Topics 2020

Topics for the course CS-E4875 Research Project in Machine Learning, Data Science and Artificial Intelligence for Autumn 2020.

Topic #1: Implementing and testing ML-based (Gaussian processes and neural networks) ODE models in Julia

Background:

Recently proposed ML-based ordinary differential equation (ODE) models have been revolutionary in continuous time modeling. Machine learning models involving ODE systems are trained via back-propagation, very much like all other ML models. Interestingly, computing the gradients of models involving ODE systems requires solving another ODE system called *adjoints*. It has been recently shown that the classical adjoint method suffers from the fact that it is not reversible. In addition, neural ODE systems are claimed to be solved using an ODE solver that is not suitable for such a system.

Both problems (irreversibility and incorrect solver) are because of the deficiencies in commonly used ODE solvers in Tensorflow and Pytorch. In this project, the goal is to implement these models in a high-level programming language called **Julia**. Since a variety of ODE solvers are implemented in Julia, we believe that state-of-the-art results obtained in Tensorflow and Pytorch would be improved with the new implementation.

Potential readings:

- Neural ordinary differential equations: https://arxiv.org/pdf/1806.07366.pdf
- DiffEqFlux.jl – A Julia Library for Neural Differential Equations: https://julialang.org/blog/2019/01/fluxdiffeq
- Bonus: https://www.juliabloggers.com/the-essential-tools-of-scientific-machine-learning-scientific-ml/
- Bonus: ODE2VAE: Deep generative second order ODEs with Bayesian neural networks: https://arxiv.org/pdf/1905.10994.pdf

Prerequisite:

- Basic machine learning knowledge
- Programming skills in Python and Julia
- Bonus: familiarity with deep learning frameworks such as PyTorch.

Instructor (name and email): Cagatay Yildiz (cagatay.yildiz@aalto.fi, contact person) and Harri Lahdesmaki (harri.lahdesmaki@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #2: Medical Natural Language Processing with Deep Learning

Background:

Natural language processing (NLP) allows the machine to understand human language. However, there remain several challenges in medical text understanding. Diagnosis notes contain complex diagnosis information, which includes a large number of professional medical vocabulary and noisy information such as non-standard synonyms and misspellings. Textual clinical notes are lengthy documents, usually from hundreds to thousands of tokens. Thus, medical text understanding requires effective feature representation learning and complex cognitive process. This topic is about investigating deep learning-based NLP techniques for medical text understanding. Specific tasks include medical concept extraction, medical natural language inference, medical code assignment, medical entity recognition, and relation extraction. This project focuses on recent advances in medical NLP and will implement deep neural networks on one or two specific tasks to reproduce the results or improve the performance. It can also be about some visualization of the medical text representation or evaluating the medical knowledge learned by the neural models.

References:
- SECNLP: A survey of embeddings in clinical natural language processing
- Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets
- Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

Prerequisite: 1) basics of deep learning; 2) programming skills with deep learning frameworks (e.g., PyTorch).

Instructor: Shaoxiong Ji (shaoxiong.ji@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #3: Federated Deep Learning

Background:

Recent advances in machine learning techniques have drawn attention from many researchers and have boosted many industrial applications. However, traditional machine learning requires to collect massive data from users and trains a centralized model for prediction. To solve large-scale learning problems, distributed machine learning is proposed for training in a distributed manner by allocating the learning process in multiple computing nodes. Federated learning is a new distributed learning paradigm that decouples data collection and model training via multi-party computation and model aggregation. The rapid development of deep neural networks facilitates the learning techniques for modeling complex problems and emerges into federated deep learning under the federated setting. This project will investigate federated averaging and its variants on the LEAF benchmark (especially the language modeling task) with the potential to propose a new aggregation method.

References:
- Communication-Efficient Learning of Deep Networks from Decentralized Data
- Federated learning: Challenges, methods, and future directions
- https://github.com/TalwalkarLab/leaf

Prerequisite: 1) basics of deep learning; 2) programming skills with deep learning frameworks (e.g., PyTorch).

Instructor: Shaoxiong Ji (shaoxiong.ji@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #4: Utilization of Liquid-chromatography (LC) System Descriptors for Retention Time Prediction

Background:

Research in Metabolomics, among other things, is concerned with the determination of the the molecular composition of biological samples, such as blood, plant extracts or waste water. In practice, a challenge thereby is the complexity of the sample. i.e. typically many (very) similar molecules are simultaneously present in the sample. However, to analyze the sample a certain isolation of the individual molecules is needed, e.g. to determine their molecular structure. A commonly applied strategy to reduce the complexity of the samples, to ease the isolation, is Liquid-chromatogray (LC). In short, LC separates molecules over *time* by differently strong interacting with the molecules in a samples, that is solved and pumped through the LC systems. The sample molecules leave the LC system at different, so called, *retention times* (RT).

The RT of a particular molecules depends on its molecular structure. A lot of research has been done to predict RTs from molecular structures and utilize such predictions, e.g. for the identification of the molecules in a samples. However, the RT *also* depends on the configuration of the LC system. The vast majority of RT prediction models, however, do not utilize this information, e.g. due to the lack of standardized LC system descriptions.

In the frame of this thesis topic, the aim of the student is to conduct a literature review on LC system *feature* descriptions that could be utilized in machine learning approaches. Some key questions:

- What RT prediction methods exists in the literature that take into account LC system configurations?
- What are the key properties of LC systems that are candidates to be modeled?
- How can those properties be represented as features ready for machine learning algorithms?
- How can existing RT prediction tools be extended to make use LC system features?

Prerequisite:

- basic knowledge of machine learning, data science and mathematics
- basic knowledge on chemoinformatics and molecular biology is beneficial
- drive to acquire domain knowledge (here chemistry / chemoinformatics)

References:

- "Introduction to Retention Time Prediction using Machine Learning", Bach (2017)
- "Cross-column retention prediction in reversed-phase high-performance liquid chromatography by artificial neural network modelling", D'Archivio et al. (2012)
- "Artificial Neural Network Prediction of Retention of Amino Acids in Reversed-Phase HPLC under Application of Linear Organic Modifier Gradients and/or pH Gradients", D'Archivio et al. (2019)

Instructor (name and email): Eric Bach (eric.bach@aalto.fi)

Language: English

Topic available: yes_one_instance

Topic available also for a group: no

Topic #5: Project in Deep (Reinforcement) Learning

Background:

In this project, the task is to reproduce one of the recently published results in deep (reinforcement) learning. The topic can be selected together with the student.

Prerequisite: knowledge of deep learning and PyTorch

Instructor (name and email): Alexander Ilin (alexander.ilin@aalto.fi)

Language: English

Topic available: yes

Topic available also for a group: yes

Topic #6: Parallel projection predictive (Bayesian) variable and structure selection

Background:

Projection predictive inference is a decision-theoretic approach whose goal is to replace a reference model (a Bayesian model that is fit taking into account all variables in the data) with a constrained model by projecting the posterior distribution. In the context of variable selection, one typically restricts the projection to include only a subset of the original variables. In practice, this is solved by individually projecting posterior draws and minimising the Kullback-Leibler (KL) divergence between the reference model's and the projection's predictive distributions. This approach has been shown to achieve a superior performance to other state of the art methods 1, 2. Nonetheless, its main bottleneck resides in needing to solve the projection for several posterior draws, which is a embarrassingly parallel subproblem that is currently solved sequentially.

Prerequisite: knowledge of R, Bayesian statistics and modelling.

References:

- https://projecteuclid.org/euclid.ejs/1589335310
- https://arxiv.org/abs/1503.08650
- https://arxiv.org/abs/2004.13118

Instructor (name and email): Alejandro Catalina (alejandro.catalina@aalto.fi)

Language: English

Topic available: yes

Topic available also for a group: no

Topic #7: Stein thinning for projection predictive (Bayesian) variable and structure selection

Background:

Projection predictive inference is a decision-theoretic approach whose goal is to replace a reference model (a Bayesian model that is fit taking into account all variables in the data) with a constrained model by projecting the posterior distribution. In the context of variable selection, one typically restricts the projection to include only a subset of the original variables. In practice, this is solved by individually projecting posterior draws and minimising the Kullback-Leibler (KL) divergence between the reference model's and the projection's predictive distributions. This approach has been shown to achieve a superior performance to other state of the art methods 1, 2. However, it is clear that the cost of solving the projection relies on the number of draws to project. Piironen et al. (2020) show that a small subset of posterior draws is enough for an accurate projection. Currently, the method uses a k-means clustering procedure to decide which subset to use. On the other hand, Stein thinning (3) addresses the problem of optimal subsample, picking the subset of draws whose empirical distribution is closest to the original, by minimising a kernel Stein discrepancy. The goal of this project is to implement Stein thinning as the subsample procedure to decide which draws to use for the projections.

Prerequisite: knowledge of R, Bayesian statistics and modelling.

References:

- https://projecteuclid.org/euclid.ejs/1589335310
- https://arxiv.org/pdf/2005.03952.pdf
- https://arxiv.org/abs/1503.08650
- https://arxiv.org/abs/2004.13118

Instructor (name and email): Alejandro Catalina (alejandro.catalina@aalto.fi)

Language: English

Topic available: yes

Topic available also for a group: no

Topic #8: Efficient LOO model comparison for projection predictive (Bayesian) variable and structure selection

Background:

Projection predictive inference is a decision-theoretic approach whose goal is to replace a reference model (a Bayesian model that is fit taking into account all variables in the data) with a constrained model by projecting the posterior distribution. In the context of variable selection, one typically restricts the projection to include only a subset of the original variables. In practice, this is solved by individually projecting posterior draws and minimising the Kullback-Leibler (KL) divergence between the reference model's and the projection's predictive distributions. This approach has been shown to achieve a superior performance to other state of the art methods 1, 2. After performing the variable selection, we decide which of all projections is the optimal one (i.e. reaches the reference model's performance with the fewest variables or additive components). For models with large number of components this comparison can be very expensive to perform. Magnusson et al. (2020) scale LOO model comparison to scenarios where we want to estimate the difference in ELPD for a large number of models by introducing the difference operator. The goal of this project is to implement this procedure for projection predictive variable selection and reduce the computational burden for comparing projections.

Prerequisite: knowledge of R, Bayesian statistics and modelling.

References:

- https://projecteuclid.org/euclid.ejs/1589335310
- http://proceedings.mlr.press/v108/magnusson20a/magnusson20a.pdf
- https://arxiv.org/abs/1503.08650
- https://arxiv.org/abs/2004.13118

Instructor (name and email): Alejandro Catalina (alejandro.catalina@aalto.fi)

Language: English

Topic available: yes

Topic available also for a group: no

Topic #9: Normalizing flows for graph structured data generation

Background: Graph structured data provides useful representation for a lot of domains such as social networks, bioinformatics or robotics. Generative models allow us to learn a latent space from the training data which can be explored for data generation such as for drug discovery with molecular graphs. Graph structure imposes additional challenges due to arbitrary shape and size of graphs and discrete structure. Normalizing flows present a promising direction to map a base distribution to a discrete graph distribution. In this project, you will review the SOTA methods for flow based generative methods and adapt them for graph generation.

Prerequisite: Basic knowledge of deep learning and mathematics (course work or self-taught), familiarity with Python and any deep learning framework of choice

Instructor (name and email): Anirudh Jain (anirudh.jain@aalto.fu)

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #10: Interpretable latent space for Variational Auto-Encoders

Background: Learning interpretable factors within the latent space of a variational auto-encoder(VAE) is a promising direction for deep generative models. Disentangled latent space allow us to learn factors for meaningful features of the data. For example, learning factors for thickness, digit label etc for the MNIST dataset. Interpretability is also useful for generation of new data by letting us explore the latent space efficiently in meaningful directions. In this project, you will perform a literature survey on existing techniques for interpretable latent space for VAEs and extend existing SOTA methods for different datasets or use cases.

Prerequisite: Basic knowledge of deep learning and mathematics (course work or self-taught), familiarity with Python and any deep learning framework of choice

Instructor (name and email): Anirudh Jain (anirudh.jain@aalto.fu)

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #11: Analyzing log data from an online learning tool

Background:
Online courses can produce vast quantities of data about the learning process and learner performance, which could be further used for developing the quality of learning. Educational data mining and Learning analytics are two overlapping fields that model and analyze learning data, such as log data, to measure and understand the learning process in an online learning platform. One important subtask is to predict or simply analyze the learner's performance in a course, thus allowing teachers prepare pedagogical interventions to support learners or, ultimately, develop intelligent learning tools that automatically adapt to learners' needs. The log data from online learning tools can be summarized using various visualizations, such as traditional graphs as well as heatmaps and animations, to make data more easily understandable.

This project focuses on analyzing a sample from learner data collected from an online course on writing skills at Aalto University. In the course, learners interact with online exercises by pointing and clicking text, and these interactions are stored in a json log file. In this project, you would also create visualizations or animations based on collected data sample.

Project aims:
1) to analyze log data (json), for example, from the following points of view:

- basic statistical analysis describing e.g., the number of attempts and time spend on different activities
- searching for interesting trends and patterns (e.g., trial-and-error strategies, relationship between spent time, number of attempts and given feedback)
- identifying exercises which are very easy or too difficult
- comparing data sets to identify whether the amount of feedback affects learner performance

2) to generate visualizations to support analysis based on the sample data, such as graphs, heatmaps or animations

References:

- Automatic Assessment of Source Code Highlighting Tasks: Investigation of Different Means of Measurement, Kramer et al. (2018): https://doi.org/10.1145/3279720.3279729

Prerequisite:

- Programming skills (e.g., Python)
- basics of computational data analysis
- Optional: basic knowledge of R or scikit-learn (https://scikit-learn.org/stable/)

Instructors (name and email):
Jan-Mikael Rybicki <jan-mikael.rybicki@aalto.fi>
Wilhelmiina Hämäläinen <wilhelmiina.hamalainen@aalto.fi>

Topic available: yes_one_instance
Topic available also for a group: yes (max 2)

Topic #12: Cognitive modelling through simulator-based inference

Background: Computational models can offer deep explanations of human cognition through correspondence between theoretical assumptions and empirical observations of human behaviour. The theoretical assumptions are usually encoded through a generative model or a simulator, and the empirical observations are collected from a real human. The connection of the two can be established through parameters of the simulator that can produce the collected data. One family of methods that solves the inverse problem of finding the simulator parameters for a specific user is called simulator-based inference. In this project the student will perform inference (using methods from opensource ELFI library) on two cognitive modelling tasks (small implementation adjustments may be required): ACTR skill acquisition, and menu visual search models.

Prerequisite: good knowledge of Python, familiarity with Bayesian Optimization and Gaussian Processes.

References:

- The frontier of simulation-based inference https://arxiv.org/pdf/1911.01429.pdf – a very nice high-level overview paper of the simulator-based inference;
- A Tutorial on Bayesian Optimization https://arxiv.org/pdf/1807.02811.pdf – a gentle introduction to BO and GPs;
- Parameter Inference for Computational Cognitive Models with Approximate Bayesian Computation https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6593436/ – cognitive modelling motivation, simulator code of the project.

Instructor (name and email): Alex Aushev (alexander.aushev@aalto.fi)
Topic available: yes
Topic available also for a group: no

Topic #13: Bayesian optimisation for calibrating agent-based simulation model on urban mobility

Background: Bayesian optimisation is an iterative surrogate-based global optimisation approach that does not consider any assumptions on the functional form of a given optimisation problem, making it suitable for black-box functions. It comprises a surrogate model that maps the data to a utility, which guide the search through the objective space. As a consequence, bayesian optimisation is suitable for optimisation problems with expensive evaluation processes (in terms of resources like computation time, budget, etc), which highly correspond to agent-based simulation models on urban mobility. Calibration of such models for a city or region involves tuning from couples of hundreds to thousands parameters, along with computationally expensive simulations. Thus, number of simulations is the bottleneck towards convergence in global optima, and requires selection of both well-informative surrogate model and utility function that efficiently manage the exploration-exploitation trade-off.

The scope of the proposed research topic aims at implementation of efficient calibration framework encompassing different surrogate models and utility functions that will support further research on coping with high-dimensional objective spaces. Depending on resources (primarily time-related), the scope may touch the topic of few-shot learning, as well. Along with implementation itself, the work may extend towards reproduction/reimplementation of studies performed in the domain of urban mobility, using available data from different cities or regions.

Prerequisite:

- Programming skills and experience with R(preferable) or Python
- Basic knowledge in mathematical optimisation, machine learning and probability (bayesian paradigm)

Prerequisite:

- https://arxiv.org/abs/1807.02811
- https://arxiv.org/abs/1810.03688
- https://www.sciencedirect.com/science/article/pii/S0167739X1830150X
- https://arxiv.org/abs/1902.10675

Instructor (name and email): Vladimir Kuzmanovski <vladimir.kuzmanovski@aalto.fi>, Jaakko Hollmén <jaakko.hollmen@aalto.fi>

Topic available: yes_one_instance

Topic available also for a group: yes (max 2)

Topic #14: Importance sampling for leave-one-group-out cross-validation

Background:

Leave-one-out cross-validation (loo-cv) is a method for assessing a statistical model on independent data by reusing existing data. With importance sampling, loo-cv can be computed without needing to repeatedly refit the model to different partitions of the data. When the data can be divided into known groups, such as countries, assessing the model on unseen observations from unseen groups requires partitioning the data by leaving out a group of observations at a time. This is called leave-one-group-out cross-validation (logo-cv) and is a more challenging task for importance sampling. To make this task easier, it is possible use an adaptive importance sampling algorithm called importance weighted moment matching (IWMM) (Paananen et al. 2020). The goal of this project is to evaluate the applicability of importance sampling and IWMM to leave-one-group-out cross-validation in different Bayesian models.

Prerequisite:

- Programming skills with R
- knowledge of Stan, Bayesian statistics and modelling

References:

- https://link.springer.com/article/10.1007/s11222-016-9696-4
- https://arxiv.org/abs/1906.08850

Instructor (name and email): Topi Paananen (topi.paananen@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #15: Accurate computation of Monte Carlo posterior summaries

Background:

In Bayesian inference, it is common to generate a Monte Carlo sample from the posterior distribution of a probabilistic model. Much of Bayesian statistics involves computing expectations of different functions over the posterior distribution of the model. A simple but important example of such expectation is summarising the posterior using statistics such as mean and variance. If the posterior distribution has heavy tails, simple Monte Carlo expectations may be inaccurate. Importance weighted moment matching (IWMM) is an algorithm that uses importance sampling to make computation of Monte Carlo expectation values more accurate (Paananen et al. 2020). The goal of this project is to use IWMM for computing posterior summaries more accurately and test its applicability in different Bayesian models.

Prerequisite:

- Programming skills with R
- knowledge of Stan, Bayesian statistics and modelling

References:

- https://arxiv.org/abs/1906.08850

Instructor (name and email): Topi Paananen (topi.paananen@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #16: Using automatic differentiation for nonlinear variable selection

Background:

Evaluating the importance of predictor variables and their interactions is a common task in supervised machine learning that helps to gain knowledge about the used model or data. Model-agnostic methods perform this evaluation based on the predictions of the model. KL-diff is a method that uses the model's predictive distribution instead of point predictions to improve the accuracy of the evaluation (Paananen et al. 2019). Its drawback is that is requires hand-calculating derivatives that depend on the structure of the used model. Automatic differentiation is a set of techniques for numerically evaluating the derivatives of functions. The goal of this project is to familiarise with automatic differentiation and implement it for KL-diff.

Prerequisite:

- Programming skills with Python
- knowledge of machine learning, Bayesian statistics, and Gaussian processes

References:

- https://www.jmlr.org/papers/volume18/17-468/17-468.pdf
- https://arxiv.org/abs/1910.07942

Instructor (name and email): Topi Paananen (topi.paananen@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #17: Adaptive importance sampling transformations

Background:

Importance sampling is a technique for computing Monte Carlo expectations over a specific distribution using a sample from a different distribution. Adaptive importance sampling is a class of techniques for improving the accuracy of importance sampling by adapting the proposal distribution according to some target. Importance weighted moment matching (IWMM) is an adaptive importance sampling algorithm that adapts the proposal distribution by matching its moments to weighted moments based on the target (Paananen et al. 2020). The goal of this project is to implement new transformations for IWMM to make the moment matching more efficient. You can test the efficiency of the transformations with importance sampling cross-validation using Bayesian models.

Prerequisite:

- Programming skills with R
- knowledge of Stan, Bayesian statistics and modelling

References:

- https://ieeexplore.ieee.org/document/7974876
- https://arxiv.org/abs/1906.08850

Instructor (name and email): Topi Paananen (topi.paananen@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #18: Quantifying uncertainty in deep learning

Background:

How to estimate the predictive uncertainty of deep learning models, i.e., provide meaningful confidence values in addition to the label predictions, is drawing more attention. Predictive uncertainty is increasingly being used to make decisions in important applications such as medical decision support, self-driving cares, and financial forecasts. There have been various approaches proposed to quantify uncertainty in deep learning ranging from frequentist approaches to Bayesian methods. In this project, the student will review the current literature about quantifying prediction uncertainty in DL and present comparison studies on real-world datasets.

Prerequisite:

- Deep learning and statistics

References:

- https://arxiv.org/abs/2007.13481
- https://openreview.net/forum?id=BJxI5gHKDr
- https://arxiv.org/abs/1906.02530

Instructor (name and email): Tianyu Cui (tianyu.cui@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #19: Automatic Machine Learning Systems

Background:

Machine learning is becoming an important tool in our daily life. To simplify the use of machine learning and reduce human effort, automated machine learning has emerged as a potential tool that alleviates the burden for both users and researchers. In addition, exploiting parallelism for effectively executing automatic machine learning (AutoML) is also necessary for such breakthroughs and others, going hand in hand with the advancement of deep learning as a field.

In this topic, students will study AutoML methods according to the pipeline consisting of data preparation, feature engineering, hyperparameter optimization, and algorithm search. As a result, the student will give a comprehensive comparison for existing AutoML libraries, for instance, Auto-sklearn, MLBox, TPOT, H20 AutoML, etc. In case students use common ML algorithms, it is recommended students use the datasets from OpenML datasets (https://www.openml.org/s/88/data). For students interested in deep learning, the library Auto-Keras is recommended and the datasets should be used for this topic are at http://deeplearning.net/datasets/. It is encouraged that the student can introduce the solutions for improving the AutoML systems in terms of performance, accuracy, etc.

Further information:

- https://arxiv.org/abs/2007.04074
- https://arxiv.org/abs/1801.06007
- https://arxiv.org/pdf/1808.06492.pdf
- https://arxiv.org/abs/1806.10282
- https://bayesopt.github.io/papers/2017/22.pdf

Prerequisites:

- Machine learning and deep learning knowledge
- Bonus: Graph Theory.

Instructor: Pham Thanh Phuong (phuong.pham@aalto.fi)

Co-instructor: Assoc. Prof. Linh Truong (linh.truong@aalto.fi)

AaltoSEA, https://rdsea.github.io

Language: English

Topic available: Yes

Topic available also for a group: Yes (max 2)

Topic #20: Machine Learning approaches for detecting COVID-19 related misinformation on Social Media: Literature Review

Background: The recent outbreak of the COVID-19 pandemic has been accompanied by an exponential outburst of misinformation and fake news. In particular, the overabundance of information on social media platforms about the pandemic (also referred as Infodemic), makes it extremely hard for people to find sources and guidance that is trustworthy and reliable. An active stream of research applies various machine learning techniques to address issues linked to COVID-19 related misinformation and fake news. The current work will specifically review the existing body of knowledge around COVID-19 related misinformation on different social media platforms. The focus will be to synthesize the methodological approaches and offer insights into the most promising ones. The overview of various machine learning algorithms will further delve into benchmark analysis of these algorithms, e.g. accuracy, robustness, speed, and effectiveness.

Further information:

Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., ... & Liang, B. (2020). Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. JMIR public health and surveillance, 6 (3), e20794.

Huang, B., & Carley, K. M. (2020). Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak. *arXiv preprint arXiv:2006.04278*.

Boukouvalas, Z., Mallinson, C., Crothers, E., Japkowicz, N., Piplai, A., Sudip, M., ... & Adal, T. (2020). Independent Component Analysis for Trustworthy Cyberspace during High Impact Events: An Application to Covid-19. *arXiv preprint arXiv:2006.01284*.

Prerequisite: Basic knowledge of machine learning and data science.

Instructor (name and email): Aqdas Malik (aqdas.a.malik@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 3)

Topic #21: Predictive maintenance

Background: The number of errors in modern systems (such as web services, factories, networks) is significant, and it becomes harder and harder to maintain those systems without any tools. But many problems could be predicted in advance. Predictive maintenance algorithms are used to predict those breaks and errors. Predictive maintenance is a new and exciting field where different branches of Artificial Intelligence are involved: recurrent neural networks, Gaussian processes, graph neural networks, etc. In this project, you will review SOTA methods of the field and implement one of them. This project can be adjusted for the student’s research preferences: theoretical machine learning research or applications to one of the real-life problems.

References:

https://arxiv.org/pdf/1912.07383.pdf

https://ieeexplore.ieee.org/document/8454498

https://www.sciencedirect.com/science/article/abs/pii/S0378775319315848

Prerequisite: Basic understanding of machine learning and mathematics, good knowledge of Python (or Julia, or R, or MatLab).

Instructor (name and email): Alexander Nikitin (firstname.lastname@aalto.fi)

Topic available: yes_many_instances

Topic available also for a group: yes (max 3)

Topic #22: Auto time series forecasting for the regression problems

Background: Modern machine learning tools become more straightforward for users without any background knowledge. AutoML techniques make it possible to automatically construct the model (choose the model type, architecture, hyperparameters). The goal of this project is to review existing methods for time series regression problems and propose a way to automatically construct solutions for any particular dataset.

References:

https://www.kaggle.com/c/m5-forecasting-accuracy/overview

https://autodl.lri.fr/competitions/163

Prerequisite: Basic understanding of machine learning and mathematics, good knowledge of Python (or Julia, or R, or MatLab).

Instructor (name and email): Alexander Nikitin (firstname.lastname@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 3)

Topic #23: Concrete problems in AI safety

Background: The rapid improvement in RL performance over the last two decades has lead to an increase in interest in AI safety. One of the question AI safety tries to answer is how we can make sure that the values of an AI will always be aligned with out own. You can read about some of the concrete problems the field of AI safety tries to tackle in the first reference below. In a recent publication (see second reference below) a team at DeepMind has introduced gridworlds which yield unsafe behaviour in some of the RL algorithms we use every day. For this project you will choose one or more of these concrete problems to study. You will identify where current algorithms go wrong, survey proposed mitigations and -- time permitting -- implement one of these mitigations to see how well it works in practice.

References:

https://arxiv.org/pdf/1606.06565.pdf

https://arxiv.org/pdf/1711.09883.pdf

Prerequisite: Basic knowledge of RL.

Instructor (name and email): Sebastiaan De Peuter (sebastiaan.depeuter@aalto.fi)

Topic available: yes_many_instances

Topic available also for a group: yes (max 2)

Topic #24: Contextual Bandit Algorithms

Background: Bandit problems are stateless decision problems where you are given a set of options (arms) to choose from but are not told how good each option is. In every iteration you choose an option and then receive a reward for that option. Over time you can use these rewards to estimate the value of every option. The goal in bandit problems is to maximize the sum of rewards over time, this requires trading off exploration (trying out options to observe their reward) and exploitation (using the best known option currently to get guaranteed high reward). This project will be about contextual bandits, which is a variant of the original problem where at every iteration you are given a "context", some additional information to make your choice, which also determines the reward of the different options. An example of a contextual bandit problem is an assistance problem where at every iteration you have to help a user whose type is given by the context. Different types of users require different kinds of assistance so the value of your assistance options depends on the user type.

In this project you will survey contextual bandit algorithms and apply them to a real problem. The project is flexible. If you don't currently have a good background in RL or bandit algorithms you can do a literature survey, supplemented with an application if time permits. If you have more background or are a quick learner then we have a real project related to the example from above which you can tackle.

References:

https://tor-lattimore.com/downloads/book/book.pdf

Prerequisite: Basic knowledge of RL.

Instructor (name and email): Sebastiaan De Peuter (sebastiaan.depeuter@aalto.fi)

Topic available: yes_many_instances

Topic available also for a group: no

Topic #25: Efficient-communication in federated learning

Background: Federated learning is an emerging machine learning field that comes to utilize the abundant diversified and naturally disperse data sources, using their resources in a safe way, without moving the data from its source to a datacenter and doing the computations locally. While providing solutions for high latency, privacy, and lower power consumption, it still suffers from challenges for heterogeneity of the data, and the intrinsic expensive communications that come from it. We would like to do a literature survey in the efficient-communication federated learning algorithms up until now and implement at least one of them on a real-life dataset.

Prerequisite: Machine learning, PyTorch.

Instructor (name and email): Khaoula El Mekkaoui (khaoula.elmekkaoui@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #26: Multilabel Classification for Music and Audio using Deep Learning

Background:

In this project, you will work with audio and music to develop multi-label classification systems that are robust to acoustically challenging conditions. Some applications include audio tagging, music recommendation, sound event detection, music genre classification, and other audio tasks. These tasks are challenging due to the large size of data, and human perception of sound which is influenced by psycho-acoustics.

The goal of the project is to explore different deep learning architectures (CNNs, RNNs, Wavenet, Capsule nets, etc), as well as data augmentation techniques, that improve the performance of specific tasks for acoustically challenging scenarios. These scenarios include poor quality loudspeakers (e.g. mobile phones) or highly reverberant rooms. The methods will be evaluated on common datasets such uch as MSD, FMA, Jamendo or Freesound (among others). Ideally, the solutions developed will beside to participate in competitions such as DCASE, MediaEval, or MIREX.

This is a vast project with many possible activities for enthusiastic students. Some activities could include literature search, coding of different parts (pre-processing, signal processing, deep learning, evaluation, auralization, acoustic simulations), or running experiments and collecting results.

Potential readings:

Purwins, H., B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath. **“Deep Learning for Audio Signal Processing.”** https://doi.org/10.1109/JSTSP.2019.2908700.

Nam, J., K. Choi, J. Lee, S. Chou, and Y. Yang. **“Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach.”** https://doi.org/10.1109/MSP.2018.2874383.

Choi, Keunwoo, György Fazekas, Kyunghyun Cho, and Mark Sandler. **“A Tutorial on Deep Learning for Music Information Retrieval.”** http://arxiv.org/abs/1709.04396.

Pons, Jordi, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. **“End-to-End Learning for Music Audio Tagging at Scale.”** http://arxiv.org/abs/1711.02520.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. **“WaveNet: A Generative Model for Raw Audio.”** http://arxiv.org/abs/1609.03499.

Lee, J.; Park, J.; Kim, K.L.; Nam, J. **SampleCNN: End-to-End Deep Convolutional Neural Networks Using Very Small Filters for Music Classification**. Appl. Sci. 2018, 8, 150. https://www.mdpi.com/2076-3417/8/1/150#cite

T. Kim, J. Lee and J. Nam, "**Comparison and Analysis of SampleCNN Architectures for Audio Classification**," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285-297, May 2019, doi: 10.1109/JSTSP.2019.2909479. https://ieeexplore.ieee.org/document/8681654

Cite as: Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., Le, Q.V. (2019) S**pecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition**. Proc. Interspeech 2019, 2613-2617, DOI: 10.21437/Interspeech.2019-2680. https://arxiv.org/abs/1904.08779

Prerequisite:

Good knowledge of deep learning, programming skills in python.
Bonus: Experience with PyTorch.
Bonus: Familiarity with audio signal processing concepts such as filtering, time-frequency transforms, impulse responses, sampling frequency, pitch, etc ...

Instructor (name and email): Ricardo, Falcon Perez (ricardo.falconperez@aalto.fi)

Topic presented by: Ricardo, Falcon Perez

Topic available: yes_many_instances

Topic available also for a group: yes (max 2)

Topic #27: Model-based Reinforcement Learning

Background:

Within Reinforcement learning we aim to choose the optimal actions to achieve a task, given knowledge obtained from previous interactions with the system. When only a small number of observations can be feasibly observed, a model of the system dynamics drastically improves sample efficiency. Whilst this has been shown to be extremely sample-efficient, a number of challenges are present when trying to effectively use this model for control.

The goal of this project is to explore Bayesian approaches for model-based Reinforcement Learning. Benefits of Bayesian approaches include incorporation of model uncertainty in the exploration-exploitation trade-off, and as a method of incorporating prior information.

You will work as part of the PML research group, which has many complementary research projects in model-based Reinforcement Learning. Link: https://research.cs.aalto.fi/pml

Prerequisite: Machine learning, Statistics, Python.

Instructor (name and email): Charles Gadd (charles.gadd@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: no

Topic #28: Model agnostic fast and robust leave-one-out cross-validation in Stan probabilistic programming framework

Background: Cross-Validation (CV) is a general-purpose method to measure out-of-sample predictive performance. Unfortunately, CV is very computationally expensive especially for Bayesian models, at models have to be fit multiple times to different subsets of the data. In the last couple of years, we have developed methods to performace approximate leave-one-out cross-valdiation (LOO-CV) in a way that does not require refitting the model at all, if the methods work out. When the approximation does not work for certain observations, one option is to perform exact LOO-CV only for these problematic observations and to amend the approximate LOO-CV object accordingly. We have implemented this procedure in our Stan related R packages brms and rstanarm but see it as beneficial to implement a model-agnostic version in our R package loo, which can then be used by other packages without duplicating substantial amounts of code. The goal of this project would be to implement this model-agnostic version.

Prerequisite: R programming experience and basic understanding of Bayesian statistics

Instructor (name and email): Aki Vehtari (aki.vehtari@aalto.fi)

Topic available: yes_one_instance

Topic available also for a group: yes (max 3)

Topic #29: Probabilistic programming frameworks - what is the state-of-the-art?

Background: We have seen a surge in probabilistic programming frameworks (PPF) for making flexible Bayesian inference at scale during the last years. Examples of PPF that have been proposed are Stan, TensorFlow probability (TFP), Pyro, PyMC3, and Turing. The large number of PPFs has led to increased discussions of different probabilistic programming frameworks' quality and speed.

Recently, posteriordb, a database with multiple probabilistic models, data, and reference posteriors, has been developed to simplify inference research using Stan. The next step is to extend the database with models from other probabilistic programming frameworks such as TensorFlow probability (TFP), Pyro, and PyMC3 and compare performance.

The project consists of three parts:
1) Implement a test structure for assessing that models from different PPF are identical.
2) Include a set of models for TFP, Pyro, and PyMC3
3) Compare the other PPFs quality and speed for the implemented collection of models with Stan.

References:
https://github.com/MansMeg/posteriordb
https://www.tensorflow.org/probability
https://docs.pymc.io/
https://pyro.ai/
https://mc-stan.org/

Prerequisite: Knowledge of Python, Bayesian statistics, and modeling. Experience with Stan, TensorFlow, Pyro, or PyMC3 is a benefit.

Instructor (name and email): Aki Vehtari (aki.vehtari@aalto.fi)

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 3)

Topic #30: Federated Learning from Big Data over Networks

Background: We have recently proposed networked exponential families as a powerful modeling paradigm for massive networked-structure data (big data over networks. Such big data over networks arise in the management of pandemics where the raw data stored in the smartphones of individuals is structured according to different networks such as social networks or contact networks. This project revolves around the application of distributed optimization methods in networked exponential families. The resulting federated learning algorithms are analyzed from different angles such as robustness and privacy protection.

References:

Clustered Federated Learning: https://arxiv.org/abs/1910.01991

Network Lasso for Linear Models: https://arxiv.org/abs/1903.11178 https://arxiv.org/abs/1805.02483

Prerequisite: Good grip on linear algebra and convex optimization. Some experience in a scientific programming framework such as Python or Matlab.

Instructor (name and email): Alex Jung (alex.jung@aalto.fi)

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 3)

Topic #31: Acoustic scene classification

Background: Sounds carry a large amount of information about our everyday environment and physical events that take place in it. We can perceive the sound scene we are within (busy street, office, etc.), and recognize individual sound sources (car passing by, footsteps, etc.). Developing signal processing methods to automatically extract this information has huge potential in several applications, for example searching for multimedia based on its audio content, making context-aware mobile devices, robots, cars etc., and intelligent monitoring systems to recognize activities in their environments using acoustic information. However, a significant amount of research is still needed to reliably recognize sound scenes and individual sound sources in realistic soundscapes, where multiple sounds are present, often simultaneously, and distorted by the environment.

References:

DCASE2020 Challenge, IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events http://dcase.community/challenge2020/

Prerequisite: Some experience with speech or audio processing.Some experience in a scientific programming framework such as Python or Matlab.

Instructor (name and email): Tom Bäckström (tom.backstrom@aalto.fi) and Pablo Pérez Zarazaga (pablo.perezzarazaga@aalto.fi)

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 3)

Topic #32: Title Deep learning in speech and language processing

Background: Deep learning is changing the ways how speech and language data can be processed and represented. Several specific topics are available either for experimenting with new model architectures in real-word data or applications, such as automatic speech and speaker recognition, translation and language learning. The topic can be selected together with the student.

Prerequisite: One of Aalto's basic course in speech recognition or natural language processing or corresponding knowledge. Knowledge in deep learning. Experience in scientific programming, e.g. in Python**.**

Instructor (name and email): Mikko Kurimo mikko.kurimo@aalto.fi and his research group

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 3)

Topic #33: Speech-based biomarking of the human health state with machine learning

Background: Speech signal provides an attractive means to predict the human health state. Increasing research interest is devoted particularly to detect neurodegenerative diseases, such as Parkinson's disease and Alzheimer's disease, from speech signals using both classical ML methods (such as SVMs) and more recent deep learning methods. Specific topics are provided in this health -related research area at the Department of Signal Processing and Acoustics.

Prerequisite: Knowledge in machine learning and deep learning. Basic knowledge in speech processing is added value. Experience in scientific programming**.**

Instructor (name and email): Paavo Alku (paavo.alku@aalto.fi), Kiran Reddy (kiran.r.mittapalle@aalto.fi), Sudarsana Kadiri (sudarsana.kadiri@aalto.fi)

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 2)

Topic #34: Implementing deep learning in processing and analysis of multiple biosignals.

Background: Physical examinations performed by doctors and nurses in the primary health care are often the first measurement based health assessment performed on patients. Depending on the cause of visit, temperature, blood pressure, oxygen saturation, pulse, respiratory rate and auscultation (listening to the heart and lungs) are performed by doctors and nurses on daily bases. Project Vestoscope at the department of Neurosciences and Biomedical Engineering (NBE) aims at automatising these examinations to save time and implementing algorithms and machine learning to improve the accuracy of the diagnosis.

Most challenging task of the project is the classification and analysis of digitalised breathing sounds, detected by multiple sensors.

Sample algorithms of the audio analysis process are available in Python to use as a model or build upon. SDKs and APIs of some of the sensors and the wireless transfer system also available.

Prerequisite: Knowledge of Python, machine learning and Deep Learning. Knowledge or experience in audio signal processing is a plus.

Instructor (name and email): Alexis Kouros (alexis.kouros@aalto.fi), Aki Laakso (aki.j.laakso@aalto.fi)

Topic available: yes_multiple_instances

Topic available also for a group: yes (max 3 students/group)