Banach-like perceptons based on semi-inner products. *Machine Learning Reports*, 13(MLR-02-2019):35–44, 2019. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/˜fschleif/mlr/mlr_02_2019.pdf.

Large Margin Learning Vector Quantization

Benjamin Paaßen

CITEC, Bielefeld University

Abstract

Learning vector quantization approaches are advantageous for their online learning capability, their low memory requirements, interpretable model, easy extensibility to multi-class problems, and their fast decision function [6]. However, a key drawback is the highly non-liner and non-convex loss of most LVQ variants, making optimization at times numerically difficult and prone to local optima. We suggest to slightly vary the usual LVQ loss in order to get closer to a convex setup. In particular, we propose to adapt the Large Margin Nearest Neighbor (LMNN) loss of Weinberger and Saul [9] to learning vector quantization models.

We obtain a dissimilarity version of large margin LVQ as a non-convex quadratic program and a kernel version as a convex quadratic program, the latter being analogous to the multi-prototype support vector machine developed by Aiolli and Sperduti [1]. First empiric results suggest that further work is required to make large margin LVQ useful in practice.

Assume we wish to learn K prototypes $w_1, \ldots, w_K \in \mathcal{X}$ with labels $z_1, \ldots, z_K \in \{1, \ldots, L\}$, such that as many data points $x_1, \ldots, x_m \in \mathcal{X}$ with labels $y_1, \ldots, y_m \in \{1, \ldots, L\}$ are correctly classified by a one-nearest neighbor assignment, i.e. data point x_i is assigned the label of the closest prototype. Then, data point x_i is classified *correctly* if and only if the closest prototype has the correct label. Now, let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be some dissimilarity on $\mathcal{X}$, let $d_{i,l} = d(x_i, w_l)^2$ be the (squared) dissimilarity to to the lth prototype, let $d_i^+ := \min_{k:z_k=y_i} d_{i,k}$ be the (squared) dissimilarity to the closest prototype with the correct label, and let $I_i := \{l | y_i \neq z_l\}$ be the set of prototype indices with different label than the ith data point. Then, the prototype-based LMNN loss for our problem is

$$\ell(w_1, \ldots, w_K) = \overbrace{\sum_{i=1}^m d_i^+}^{\text{pull}} + \frac{1}{2C} \cdot \overbrace{\sum_{i=1}^m \sum_{l \in I_i} \text{ReLU}\big(d_i^+ - d_{i,l} + \gamma\big)^2}^{\text{push}} \tag{1}$$

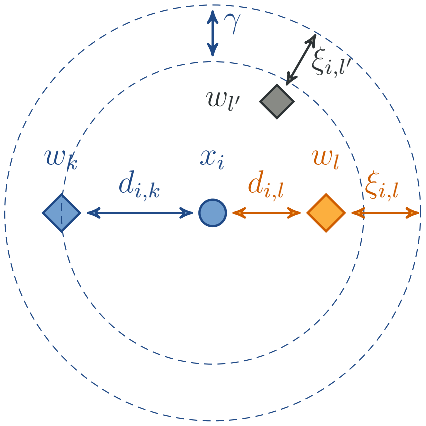


Figure 1: An illustration of the large-margin learning vector quantization loss 1. From the perspective of data point x_i (blue circle), the closest prototype with the same label is w_k (blue diamond). The pull loss tries to move w_k closer to x_i. Further, w_l (orange diamond) is closer to x_i and has a different label, which would mean that x_i would be misclassified, which punish via the push loss. Moreover, we punish all prototypes that are within a margin of safety of γ (outer dashed circle), e.g. $w_{l'}$ (grey diamond). The contribution to the push loss is $\xi_{i,l}$ and $\xi_{i,l'}$ respectively, i.e. the *slack* that is needed to satisfy margin constraints.

where $\text{ReLU} : \mathbb{R} \to \mathbb{R}^+$ is the rectified linear unit aka the hinge loss, defined as $\text{ReLU}(\mu) = \max\{0, \mu\}$. We call this loss the *large margin learning vector quantization* (LM-LVQ) loss.

Intuitively, the first term pulls prototypes closer to data points in their receptive field and the second term pushes prototypes with wrong labels away if they invade a margin of safety around the data point. Also refer to Figure 1 for an illustration of this loss. The hyper-parameter $C \in \mathbb{R}^+$ weighs between the push and pull forces.

Note that the push force can be seen as a variant of generalized learning vector quantization [8] with the nonlinearity $\Phi(\mu) = \text{ReLU}(\mu + \gamma)^2$, albeit without the dissimilarity normalization.

Also note that, in contrast to LMNN, we square the push loss contributions, which has the advantage that the gradient is smooth at point 0 and supports our later optimization steps.

To optimize loss 1, we re-write it first as an optimization problem with slack variables $\xi_{i,l}$, which express how far prototype w_l invades the margin of data point i.

$$\begin{aligned} \min_{\substack{w_1,\ldots,w_K,\\ \Xi \in \mathbb{R}^{m \times K}}} \quad & \frac{1}{2 \cdot C} \sum_{i=1}^{m} \sum_{l \in I_i} \xi_{i,l}^2 + \sum_{i=1}^{m} d_i^+ & (2) \\ \text{s.t.} \quad & \xi_{i,l} \geq d_i^+ - d_{i,l} + \gamma & \forall i \in \{1, \ldots, m\} \quad \forall l \in I_i \\ & \xi_{i,l} \geq 0 & \forall i \in \{1, \ldots, m\} \quad \forall l \in \{1, \ldots, K\} \end{aligned}$$

Note that this problem is a quadratically constrained quadratic program [2]. Unfortunately, this program is not convex due to the negative sign in front of $d_{i,l}$. Still, we can make a solution attempt by constructing the Wolfe dual of this problem.

Disregarding the non-negativity constraints for the slack variables for the moment, we obtain the following Lagrangian.

$$\mathcal{L}(\boldsymbol{W}, \Xi, \boldsymbol{\Lambda}) = \frac{1}{2 \cdot C} \sum_{i=1}^{m} \sum_{l \in I_i} \xi_{i,l}^2 + \sum_{i=1}^{m} d_i^+ - \sum_{i=1}^{m} \sum_{l \in I_i} \lambda_{i,l} \cdot \Big(\xi_{i,l} + d_{i,l} - d_i^+ - \gamma \Big)$$

Let now P_k be the set of all points x_i to which w_k is the closest prototype with the same label, and let $N_k = \{j | y_j \neq z_k\}$ be the set of data points x_j with a different label than w_k. We also call these sets the *positive* and *negative receptive field* of prototype w_k. Then, we can re-write the Lagrangian as follows.

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}, \Xi, \boldsymbol{\Lambda}) = & \sum_{k=1}^{K} \sum_{i \in P_k} \big(1 + \sum_{l \in I_i} \lambda_{i,l}\big) \cdot d_{i,k} + \sum_{i \in N_k} (-\lambda_{i,k}) \cdot d_{i,k} & (3) \\ & + \frac{1}{2 \cdot C} \sum_{i=1}^{m} \sum_{k \in I_i} \xi_{i,k}^2 - \sum_{i=1}^{m} \sum_{k \in I_i} \lambda_{i,k} \cdot \xi_{i,k} + \gamma \cdot \sum_{i=1}^{m} \sum_{k \in I_i} \lambda_{i,k} \end{aligned}$$

We further introduce the auxiliary variables $\beta_{k,i}$, defined as $1 + \sum_{l \in I_i} \lambda_{i,l}$ if $i \in P_k$, as $-\lambda_{i,k}$ if $i \in N_k$, and as zero otherwise. Accordingly, the first line of the Lagrangian simplifies to $\sum_{k=1}^{K} \sum_{i=1}^{m} \beta_{k,i} \cdot d_{i,k}$.

Our next step is to compute the gradient of our Lagrangian with respect to the optimization variables and set it to zero. To do so, we need to make an assumption about the underlying data space. In particular, we assume that the dissimilarity d is self-equal and symmetric, i.e. for any two data points $x, y \in \mathcal{X}$ we obtain $d(x, x) = 0$ and $d(x, y) = d(y, x)$. Under these assumptions, Pekalska and Duin [7] guarantee a *pseudo-Euclidean embedding*, i.e. there exist two mappings, $\phi^+ : \mathcal{X} \to \mathbb{R}^n$ and $\phi^- : \mathcal{X} \to \mathbb{R}^n$ from the data space into $\mathbb{R}^n$ for some n, such that for any $x, y \in \mathcal{X}$ it holds:

$$d(x, y)^2 = ||\phi^+(x) - \phi^+(y)||^2 - ||\phi^-(x) - \phi^-(y)||^2$$

Accordingly, we obtain the following gradient of the Lagrangian with respect to $\phi^+(w_k)$ and $\phi^-(w_k)$:

$$\nabla_{\phi^+(w_k)}\mathcal{L}(\boldsymbol{W}, \Xi, \boldsymbol{\Lambda}) = \sum_{i=1}^{m} \beta_{k,i} \cdot 2 \cdot \big(\phi^+(w_k) - \phi^+(x_i)\big)$$

$$\nabla_{\phi^-(w_k)}\mathcal{L}(\boldsymbol{W}, \Xi, \boldsymbol{\Lambda}) = \sum_{i=1}^{m} \beta_{k,i} \cdot (-2) \cdot \big(\phi^-(w_k) - \phi^-(x_i)\big)$$

Setting these to zero yields:

$$\phi^+(w_k) = \frac{\sum_{i=1}^m \beta_{k,i} \cdot \phi^+(x_i)}{\sum_{i=1}^m \beta_{k,i}}, \quad \text{and} \quad \phi^-(w_k) = \frac{\sum_{i=1}^m \beta_{k,i} \cdot \phi^-(x_i)}{\sum_{i=1}^m \beta_{k,i}}$$

In other words, the prototype is an affine combination of data points. This, in turn, lets us re-write the squared dissimilarities $d_{i,k}$ as follows [4]:

$$d_{i,k} = \frac{\sum_{j=1}^m \beta_{k,j} \cdot d(x_i, x_j)^2}{\sum_{j=1}^m \beta_{k,j}} - \frac{\vec{\beta}_k^T \cdot \boldsymbol{D}^2 \cdot \vec{\beta}_k}{2 \cdot \big(\sum_{j=1}^m \beta_{k,j}\big)^2}$$

where $\boldsymbol{D}$ is the matrix of all squared pairwise dissimilarities $d(x_i, x_j)^2$.

Finally, consider the derivative of our Lagrangian with respect to $\xi_{i,k}$:

$$\frac{\partial}{\partial \xi_{i,l}}\mathcal{L}(\boldsymbol{W}, \Xi, \boldsymbol{\Lambda}) = \frac{1}{C} \cdot \xi_{i,l} - \lambda_{i,l} \overset{!}{=} 0 \qquad \iff \qquad \xi_{i,l} = C \cdot \lambda_{i,l}$$

Note that this result ensures that the non-negativity constraints for $\xi_{i,l}$ hold because we already have non-negativity constraints for the Lagrange multipliers $\lambda_{i,l}$.

Plugging these results back into our Lagrangian 3, we obtain:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\Lambda}) =& \sum_{k=1}^{K}\sum_{i=1}^{m} \beta_{k,i} \cdot \Big(\frac{\sum_{j=1}^m \beta_{k,j} \cdot d(x_i, x_j)^2}{\sum_{j=1}^m \beta_{k,j}} - \frac{\vec{\beta}_k^T \cdot \boldsymbol{D}^2 \cdot \vec{\beta}_k}{2 \cdot \big(\sum_{j=1}^m \beta_{k,j}\big)^2}\Big) \\
&+ \frac{1}{2 \cdot C} \sum_{i=1}^{m}\sum_{k \in I_i} C^2 \cdot \lambda_{i,k}^2 - \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k} \cdot C \cdot \lambda_{i,k} + \gamma \cdot \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k} \\
=& \sum_{k=1}^{K} \Big(\frac{\sum_{i=1}^m \sum_{j=1}^m \beta_{k,i} \cdot \beta_{k,j} \cdot d(x_i, x_j)^2}{\sum_{j=1}^m \beta_{k,j}}\Big) - \Big(\frac{\vec{\beta}_k^T \cdot \boldsymbol{D}^2 \cdot \vec{\beta}_k}{2 \cdot \big(\sum_{j=1}^m \beta_{k,j}\big)^2} \cdot \sum_{i=1}^{m} \beta_{k,i}\Big) \\
&- \frac{C}{2} \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k}^2 + \gamma \cdot \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k} \\
=& \frac{1}{2} \sum_{k=1}^{K} \frac{\vec{\beta}_k^T \cdot \boldsymbol{D}^2 \cdot \vec{\beta}_k}{\sum_{j=1}^m \beta_{k,j}} - \frac{C}{2} \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k}^2 + \gamma \cdot \sum_{i=1}^{m}\sum_{k \in I_i} \lambda_{i,k}
\end{aligned}$$

Accordingly, we obtain the following Wolfe dual form.

$$\min_{\mathbf{\Lambda} \in \mathbb{R}^{m \times K}} \quad -\frac{1}{2} \sum_{k=1}^{K} \frac{\vec{\beta}_k^T \cdot \boldsymbol{D} \cdot \vec{\beta}_k}{\sum_{j=1}^m \beta_{k,j}} + \frac{C}{2} \sum_{i=1}^{m} \sum_{k \in I_i} \lambda_{i,k}^2 - \gamma \cdot \sum_{i=1}^{m} \sum_{k \in I_i} \lambda_{i,k} \tag{4}$$

$$\text{s.t.} \quad \lambda_{i,k} \geq 0, \quad \beta_{k,i} = \begin{cases} 1 + \sum_{l \in I_i} \lambda_{i,l} & \text{if } i \in P_k \\ -\lambda_{i,k} & \text{if } i \in N_k \\ 0 & \text{otherwise} \end{cases} \qquad \begin{matrix} \forall k \in \{1, \ldots, K\}, \\ \forall i \in \{1, \ldots, m\} \end{matrix}$$

Note that solving the Wolfe dual does *not* guarantee an optimal solution for the primal problem due to non-convexity [2]. Still, we can hope to achieve *good* solutions, given that the dual in itself has a rather intuitive interpretation: If all Lagrange multipliers are zero, prototypes gracefully degenerate to the means of their receptive fields. If this is sufficient to classify the data set correctly, this is the single optimal solution. However, if any prototype l invades the margin of a data point i, the Lagrange multiplier $\lambda_{i,l}$ must be increased, which means that prototype l moves *away* from data point i, whereas this multiplier is *added* to $\beta_{k,i}$ for the closest correct prototype k, such that this prototype moves *closer* to data point i, resembling the typical LVQ behavior.

Another problem is that the dual is numerically problematic due to the sum over all $\beta_{k,i}$ coefficients in the denominator of the quadratic term. Fortunately, this can be addressed by imposing that $\sum_{i=1}^m \beta_{k,i} = |P_k|$ for all k, i.e. that the sum of all β coefficients should sum to the size of the receptive field. This side constraint enforces that the sum of all coefficients stays the same as in its initial state when all Lagrange multipliers are zero. Interestingly, this side constraint is also equivalent to introducing bias terms (we omit the full derivation here for brevity).

We further note that this dual can be transformed into a standard quadratic problem by expressing the vectors $\vec{\beta}_k$ as a (sparse) affine transformation from the Lagrange multipliers. In particular, we can re-write $\vec{\beta}_k = \boldsymbol{A_k} \cdot \vec{\lambda} + \vec{1}_{P_k}$, where $\vec{\lambda}$ is the concatenation of all rows in $\mathbf{\Lambda}$, where $a_{k,i,(i-1)\cdot K+k} = -1$ if $i \in N_k$, $a_{k,i,(i-1)\cdot K+l} = +1$ if $i \in P_k$ and $l \in I_i$ and zero otherwise, and where $\vec{1}_{P_k}$ is a m-dimensional vector which is 1 at entries $i \in P_k$ and zero otherwise. Then, problem 4 becomes:

$$\min_{\vec{\lambda} \in \mathbb{R}^{m \cdot K}} \quad \frac{1}{2} \vec{\lambda}^T \cdot \Big(C \cdot \boldsymbol{I} - \sum_{k=1}^{K} \boldsymbol{A_k^T} \cdot \frac{\boldsymbol{D}}{|P_k|} \cdot \boldsymbol{A_k} \Big) \cdot \vec{\lambda} - \Big(\gamma \cdot \vec{1}^T + \sum_{k=1}^{K} \vec{1}_{P_k}^T \cdot \frac{\boldsymbol{D}}{|P_k|} \cdot \boldsymbol{A_k} \Big) \cdot \vec{\lambda}$$

$$\begin{aligned} \text{s.t.} \quad & \vec{\lambda} \geq 0 \\ & \vec{1}^T \cdot \boldsymbol{A_k} \cdot \vec{\lambda} = 0 \qquad \forall k \in \{1, \ldots, K\} \end{aligned} \tag{5}$$

Still, this quadratic problem is not necessarily convex because $\boldsymbol{D}$ is indefinite, yielding a convex shape in some search directions and a concave shape in other directions. The regularization term and the equality side constraints may ensure a convex shape, but this appears to not be guaranteed in general. Still, we expect a more smooth loss function compared to relational GLVQ [5] due to the absent dissimilarity normalization.

Finally, we note that our problem formulation relies on the assumption that the positive receptive fields P_k stay fixed, similar to LMNN. However, we can relax this assumption by running a multi-pass scheme where the positive receptive fields are updated after each optimization of problem 5 until the fields do not change anymore. Because any update in positive receptive field is guaranteed to reduce the loss, this scheme is guaranteed to converge, similar to Multi-pass LMNN [3].

If we wish to guarantee proper convexity, we can switch from a dissimilarity formulation to a kernel formulation. In that case, our primal loss changes to:

$$\ell(w_1, \ldots, w_K) = \sum_{i=1}^{m} d_i^+ + \frac{1}{2C} \cdot \sum_{l \in I_i} \text{ReLU}\big(s_{i,l} - s_i^+ + \gamma\big)^2 \tag{6}$$

where $s_{i,l}$ is the kernel value between data point i and prototype l and d_i^+ refers to the dissimilarity between data point i and the closest prototype with the same label, where the dissimilarity measure is given as $d(x, y)^2 = s(x, x) - 2s(x, y) + s(y, y)$. Via a very similar derivation as before, we obtain the following Wolfe dual:

$$\begin{aligned}
\min_{\vec{\lambda} \in \mathbb{R}^{m \cdot K}} \quad & \frac{1}{2}\vec{\lambda}^T \cdot \Big(\sum_{k=1}^{K} \boldsymbol{A_k}^T \cdot \frac{\boldsymbol{S}}{|P_k|} \cdot \boldsymbol{A_k} + C \cdot \boldsymbol{I}\Big) \cdot \vec{\lambda} + \Big(\sum_{k=1}^{K} \vec{1}_{P_k}^T \cdot \frac{\boldsymbol{S}}{|P_k|} \cdot \boldsymbol{A_k} - \gamma \cdot \vec{1}^S\Big) \cdot \vec{\lambda} \\
\text{s.t.} \quad & \vec{\lambda} \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (7) \\
& \vec{1}^T \cdot \boldsymbol{A_k} \cdot \vec{\lambda} = 0 \qquad \forall k \in \{1, \ldots, K\}
\end{aligned}$$

where $\boldsymbol{S}$ is the matrix of pairwise kernel values $s(x_i, x_j)$. This problem is a standard convex quadratic program with linear side constraints. Indeed, the kernel matrix may even be slightly indefinite because the regularization term $C \cdot \boldsymbol{I}$ applies an implicit shift eigenvalue correction.

This latter form is almost equivalent to the multi-prototype SVM suggested by Aiolli and Sperduti [1], except for the regularization, which here pushes prototypes to the center of the receptive field, whereas the multi-prototype SVM pushes prototypes toward the origin.

Closing Remarks: Merging LMNN and LVQ concepts appears promising from these first attempts. In particular, we obtained a non-convex, but "wellbehaved" dissimilarity formulation for pseudo-Euclidean dissimilarities, and a convex quadratic program formulation for kernels. Unfortunately, first empiric experiments suggest that the approach may not be able to compete with a simple one-versus-one SVM. Likely, further improvements are required - be it in terms of concept or optimization - in order to make this approach useful in practical applications. Still, large margin LVQ offers a conceptual bridge, shedding further light on the strong relations between LVQ, LMNN, and SVMs.

References

[1] Fabio Aiolli and Alessandro Sperduti. "Multiclass classification with multi-prototype support vector machines". In: *Journal of Machine Learning Research* 6 (2005), pp. 817–850. URL: `http://www.jmlr.org/papers/volume6/aiolli05a/aiolli05a.pdf`.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004. URL: `https://web.stanford.edu/~boyd/cvxbook/`.

[3] Christina Göpfert, Benjamin Paaßen, and Barbara Hammer. "Convergence of Multi-pass Large Margin Nearest Neighbor Metric Learning". In: *Proceedings of the 25th International Conference on Artificial Neural Networks (ICANN 2016)*. (Barcelona, Spain). Ed. by Alessandro E.P. Villa, Paolo Masulli, and Antonio Javier Pons Rivero. Vol. 9886. Springer Nature, Aug. 2016, pp. 510–517. DOI: `10.1007/978-3-319-44778-0_60`. URL: `https://pub.uni-bielefeld.de/record/2905729`.

[4] Barbara Hammer and Alexander Hasenfuss. "Topographic Mapping of Large Dissimilarity Data Sets". In: *Neural Computation* 22.9 (2010), pp. 2229–2284. DOI: `10.1162/NECO_a_00012`.

[5] Barbara Hammer et al. "Learning vector quantization for (dis-)similarities". In: *Neurocomputing* 131 (2014), pp. 43–51. DOI: `10.1016/j.neucom.2013.05.054`.

[6] David Nova and Pablo A. Estévez. "A review of learning vector quantization classifiers". In: *Neural Computing and Applications* 25.3 (2014), pp. 511–524. DOI: `10.1007/s00521-013-1535-3`. URL: `https://arxiv.org/abs/1509.07093`.

[7] Elzbieta Pekalska and Robert Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2005. ISBN: 9812565302.

[8] Atshushi Sato and Keiji Yamada. "Generalized Learning Vector Quantization". In: *Proceedings of the 7th conference on Advances in Neural Information Processing Systems (NIPS 1995)*. Ed. by G. Tesauro, D. Touretzky, and T. Leen. Cambridge, MA, 1995, pp. 423–429. URL: `https://papers.nips.cc/paper/1113-generalized-learning-vector-quantization`.

[9] Kilian Weinberger and Lawrence Saul. "Distance Metric Learning for Large Margin Nearest Neighbor Classification". In: *Journal of Machine Learning Research* 10 (2009), pp. 207–244. URL: `http://www.jmlr.org/papers/v10/weinberger09a.html`.

Generalized Multilayer Perceptrons Using Banach-like Perceptons Based on Semi-Inner Products

Thomas Villmann[1], Andrea Villmann[2], and Marika Kaden[1]

[1]University of Applied Sciences Mittweida – SICIM, Germany
[2]Berufliches Schulzentrum Döbeln-Mittweida, Germany

1 Motivation

Multilayer perceptrons (MLP) are nowadays the standard networks in machine learning for classification and regression tasks [1, 8]. Motivated by pyramid cells in biological neural networks the mathematical perceptron is the basis of those networks [14], see Fig. 1.

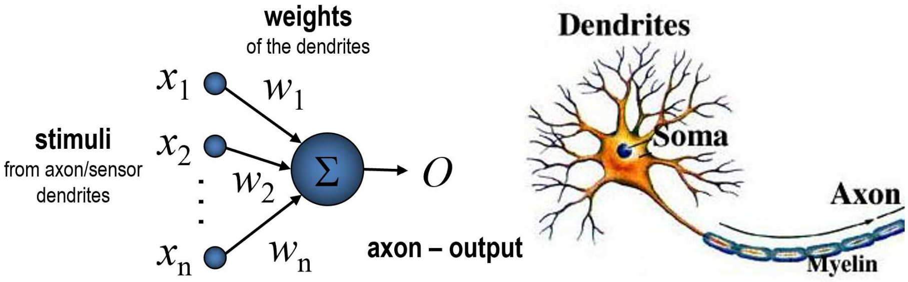


Figure 1: Schematic illustration of a mathematical perceptron (left) according to a pyramid cell (right). The input vector $\mathbf{x} = (x_1, \ldots, x_n)^T$is weighted by the weight vector $\mathbf{w} = (w_1, \ldots, w_n)^T$ to generate the output O.

The capability for these networks is justified by Cybenko's theorem with states the universal approximation property. One key ingredient in the proof of this theorem is the Hilbert-space-property which relates to the standard mathematical perceptron based on the Euclidean inner product.

In this paper we consider multilayer networks consisting of Banach-like perceptrons. Those perceptrons are based on semi-inner products (SIP) instead of the Euclidean inner product. Semi-inner products are related to Banach-spaces.

We show that for those networks Cybenkos theorem still can be applied if it is modified slightly. For this purpose, we have to consider special cases of Banach spaces. Particularly, it turns out that the SIP, which is related to the l_p-space, can be used to replace the inner product in standard perceptrons.

2 Introduction and Basic Concepts

The mathematical modeling of standard perceptrons assumes stimulus vectors $\mathbf{x} \in \mathbb{R}^n$ and a weight vector $\mathbf{w} \in \mathbb{R}^n$ to generate the output according to

$$O(\mathbf{w}, \mathbf{x}) = f(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{1}$$

where $b \in \mathbb{R}$ is the bias and f is the so-called activation function. The quantity $\langle \mathbf{w}, \mathbf{x} \rangle = \sum_{k=1}^{n} x_k \cdot w_k$ is the (real) Euclidean inner product, which is motivated biologically by the weighted sum of inputs, see Fig. 1. The activation function f usually is a monotonically increasing function. Frequent choices are the Heaviside function

$$H(z) = \begin{cases} 1 & \text{for} \quad z > 0 \\ 0 & \text{else} \end{cases}$$

and the identity $\mathrm{id}(z) = z$ or

$$f_\theta(z) = \frac{1}{1 + \exp(\theta z)}$$

as standard sigmoid function.

MLPs are directed graphs with mathematical perceptrons as nodes. The perceprons are arranged in layers. Only the first layer (input layer) receives direct data inputs. The last layer is denoted as output layer and delivers the network response $\mathbf{o}$ for a given data vector $\mathbf{x}$. The stimulus vectors of perceptrons in all layers except the input layer are output vectors of previous layers.

Mathematically speaking, MLPs realize a mapping

$$F_{W,B} : \mathbb{R}^n \ni \mathbf{x} \longmapsto \mathbf{o} \in \mathbb{R}^m \tag{2}$$

if m output units are available and W is the set of all weights $\mathbf{w}$ and B is the set of all biases in the network. It was shown by Cybenko that under certain conditions MLP's are universal approximators [4].

In this note we discuss generalizations of this statement. Particularly, we discuss the replacement of the Euclidean inner product in (1) by kernels or semi-inner products. We will give conditions such that the statements regarding the universal approximation properties given by Cybenko remain valid.

3 MLPs as Universal Approximators

3.1 Cybenkos Results for Standard MLP

We start with a brief explanation of the main result provided by Cybenko in [4] regarding the approximation completeness of standard MLPs .

Definition 1. The function σ is discriminatory with respect to the inner product $\langle\cdot,\cdot\rangle$ if for a measure $\mu \in \mathcal{M}(I_n)$ of the closed (compact) subset $I_n = [0,1]^n \subset \mathbb{R}^n$ with the property

$$\int_{I_n} \sigma(\langle\mathbf{w},\mathbf{x}\rangle + b)\, d\mu(\mathbf{x}) = 0$$

for all $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ the implication $\mu \equiv 0$ follows.

Further we define what a sigmoidal function should be.

Definition 2. The function σ is sigmoidal if

$$\sigma(z) \longrightarrow \begin{cases} 1 & \text{for} \quad z \to \infty \\ 0 & \text{for} \quad z \to -\infty \end{cases}$$

holds.

The following Lemma relates sigmoidal functions to discriminatory functions:

Lemma 3. *Any bounded, measurable sigmoidal function is discriminatory and, hence, any continuous sigmoidal function is discriminatory.*

The main statement regarding the universal approximation is given by the following theorem. For the sake of later considerations we also give the proof of the theorem as provided in [4].

Theorem 4. *Let σ be a continuous discriminatory function and*

$$G(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \cdot \sigma(\langle\mathbf{w}_j,\mathbf{x}\rangle + b_j) \tag{3}$$

be the finite sum of perceptrons (1) with activation function $f = \sigma$. Let $I_n = [0,1]^n \subset \mathbb{R}^n$ be the closed hypercube equipped with the Euclidean metric. Then the set $\mathcal{G}$ of functions $G(\mathbf{x})$ is dense in the space $\mathcal{C}(I_n)$ of continuous functions over I_n.

Proof. The set $\mathcal{G}$ is dense in $\mathcal{C}(I_n)$ iff for any function $g(\mathbf{x}) \in \mathcal{C}(I_n)$ and $\varepsilon > 0$ exists a function $G(\mathbf{x}) \in \mathcal{G}$ with $|G(\mathbf{x}) - g(\mathbf{x})| < \varepsilon$ for all $\mathbf{x} \in I_n$. This statement is proven if we can show that for the closure $\overline{\mathcal{G}}$ of $\mathcal{G}$ the equality $\overline{\mathcal{G}} = \mathcal{C}(I_n)$ holds. We apply a proof by contradiction:

Obviously, $\mathcal{G}$ is a linear subspace of $\mathcal{C}(I_n)$. Thus, the closure $\overline{\mathcal{G}}$ is a closed subspace of $\mathcal{C}(I_n)$. We remark that I_n is equipped with the Euclidean norm such that it is a Banach-space or, more precisely, a Hilbert space. Now we assume that $\overline{\mathcal{G}} \neq \mathcal{C}(I_n)$, i.e. $\mathcal{G}$ is not dense in $\mathcal{C}(I_n)$. Hence, according to the Hahn-Banach-theorem [15], there is a bounded linear functional L on $\mathcal{C}(I_n)$ with $L(h) \neq 0$, i.e. it is not completely vanishing for $h \in \mathcal{C}(I_n)$ but $L(\mathcal{G}) = L\left(\overline{\mathcal{G}}\right) = 0$ is valid. We remark that L is continuous and we have $L \in \mathcal{C}^*(I_n)$ being the dual space of $\mathcal{C}(I_n)$.

According to the Hilbert-space property of I_n we can apply the Riesz Representation Theorem (RRT, [13]) which states that the functional L can be written in the form

$$L(h) = \int_{I_n} h(\mathbf{x})\, d\mu(\mathbf{x}) \tag{4}$$

for some measure $\mu \in \mathcal{M}(I_n)$ and a continuous function $h \in \mathcal{C}(I_n)$. Yet, so far μ is unspecified.

Because for the continuous function $\sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b) \in \overline{\mathcal{G}}$ is valid for all $\mathbf{w}$ and b we must have that

$$L(\sigma) = \int_{I_n} \sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)\, d\mu(\mathbf{x}) = 0$$

holds for all $\mathbf{w}$ and b according to $L\left(\overline{\mathcal{G}}\right) = 0$. Since σ is assumed to be discriminatory, the zero integral implies that $\mu \equiv 0$ has to be valid, which further implies, however, that $L(h) \equiv 0$ for any $h \in \mathcal{C}(I_n)$. This contradicts the assumption $\overline{\mathcal{G}} \neq \mathcal{C}(I_n)$. Hence, $\mathcal{G}$ is dense in $\mathcal{C}(I_n)$ which completes the proof. □

We remark the following:

Remark 5. In the proof of the theorem the Hilbert-space property of I_n was explicitly used which is guaranteed by the Euclidean metric/norm. Further, the Euclidean norm in I_n is consistent with the mathematical structure of the discriminatory functions $\sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)$ containing the Euclidean inner product as argument.

3.2 Generalizations

3.2.1 Kernels for Hilbert-Spaces

Obviously, the proof of the Cybenko-theorem remains valid if we replace the Euclidean inner product $\langle \mathbf{w}, \mathbf{x}\rangle$ in the standard perceptron (1) by an arbitrary inner product and use the resulting norm as norm for the n-dimensional real space $\mathbb{R}^n$. We can continue approach and, more generally, replace the inner product by a kernel κ, i.e. we consider

$$\kappa(\mathbf{w}, \mathbf{x}) = \langle \phi(\mathbf{w}), \phi(\mathbf{x})\rangle$$

with $\phi(\mathbf{w}) \in \mathcal{H}$ and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). Then $\mathcal{I}_n = \phi(I_n)$ is a compact Hilbert space and the Cybenko's theorem is still applicable but now for $\mathcal{I}_n$.

3.2.2 Semi-Inner Products

In the second case we want to exchange in the perceptron (1) the inner product $\langle \mathbf{w}, \mathbf{x} \rangle$ by a semi-inner product (SIP) $[\mathbf{w}, \mathbf{x}]$ [11].

Definition 6. A mapping $[\cdot, \cdot] : \mathcal{B} \times \mathcal{B} \to \mathbb{C}$ is called a semi-inner product (SIP) if the following relations are fulfilled:

1. linearity: $[\lambda \mathbf{x} + \mathbf{z}, \mathbf{y}] = \lambda [\mathbf{x}, \mathbf{y}] + [\mathbf{z}, \mathbf{y}]$ for $\lambda \in \mathbb{C}$

2. positiveness: $[\mathbf{x}, \mathbf{x}] > 0$ for $\mathbf{x} \neq \mathbf{0}$

3. Cauchy-Schwarz-inequality: $|[\mathbf{x}, \mathbf{y}]|^2 \leq [\mathbf{x}, \mathbf{x}] [\mathbf{y}, \mathbf{y}]$

Lumer has shown that a SIP always generates a norm by $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ as well as he has proofed that every Banach-space with norm $\|\mathbf{x}\|_{\mathcal{B}}$ is equipped with a SIP generating this norm [11]. One can show that the relation $[\mathbf{x}, \lambda \mathbf{y}] = \overline{\lambda} [\mathbf{x}, \mathbf{y}]$ follows immediately [7].

Now we equip I_n with the norm $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ denoted as $I_n^{\mathcal{B}} \subset \mathbb{R}_{\mathcal{B}}^n$. Thus $\mathbb{R}_{\mathcal{B}}^n$ is an n-dimensional real Banach-space. Considering now *Banach-like perceptrons* with output

$$O(\mathbf{w}, \mathbf{x}) = f([\mathbf{w}, \mathbf{x}] + b) \tag{5}$$

using SIPs, we cannot simply apply the Cybenko-theorem, because the Hilbert-space property needed in the proof for the RRT is violated for $I_n^{\mathcal{B}}$. However, there exist variants of the RRT which suppose special Banach-spaces instead of a Hilbert-spaces. In the following we will characterize those Banach-spaces.

Theorem 7. *Let $\mathcal{B}$ be an uniformly convex Banach space with continuous SIP $[\cdot, \cdot]$. Then a RRT analogously to (4) is valid.*

Proof. [7, Theorem 6] □

The theorem can be extended to:

Theorem 8. *Let $\mathcal{B}$ be a reflexive Banach space. Then a RRT analogously to (4) is valid.*

Proof. Let $\mathcal{B}$ be a reflexive Banach space and $h \in \mathcal{B}^* = \mathcal{C}(\mathcal{B})$. Then exists a SIP $[\cdot, \cdot]$ and an element $\beta \in \mathcal{B}$ such that $\varphi(\mathbf{x}) = [\mathbf{x}, \beta]$ is a continuous linear functional [6]. Hence, the respective SIP determines a RRT analogously to (4). □

Both theorems are related according to the following lemma

Lemma 9. *Every smooth (continuous) uniformly convex Banach space is also reflexive and strictly convex. The reverse direction is not valid. Hence, Theorem 7 is a special case of Theorem 8.*

Proof. [6] □

Theorem 10. *Let σ be a continuous discriminatory function with respect to the SIP $[\cdot,\cdot]$ for $I_n^{\mathcal{B}} \subset \mathbb{R}_{\mathcal{B}}^n$ equipped with the norm $\|\mathbf{x}\| = \sqrt{[\mathbf{x},\mathbf{x}]}$ such that $\mathbb{R}_{\mathcal{B}}^n$ is a reflexive n-dimensional real Banach-space. Additionally, let*

$$G_{\mathcal{B}}\left(\mathbf{x}\right) = \sum_{j=1}^{N} \alpha_j \cdot \sigma\left(\left[\mathbf{w}_j, \mathbf{x}\right]_p + b_j\right) \tag{6}$$

be the finite sum of Banach-like perceptrons (5) with activation function $f = \sigma$. Then $G_{\mathcal{B}}\left(\mathbf{x}\right)$ is an universal approximator.

Proof. The proof is in complete analogy to the proof of the Cybenko-theorem. The application of the Hahn-Banach-theorem is not affected by the weaker assumption regarding the Banach-space. The existence of a respective RRT is guaranteed by the previous theorems. □

Most famous examples for Banach-spaces are the spaces L^p and l^p. The latter one is equipped with the unique SIP

$$\left[\mathbf{w},\mathbf{x}\right]_p = \frac{1}{\|\mathbf{x}\|_p^{p-2}} \sum_k w_k \cdot |x_k|^{p-1} \cdot \operatorname{sgn}\left(x_k\right) \tag{7}$$

with $1 \leq p \leq \infty$ [7]. Thus we can equip $I_n^{\mathcal{B}}$ with the SIP $\left[\mathbf{w},\mathbf{x}\right]_p$. Further we can state the following lemma:

Lemma 11. *Both L^p and l^p are uniformly convex for $1 < p < \infty$.*

Proof. [9] □

Corollary 12. *The compact set $I_n^{\mathcal{B}}$ with the SIP $\left[\mathbf{w},\mathbf{x}\right]_p$ from (7) is an uniformly reflexive Banach space for $1 < p < \infty$. Hence, a RRT analogously to (4) is valid.*

Proof. Just apply Theorem 8. □

The last corollary leads to the following statement:

Lemma 13. *A MLP using Banach-like perceptrons with output*

$$O_p\left(\mathbf{w},\mathbf{x}\right) = f\left(\left[\mathbf{w},\mathbf{x}\right]_p + b\right)$$

according to (5) generated by the SIP $\left[\mathbf{w},\mathbf{x}\right]_p$ from (7) is an universal approximator for $1 < p < \infty$.

Proof. The previous corollary about uniform convexity of l_p together with Lemma (9) guarantee that Theorem (10) is applicable. □

According to statement in [17] a RRT is also valid for generalized SIP-spaces.

Zhang & Zhang considered generalized SIPs (gSIP) [17] extending a first attempt by Nath [12]. They considered SIPs $\left[\mathbf{w},\mathbf{x}\right]_\xi$ for a function $\xi : \mathbb{R}_+ \to \mathbb{R}_+$

fulfilling the requirements 1) and 2) of Def. 6. The Cauchy-Schwarz-inequality is replaced by

$$\left|[\mathbf{w},\mathbf{x}]_{\xi}\right| \leq \xi\left([\mathbf{w},\mathbf{w}]_{\xi}\right)\cdot\psi\left([\mathbf{x},\mathbf{x}]_{\xi}\right)$$

for a conjugate function $\psi:\mathbb{R}_{+}\to\mathbb{R}_{+}$, i.e. $\xi(t)\cdot\psi(t)=t$ has to be valid. For a RRT regarding those gSIPs it is assumed that $\xi(t)$ is a so-called gauge function, i.e. $\xi(0)=0$ and $\lim_{t\to\infty}\xi(t)=\infty$. If $\xi(t)$ is surjective onto $\mathbb{R}_{+}$ and $\zeta(t)=\frac{\xi^{-1}(t)}{t}$ is a gauge function on $\mathbb{R}_{+}$ then a RRT can be formulated, because the resulting Banach-space is reflexive and strictly convex [17].

3.2.3 Kernels for Banach-Spaces

In this step we extend Cybenkos theorem to the case of kernels regarding reproducing kernel Banach spaces (RKBS). As stated in [16, Theorem 4], a RKBS is always reflexive. Thus, we suppose a kernel $\kappa_{\mathcal{B}}$ corresponding to the kernel feature map $\phi_{\mathcal{B}}:I_n\to\mathcal{I}_n\subset\mathcal{B}$ with $\mathcal{B}$ being a RKBS. From Theorem 8 we can conclude that Cybenko's theorem is applicable accordingly.

3.2.4 Indefinite Inner Product Spaces

Definition 14. An indefinite inner product (IIP) is a mapping $[\![\cdot,\cdot]\!]:X\times X\to\mathbb{C}$ from a vector space X into the complex plane if the following relations are fulfilled [2]:

1. linearity: $[\![\lambda\mathbf{x}+\mathbf{z},\mathbf{y}]\!]=\lambda[\![\mathbf{x},\mathbf{y}]\!]+[\![\mathbf{z},\mathbf{y}]\!]$ for $\lambda\in\mathbb{C}$
2. Hermitean symmetry: $[\![\mathbf{x},\mathbf{y}]\!]=\overline{[\![\mathbf{y},\mathbf{x}]\!]}$

The quantity $\widehat{n}(\mathbf{x})=[\![\mathbf{x},\mathbf{x}]\!]$ might become negative. Hence, it defines neither a norm nor it is sub-additive. Therefore, we consider the quantity $n(\mathbf{x})=\sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|}$ instead. Then we can state the following lemma:

Lemma 15. *The functional* $n(\mathbf{x})=\sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|}$ *is sub-linear.*

Proof. We have for $\lambda>0$

$$\begin{aligned} n(\lambda\mathbf{x}) &= \sqrt{|[\![\lambda\mathbf{x},\lambda\mathbf{x}]\!]|}\\ &= \sqrt{|\lambda^{2}[\![\mathbf{x},\mathbf{x}]\!]|}\\ &= \lambda\sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|}\end{aligned}$$

which shows the homogenity and

$$\begin{aligned} n(\mathbf{x}+\mathbf{y}) &= \sqrt{|[\![\mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}]\!]|} \\ &= \sqrt{|[\![\mathbf{x},\mathbf{x}]\!]+[\![\mathbf{x},\mathbf{y}]\!]+[\![\mathbf{y},\mathbf{x}]\!]+[\![\mathbf{y},\mathbf{y}]\!]|} \\ &\leq \sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|+|[\![\mathbf{x},\mathbf{y}]\!]|+|[\![\mathbf{y},\mathbf{x}]\!]|+|[\![\mathbf{y},\mathbf{y}]\!]|} \\ &\leq \sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|+|[\![\mathbf{y},\mathbf{y}]\!]|} \\ &\leq \sqrt{|[\![\mathbf{x},\mathbf{x}]\!]|}+\sqrt{|[\![\mathbf{y},\mathbf{y}]\!]|} \\ &= n(\mathbf{x})+n(\mathbf{y}) \end{aligned}$$

verifying the sub-additivity. This completes the proof. □

In the next step we consider the vector space equipped with the introduce sub-linear function $n(\mathbf{x})$. [5]

Appendix

In this appendix we give some useful definitions regarding Banach-spaces used in the text as well as some basic properties.

Definition 16. Let X be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $\varphi : X \to \mathbb{K}$ be a functional. It is denoted as sub-linear if both

- positive homogeneity: $\varphi(\lambda\mathbf{x}) = \lambda\varphi(\mathbf{x})$ for $\lambda \in \mathbb{R}_+$ and $\varphi(i\mathbf{x}) = i\varphi(\mathbf{x})$ is valid in the complex case
- sub-additivity: $\varphi(\mathbf{x}+\mathbf{y}) \leq \varphi(\mathbf{x}) + \varphi(\mathbf{y})$

hold.

We remark that every norm on X is sub-linear. The *Hahn-Banach-Theorem* is stated as follows [10, 13, 15]:

Theorem 17. *Variant* ***a****) Let X be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $Y \subseteq X$ a subspace. Let $\varphi : X \to \mathbb{R}$ be a sub-linear functional and $f : Y \to \mathbb{K}$ be a linear functional with $\Re(f(\mathbf{y})) \leq \varphi(\mathbf{y})$ for all $\mathbf{y} \in Y$. Then there exists a linear functional $F : X \to \mathbb{K}$ with $F|_Y = f$ and $\Re(F(\mathbf{x})) \leq \varphi(\mathbf{x})$ is valid for all $\mathbf{x} \in X$.*

An alternative formulation is variant ***b****): Let X be a normed space and Y is a subspace $Y \subset X$. Let be $f \in X^*$ with $f|_Y = 0$. The subspace Y is dense in X iff under these assumptions always follows $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in X$.*

Definition 18. A Banach space $\mathcal{B}$ is denoted as *strictly convex* iff for $\mathbf{x}, \mathbf{y} \neq 0$ with $\|\mathbf{x}\| + \|\mathbf{y}\| = \|\mathbf{x} + \mathbf{y}\|$ we can always conclude that $\mathbf{x} = \lambda \mathbf{y}$ for some $\lambda > 0$.

Lemma 19. *A Banach space $\mathcal{B}$ with SIP $[\cdot, \cdot]$ is strictly convex iff for $\mathbf{x}, \mathbf{y} \neq 0$ with $[\mathbf{x}, \mathbf{y}] = \|\mathbf{x}\| \cdot \|\mathbf{y}\|$ we can always conclude that $\mathbf{x} = \lambda \mathbf{y}$ for some $\lambda > 0$.*

Proof. [7] □

The following definition for the uniform convexity was introduced in [3]:

Definition 20. A Banach space $\mathcal{B}$ is denoted as *uniformly convex* iff for each $\varepsilon > 0$ exist a $\delta(\varepsilon) > 0$ such that if $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ with $\|\mathbf{x} - \mathbf{y}\| > \varepsilon$ then $\frac{\|(\mathbf{x}+\mathbf{y})\|}{2} < 1 - \delta(\varepsilon)$ is valid.

Definition 21. A Banach space $\mathcal{B}$ with SIP $[\cdot, \cdot]$ is denoted as *continuous* iff

$$\Re\{[\mathbf{x}, \mathbf{y} + \lambda \mathbf{x}]\} \underset{\lambda \to 0}{\longrightarrow} \Re\{[\mathbf{x}, \mathbf{y}]\}$$

is valid for $\lambda \in \mathbb{R}$. The space is *uniformly continuous* iff this limit is approached uniformly.

Definition 22. A Banach space $\mathcal{B}$ is denoted as *reflexive* iff the mapping $J : \mathcal{B} \to \mathcal{B}^{**}$ is surjective.

Theorem 23. *Let $\mathcal{B}$ be a Banach space. Then a necessary and sufficient condition for $\mathcal{B}$ to be reflexive is that for every $f \in \mathcal{B}^*$ exists an SIP $[\cdot, \cdot]$ and an element $\mathbf{y} \in \mathcal{B}$ with $f(\mathbf{x}) = [\mathbf{x}, \mathbf{y}]$ for all $\mathbf{x} \in \mathcal{B}$. If $\mathcal{B}$ is strictly convex then $\mathbf{y}$ is unique.*

Proof. [6, Theorem 2] □

Definition 24. A Banach space $\mathcal{B}$ is denoted as *smooth* iff for each $\mathbf{x} \in \mathcal{B}$ with $\|\mathbf{x}\| = 1$ there exist a linear functional $f_{\mathbf{x}} \in \mathcal{B}^*$ with $f_{\mathbf{x}}(\mathbf{x}) = \|f_{\mathbf{x}}\|$. The existence of $f_{\mathbf{x}}$ is guaranteed by the Hahn-Banach-Theorem.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning.* Springer Science+Business Media, LLC, New York, NY, 2006.

[2] J. Bognár. *Indefinite Inner Product Spaces.* Springer-Verlag, 1974.

[3] J. Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40:396–414, 1936.

[4] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[5] M. Dritschel and J. Rovnyak. Operators in indefinite inner product spaces. In A. Böttcher, A. Dijksma, H. Langner, M. Dritschel, J. Robnyak, M. Kaashoek, and P. Lancaster, editors, *Lectures on Operator Theory and Its Applications*, number 3 in Fields Institute Monographs, pages 143–234. American Mathematical Society, Fields Institute, 1995.

[6] G. D. Faulkner. Representation of linear functionals in a Banach space. *Rocky Mountain Journal of Mathematics*, 7(4):789–792, 1977.

[7] J. Giles. Classes of semi-inner-product spaces. *Transactions of the American Mathematical Society*, 129:436–446, 1967.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[9] O. Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(19):239–244, 1956.

[10] A. Kolmogorov and S. Fomin. *Reelle Funktionen und Funktionalanalysis*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.

[11] G. Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[12] B. Nath. Topologies on generalized semi-inner product spaces. *Composito Mathematica*, 23(3):309–316, 1971.

[13] F. Riesz and B. Sz.-Nagy. *Vorlesungen über Functionalanalysis*. Verlag Harri Deutsch, Frankfurt/M., 4th edition, 1982.

[14] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–408, 1958.

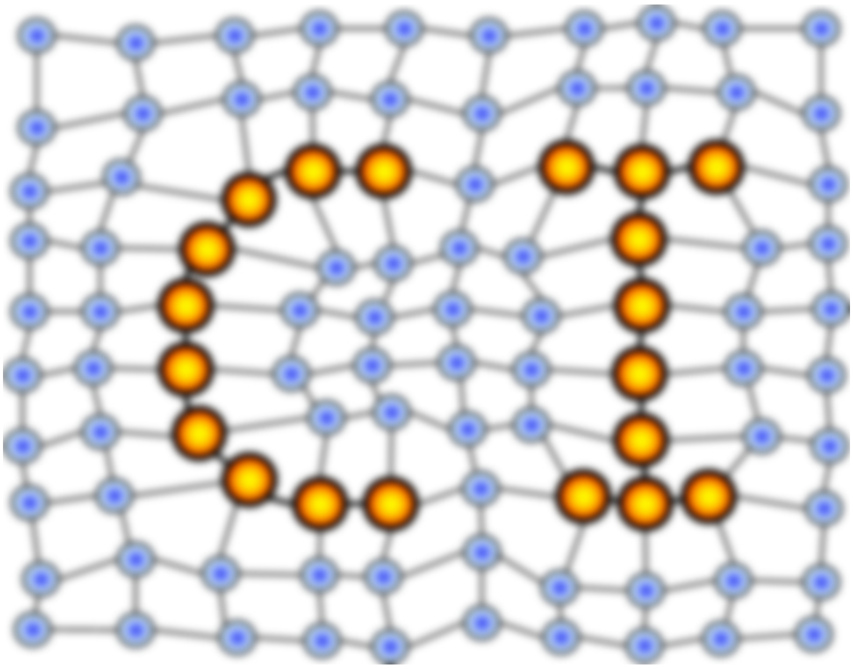
[15] H. Triebel. *Analysis und mathematische Physik*. BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 3rd, revised edition, 1989.

[16] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

[17] H. Zhang and J. Zhang. Generalized semi-inner products with applications to regularized learning. *Journal of Mathematical Analysis and Application*, 372:181–196, 2010.

Machine Learning Reports

Report 02/2019



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.