DS3 DATA SCIENCE SUMMER SCHOOL

JUNE 25-29 2018
AT CAMPUS POLYTECHNIQUE

IN-DEPTH TUTORIALS WITH PRACTICAL SESSIONS JUNE 28

1. Approximate Bayesian Inference: old and new (Part 1/2) - Emtiyaz KHAN [RIKEN]

Abstract:

Approximate Bayesian inference (ABI) offers many promising solutions to advance modern machine-learning methods such as deep learning and reinforcement learning. This tutorial will give an overview of old and new methods for ABI. We will start with motivating applications of ABI methods in modern machine learning, and discuss the computational challenges associated with them. We will review traditional methods (such as the Laplace approximation, Markov chain Monte Carlo, Expectation Propagation, and Variational Inference) as well as modern methods that are motivated by deep learning (e.g., stochastic-gradient variational inference and variational auto-encoders). Overall, the tutorial will aim to motivate and empower the audience to pursue research in ABI, as well as to apply the ABI methods to real-world problems.

Requirements:

Basic knowledge of Machine Learning and Programming in Python.

2. Causality and Machine Learning (1 day repeated on June 29) - Kun ZHANG [Carnegie Mellon University]

Abstract:

Does smoking cause cancer? Can we find the causal direction between two variables by analyzing their observed values? In our daily life and science, people often attempt to answer such causal questions, for the sake of understanding and manipulating systems properly. On the other hand, we are also often faced with problems of how to properly make use of causal knowledge for machine learning. For instance, how can we make optimal predictions in non-stationary environments? In the past decades, interesting advances were made in machine learning, statistics, and philosophy for tackling long-standing causality problems, such as how to discover causal knowledge from purely observational data and how to infer the effect of interventions using such data. Furthermore, recently it has been shown that causal information can facilitate understanding and solving various machine learning problems, including transfer learning and semi-supervised learning. This tutorial reviews essential concepts in causality studies and is focused on how to learn causal relations from observation data and why and how the causal perspective helps in machine learning and other tasks.

Requirements:

We encourage the usage of the free software package Tetrad, to ensure a unified working environment and maximize compatibility.

3. Crash Course In Deep Learning And Pytorch (1 day repeated on June 29) - Francisco MASSA [Facebook]

Abstract:

You shall learn the basics in deep learning with examples in pytorch.
After this workshop, you will have a basic understanding of convolutional networks, standard gradient based optimization methods, pytorch tensors, autograd, and deep-learning specific modules.

Requirements:

Knowledge of python programming
Basics of linear algebra and statistics
Environment : Python Jupyter
Packages: numpy, pytorch, torchvision, matplotlib.
PyTorch and torchvision wheels are available on http://pytorch.org

4. Introduction to Deep Learning with Keras (1 day repeated on June 29) - Olivier GRISEL [Inria]

Abstract:

This session will introduce the main deep learning concepts with worked examples using Keras. In particular, we will cover the following concepts:
- feed-forward fully connected network trained with stochastic gradient descent,
- convolution networks for image classification with transfer learning,
- embeddings (continuous vectors as a representation for symbolic/discrete variables such as words, tags…),
- if time allows: Recurrent Neural Networks for NLP.

Requirements:

- Working of Python programming with NumPy
- Basics of linear algebra and statistics
- Environment: Python Jupyter
- Packages: numpy, matplotlib, keras (2.1.6) with the tensorflow backend (tensorflow 1.5 or later).
- Follow the instructions here: https://github.com/m2dsupsdlclass/lectures-labs/blob/master/installation_instructions.md
- Optionally pytorch 0.4.0 or later for a short intro to pytorch at the end of the session if the audience requests it.

5. Introduction To Reinforcement Learning (Part 1/2) - Olivier PIETQUIN [Google Brain]

Abstract:

In this session session we will address the fundamentals of Reinforcement Learning. Reinforcement learning is the machine learning answer to sequential decision making and control. It has known an increasing interest since its recent success in learning to play Atari games from raw pixels or helping mastering the game of Go. We will first describe the underlying model of Markov Decision Processes and describe the fundamental principles of Dynamic Programming. From there, we will derive algorithms able to learn a control policy through interactions with their environment in the case of discrete state and actions spaces. The second day, we will present methods allowing to scale up reinforcement learning algorithms so as to address continuous state and action spaces and real world problems. This will lead us all the way to deep Reinforcement Learning and its applications to video games, robotics and Go.

Requirements:

Python coding
Have the packages Numpy, matplotlib et OpenAI Gym installed

6. Machine Learning For Genetic Data (1 day only) - Chloé-Agathe AZENCOTT – Jean-Philippe VERT – Thomas WALTER [CBIO]

Abstract:

Many complex traits of living organisms, from plant growth to resistance to infection, bone density, or disease risk, are driven in part by genetic factors. Technical advances make it possible to accumulate data sets in which tens or hundreds of thousands of molecular features (such as gene expression, methylation status, copy number variation or single point mutations) have been measured for thousands of individuals or more. Applying machine learning techniques to such data poses several challenges: the number of features is high, usually greater than the sample size; these features are noisy and correlated; and a large emphasis is put on the interpretability of the models. In this session, participants will explore the behavior of classical machine learning tools on high-dimensional genetic data and discover how regularization can help build interpretable models to further our understanding of complex genetic traits.

Requirements:

scikit-learn+jupyter

More information on: **www.ds3-datascience-polytechnique.fr**

DS3 DATA SCIENCE SUMMER SCHOOL

JUNE 25-29 2018 AT CAMPUS POLYTECHNIQUE

IN-DEPTH TUTORIALS WITH PRACTICAL SESSIONS - JUNE 28

7. Missing Data Imputation (1 day only) - Julie JOSSE [Ecole polytechnique]

Abstract:

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst. In this tutorial, we give an overview of the missing values literature (EM algorithm, imputation based on SVD, imputation with random forests, multiple imputation, ...) as well as the recent improvements that caught the attention of the community due to their ability to handle large matrices with large amount of missing entries. We will illustrate the methods on medical, environmental and survey data.

Requirements:

Knowledge in PCA and in linear regression.
Environment : R and Rstudio
Packages: missMDA, missForest, Amelia, mice, naniar, VIM, norm

8. Optimal Transport And Machine Learning (1 day repeated on June 29) - Marco CUTURI [ENSAE] – Nicolas COURTY [Irisa] – Rémi FLAMARY [University of Nice]

Abstract:

Optimal transport (OT) provides a powerful and flexible way to compare probability measures, of all shapes: absolutely continuous, degenerate, or discrete. This includes of course point clouds, histograms of features, and more generally datasets, parametric densities or generative models. Originally proposed by Monge in the eighteenth century, this theory later led to Nobel Prizes for Koopmans and Kantorovich as well as Villani's Fields Medal in 2010.

After having attracted the interest of mathematicians for several years, OT has recently reached the machine learning community, because it can now tackle (both in theory and numerically) challenging learning scenarios, including for instance dimensionality reduction and structured prediction problems that involve histograms or point clouds, and estimation of parametric densities or generative models in highly degenerate / high-dimensional problems.

We will present in this course a brief introduction to all the important elements needed to grasp this new tool, with an emphasis on algorithmics (LP and regularized formulations) as well as applications (barycenters, distance between texts, topic models, generative models)

Requirements:

– Python/numpy/matplotlib with jupyter notebook or spyder
– POT Python optimal transport toolbox (easy install through anaconda or pip)

9. Representing and Comparing Probabilities with Kernels (1 day repeated on June 29) - Arthur GRETTON [University College London]

Abstract:

We provide an introduction to kernel distribution embeddings, with a focus on practical applications in hypothesis testing and machine learning. The first part of the talk will focus entirely on the representation of probability distributions. The emphasis will be on designing kernels or features to make two distributions as distinguishable as possible. Applications include two-sample testing (of whether two samples are from the same distribution), and critic design for generative adversarial networks (where a neural network is trained to generate samples that match a target distribution). The second part of the talk will focus on more sophisticated applications of distribution representations: model criticism (using Stein's method to test against a parametric model), and testing for statistical dependence between two signals, even when no explicit mapping between the signals is known.

The practical sessions will cover examples how to embed the presented tests into scientific workflows. For this, participants will first learn fundamental implementation details of all major classes of kernel hypothesis tests. This includes including implementation of the test statistics, computing test thresholds (via sampling and parametric approximations), and scalability concerns (quadratic vs linear time, efficient implementations). Next, we will demonstrate how to tune critical (kernehyper-parameters of the tests, for increased test power and for feature selection. The session will conclude with applications of all test classes on a number of real-world examples.

Requirements:

Computer with working Python installation + scipy stack (numpy, scipy, matplotlib, jupyter notebook, ...) + TensorFlow. Web-browser and internet access for course material and additional software.

10. Submodularity In Data Science (Part 1/2) - Andreas KRAUSE [ETH Zürich]

Abstract:

Many problems in Machine Learning and Data Science require solving large-scale discrete optimization problems under uncertainty. In the recent years, a fundamental problem structure has emerged as extremely useful for addressing such problems: Submodularity is an intuitive diminishing returns property, closely related to convexity, with similarly important implications for optimisation. Submodularity naturally occurs in numerous applications such as data summarisation, information retrieval, network analysis, active learning and experimental design, sparse modelling and inference in probabilistic models. Exploiting submodularity allows to devise efficient algorithms with strong theoretical guarantees. In this tutorial, I will give an introduction to the concept of submodularity and discuss basic algorithms for minimising and maximising submodular functions. I will
also discuss current research directions, such as tackling sequential decision problems via online and adaptive submodular optimisation, learning submodular functions from data, and solving large-scale submodular optimisation problems in distributed and streaming settings, as well as applications in probabilistic inference and deep learning. A particular emphasis is on demonstrating their usefulness in solving concrete data science problems.

Requirements:

- basic background in discrete algorithms, probability, linear algebra and Python programming as covered, e.g., in undergraduate computer science curricula. Familiarity with basic concepts in machine learning is very helpful but not required.
- Laptop with Python environment (including numpy, scipy, matplotlib, jupyter notebook)

More information on: **www.ds3-datascience-polytechnique.fr**

DS3 DATA SCIENCE SUMMER SCHOOL

JUNE 25-29 2018 AT CAMPUS POLYTECHNIQUE

IN-DEPTH TUTORIALS WITH PRACTICAL SESSIONS - JUNE 28

11. Text as Data in the Social Sciences (1 day repeated on June 29) - Brandon STEWART [Princeton University]

Abstract:

The evidence base of the social sciences is rapidly expanding. Text is becoming an increasingly important part of the social scientist's toolkit. In this tutorial we cover the role for computer-assisted text analysis in four basic social science tasks: discovery, measurement, causal inference and prediction. We will cover a variety of different techniques for analyzing text and illustrate with examples from social science research. Although the tutorial will cover practical implementation details for a few approaches, we will focus on the social science research process and how we can make inferences from language.

Requirements:

R and an IDE such as RStudio

12. Topological Algorithms for Data Skeletonisation in Python (1 day only) - Vitaliy KURLIN [University of Liverpool]

Abstract:

Hour 1. A lecture with slides: point clouds as metric spaces; a review of clustering; a minimum spanning tree; MST-based clusterings, e.g. the single-edge clustering.
Hour 2. Demos and practical exercises on a Minimum Spanning Tree.
Hour 3. A lecture with slides: representing a point cloud by an abstract graph, the graph reconstruction problem; the Mapper algorithm.
Hour 4. Demos and practical exercises on the Mapper algorithm.
Hour 5. A lecture with slides: a Delaunay triangulation of a point cloud in the plane, 1-dimensional persistence for the filtration of alpha-complexes, a Homologically Persistent Skeleton.
Hour 6. Demos and practical exercises on a Homologically Persistent Skeleton.

Requirements:

Basic programming knowledge in Python

More information on: **www.ds3-datascience-polytechnique.fr**

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

DATA SCIENCE INITIATIVE ÉCOLE POLYTECHNIQUE

INSTITUTE DATAIA Data Science, Intelligence & Society

DS3 DATA SCIENCE SUMMER SCHOOL

JUNE 25-29 2018
AT CAMPUS POLYTECHNIQUE

IN-DEPTH TUTORIALS WITH PRACTICAL SESSIONS - JUNE 29

1. Approximate Bayesian Inference: old and new (Part 2/2) - **Emtiyaz KHAN [RIKEN]**

2. Causality and Machine Learning (same session as June 28) - **Kun ZHANG [Carnegie Mellon University]**

3. Crash Course In Deep Learning And Pytorch (same session as June 28) - **Francisco MASSA [Facebook]**

4. Introduction to Deep Learning with Keras (same session as June 28) - **Olivier GRISEL [Inria]**

5. Introduction To Reinforcement Learning (Part 2/2) - **Olivier PIETQUIN [Google Brain]**

6. Machine Learning For Biomedical Images (1 day only) - **Chloé-Agathe AZENCOTT – Jean-Philippe VERT – Thomas WALTER [CBIO]**

Abstract:
Next generation sequencing techniques and methodological advances in the field of bioinformatics have revolutionized modern biology and have given us unprecedented insights into the molecular basis of life. These efforts have been paralleled by advances in large-scale imaging, allowing us to study spatial arrangements of cells, compartments and molecules, to analyze intact biological specimen with respect to their morphological phenotypes and to perform analyses at different scales of organization, such as the cellular, tissular and organism scale.
Such systematic imaging approaches generate extremely large and complex image data sets. For instance, in High Content Screening (HCS), a large number – typically tens or hundreds of thousands – of different experimental conditions can be tested with respect to their effect on cells and organisms, as measured by microscopy. Another example is histopathology, where we analyze stained tumor sections typically containing hundreds of thousands of cells.
In this session, I will first give an overview over different applications in the field of bioimage informatics, the counterpart of bioinformatics. I will then show in detail applications in computational phenotyping, in particular the analysis of cellular morphologies and the spatial aspects of gene expression. Finally, I will show recent trends in the field, with a strong focus on deep learning approaches.
Requirements:
scikit-learn+jupyter + scikit-image

7. Optimal Transport And Machine Learning (same session as June 28) - **Marco CUTURI [ENSAE] – Nicolas COURTY [Irisa] – Rémi FLAMARY [University of Nice]**

8. Recommendation in the Real World (1 day only) - **Olivier KOCH – Flavian VASILE [Criteo]**

Abstract:
In this course we will cover a variety of machine learning-based methods for recommendation, ranging from classical approaches to modern Deep Learning-based techniques. The focus of the course in on real-world recommendation and the two big resulting questions:
- How to scale Recommender Systems both in space and time (how to make them work for large sets of items and user profiles and how to make them be able to take into account real-time user activity signals) and
- How to mitigate the inherent discrepancy between offline technical metrics and the actual online performance.
The course will be divided in five parts:
- In part 1 we will offer a quick introduction to the field of Recommendation, offer examples of state-of-the-art Recommender Systems in the wild and review the main ML approaches powering them.
- In part 2 we will review classical ML approaches for Recommendation starting with Collaborative Filtering and continuing with Matrix Factorization, Content-Based Recommendation and Hybrid Solutions
- In part 3 we will go over modern Deep Learning based models and introduce RNN-based user modelling for Recommendation
- In part 4 we will go over one of the most promising upcoming change in Recommendation, which is Causal Recommendation, that merges ideas from Machine Learning with ideas from Causal Inference and go over cutting-edge emerging techniques
- In part 5 we will wrap-up with conclusions and have Q&A sessions with the participants in the course.
All theoretical parts (2-4) will have a practical session at the end, where the participants will be able to get hands-on experience on the methods introduces in the theoretical section.
Requirements:
Hardware: laptop
Software:
jupyter notebooks
python (version TBD)
tensorflow, pytorch, matplotlib, pandas, numpy
we will provide a web page with install instructions and sanity checks
M1-level knowledge in linear algebra and stats

More information on: **www.ds3-datascience-polytechnique.fr**

ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY
DATA SCIENCE INITIATIVE ÉCOLE POLYTECHNIQUE
INSTITUTE DATAIA Data Science, Intelligence & Society

IN-DEPTH TUTORIALS WITH PRACTICAL SESSIONS - JUNE 29

9. Representing and Comparing Probabilities with Kernels (same session as June 28) - **Arthur GRETTON [University College London]**

10. Submodularity In Data Science (Part 2/2) - **Andreas KRAUSE [ETH Zürich]**

11. Text as Data in the Social Sciences (same session as June 28) - **Brandon STEWART [Princeton University]**

12. Topological Data Analysis (1 day only) - **Pawel DLOTKO [Swansea University], Vincent ROUVREAU [Inria]**

Abstract:

Hour 1. Followup and generalization of simplicial complexes. Recovering shapes from point clouds. Rips, Cech and Alpha complexes and the comparisons of those constructions. Practical session on generating some basic point clouds (like circle, sphere, torus) and reconstruction of the sets with simplicial complexes using Gudhi (C++ or python version).
Hour 2. Euler characteristics, homology group, Betti numbers – definition and basic examples. Computations of Betti numbers of sets obtained in Hour 1.
Hour 3. Filtered complexes and persistent homology. Extending on the previous studies to build parameter-free descriptor of the data.
Hour 4. Example of cubical complexes obtained form image data, numerical simulation and point cloud data. Experiments with homology and persistent homology of cubical complexes (theory with practical experiments).
Hour 5. Persistence representations. How to tunnel persistent homology as an input for machine learning. Experiments with Gudhi.
Hour 6. If time permits, experiments with detection of (semi) periodicity using persistent homology.

Requirements:

Basic programming knowledge in Python