Astronomical Data Analysis Software and Systems XXX
ASP Conference Series, Vol. 532
J. E. Ruiz, F. Pierfederici, and P. Teuben, eds.

Efficient Machine Learning Methods for Cosmology

Eric Howard1, 2

[1]*Macquarie University, Sydney, NSW, 2000, Australia;* eric.howard@mq.edu.au

[2]*Griffith University, Brisbane, QLD, Australia*

Abstract. Machine learning-based analysis techniques are gaining interest in cosmology due to their computational ability to generate complex models in order to analyze and interpret large scale structure data sets, such as the matter density fields comprised of nonlinear complex features, like halos, filaments, sheets and voids. We present a number of powerful machine learning algorithms (classification, regression, reinforcement learning) and data-analysis tools that can be used to predict the non-perturbative cosmological structure and non-Gaussian features hierarchically formed over all scales in the Universe, justifying the advantage of employing such methods for use in cosmology. This paper focuses on explicitly analyzing the machine learning methods that can be applied to existing cosmological problems.

1. Introduction

Current cosmology research is experiencing a rapid increase in data volume and complexity. Data-driven cosmological discovery has seen a remarkable rise in the last decade, leading to unprecedented improvements in the ways we can gain knowledge and extract novel information. Nevertheless, the study of the evolution of the Universe at cosmological scales requires accurate observations of the sky and fast prediction of the structures in the Universe.

Machine learning introduces cosmology to the era of big data science, as a useful companion to common traditional tools and data analysis techniques, employed to facilitate new discoveries, interpret and extract new cosmological features from existing large datasets, constraining astrophysical (Howard 2017) and cosmological parameters and modelling the large structure formation of the Universe. Modern algorithms in machine learning and statistics can play an increasingly significant role in current cosmology research. As data sets become larger and more difficult to process, the cross-fertilization between cosmology and machine learning requires the integration of traditional statistical techniques with modern machine learning tools, providing promising opportunities with significant advantages for state-of-the art cosmological simulations. Multiple applications, from cosmic web simulations and predicting the cosmic structure in the non-linear regime to multi-wavelength structure identification, 21cm reionization models or predicting dark matter annihilation and halo formation may benefit from robust and efficient data analysis methods, such as convolutional neural networks or generative adversarial networks.

2. Machine learning as a cosmology tool

We here assess the success and challenges of employing machine learning methods and algorithms, in order to accurately estimate cosmological parameters and efficiently extract complex knowledge for use in cosmological models, that can significantly reduce the computational time for parameter estimation, greatly reduce the computational power required and ultimately taking advantage of parallelism. Modern cosmology is entering an era of massive data sets and deep, wide-field surveys, making necessary the use of novel automated techniques for data analysis and reduction.

Current machine learning and deep learning techniques have the capacity to distinguish between different cosmological and gravitational models using specific cosmological features as discriminants Pan et al. (2020) within very large volumes of data, in order to extract large scale information about the Universe and deeply probe the fundamental properties of cosmology. Firstly, the estimation of such discriminants from the large-scale cosmological structure (Pan et al. 2020) of our Universe uses statistical computations of the observed structures in galaxy surveys such as correlation functions or power-spectrum and compares with the theoretical models. Machine learning can here help in the process of finding stronger constraints and accurately estimating the cosmological parameters directly from the distribution of matter. Deep learning algorithms were used to infer the masses of galaxy clusters directly from images of the microwave sky, by determining the scaling relation between a cluster's Sunyaev-Zeldovich effect signal and its mass. Supervised regression algorithms can use multiple observables to predict key features for cosmological analysis. Current machine learning techniques, such as convolutional neural networks are a promising tool for the analysis of astrophysical data in parametric models, for image recognition and classification tasks, such as searching for strong lensing signatures in cosmic structures. Deep learning algorithms can be used in fast and efficient estimation of cosmological parameters, obtaining uncertainties for these parameters as well as classification tasks for strong gravitational lensing systems. Quantifying the image distortions due to strong gravitational lensing effects and analyzing the matter distribution in lensing galaxies generally employs maximum likelihood procedures, while fast and automated deep learning algorithms, such Independent Component Analysis of multi-filter imaging data can recover the model parameters in strong lensing systems with highly nonlinear image distortion.

Another current effective approach to predicting the large scale structure formation of the Universe are N-body simulations, that evolve the Universe from its birth at the Big Bang to the present day. Deep learning techniques currently provide an alternative to existing numerical simulations in cosmology and conventional higher-order statistics. Deep neural network algorithms were used for N-body-like simulations with positions and velocities. Deep generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders produce images of high visual quality and are a possible method to emulate such datasets in the vast amount of data collected by cosmological surveys and replacing computationally expensive simulations. Machine learning methods, in particular supervised regression algorithms, are a promising technique to handle multiple observables in order to predict a galaxy cluster's mass, or masses for large samples of clusters, which is a vital requirement for cosmological analysis with future surveys. Properties of galaxy clusters depend on the underlying cosmological model and also provide significant information for constraining cosmological parameters. Deep machine learning techniques can be used to optimally extract cosmological parameters using galaxy clusters detected from large-scale surveys.

Deep Neural Networks are capable of extracting non-Gaussian information from weak lensing data sets (Zorrilla Matilla et al. 2020), leading to a better understanding of the fundamental nature of cosmic acceleration. The weak gravitational lensing by the large-scale matter structure, also known as cosmic shear is an important cosmological probe for studying dark matter and dark energy. Machine learning methods for weak-lensing analysis of galaxy catalogs and convergence maps are employed to distinguish between different gravity models and generate statistically similar observations. Convolutional neural networks trained on simulated convergence maps can be used to classify different cosmological scenarios based on the statistically similar weak-lensing maps they generate, in terms of Gaussian weak-lensing observables.

Additionally, extragalactic distance measurements based on type Ia supernovae are the first evidence that the Universe is undergoing an accelerated expansion (Arjona & Nesseris 2020). Photometric supernovae classification is an essential ingredient for extracting cosmological constraints from type Ia supernovae. The mapping of the light curve shapes into different classes according to their spectroscopic features, is a challenging task that can be described as a supervised machine learning problem. In recent years, in a remarkable community effort, automated photometric supernova classification research, including extracting descriptive features from the light curves and classification tasks using machine learning algorithms, has recently become an active area in cosmology. For supernova cosmology, the big data challenge is approaching very fast, given the overwhelming volume of data of the massive photometric data sets to be delivered by current and future surveys, leading to the development of automated classification tools in modern cosmology. On the other hand, topological data analysis is currently used for summarizing the shape of data, in particular for distinguishing between different dark energy models using machine learning trained data sets, identifying structures of the cosmic web or new types of structures in the large-scale distribution of matter. Data exhibiting complex spatial structures with persistent homology can be difficult to analyze but topological Data Analysis offers a simple method to represent, visualize, and interpret data by extracting the topological features used to infer the dynamics and properties of the underlying structures. This method can be exploited to find clusters, cosmic voids and loops of filaments in cosmological datasets of dark matter halo catalogs and galaxy surveys and assess their statistical significance. Topological data analysis is a statistically rigorous technique useful for locating informative generators in large-scale cosmological datasets, providing further cosmological constraints on the sum of neutrino masses.

The Epoch of Reionization (EoR) also provides important cosmological information about the Universe's history characterized by large scale phase change of the neutral intergalactic medium ionized by the emergence of the first luminous sources. Reionization provides important information about the process of structure formation in the universe and the evolutionary connection between the smooth matter distribution at early times from by CMB studies, and the highly structures of galaxies and clusters at smaller redshifts. Deep learning techniques with convolutional neural networks predict the EoR history from the 21-cm differential brightness temperature tomography images using sliced-averaged neutral hydrogen fraction in a given 21-cm map. The rich interplay between the first luminous sources and the low-density gas of the intergalactic medium during the EoR contains key observational constraints in the 21 cm power spectrum such as the midpoint and duration of reionization that has to be extracted from the power spectrum. Machine learning methods such as convolutional neural networks are a good alternative method (Kwon et al. 2020) to help with extracting two-dimensional

information from images of reionization at a series of redshift values and generate similar image cubes to MWA and HERA. Deep learning methods show great potential to efficiently reconstruct the EoR evolution from the 21-cm tomography surveys in the near future. The 21 cm EoR field is expected to have a highly non-Gaussian character and machine learning approaches may help to extract non-Gaussian information present in the maps, sufficient to characterize many of the valuable features of the reionization history.

One of the major cornerstones in cosmology is the Cosmic Microwave Background (CMB) and the CMB temperature maps are a viable resource for cosmological analysis. Traditional computational models are employed to generate CMB temperature anisotropy maps but require a large amount of CMB data for analysis. Deep generative algorithms can be used generate synthetic samples of CMB all-sky maps which can be used for cosmological analysis. Neural networks such as GANs can be employed (Vafaei Sadr & Farsian 2020) to train deep generative models to learn the complex distribution of CMB maps and accurately generate new sets of CMB data as 2-dimensional patches of anisotropy maps. Multilayer perceptron model was also employed for estimating the baryon density from a CMB map and correlate the baryon density from the power spectrum of simulated CMB temperature maps with the map image, forming datasets for training the neural network model. In this scenario, the isotropy of the CMB can be analyzed by training the model with CMB maps for different galactic coordinates and compare with the results from neural network models.

3. Conclusion

Due to the rapid growth and dramatic increases of data sets in volume and complexity, as detectors, telescopes, and computers become more powerful, machine learning brings great potential for future data-driven discovery and address the statistical needs of the next generation of cosmological surveys. This opens the way to estimating the cosmological parameters of the Universe with higher accuracy than traditional data analysis techniques. For many theoretical models with a large number of parameters, the standard algorithms and methods of data processing and knowledge extraction can be computationally intensive and machine learning may extrapolate beyond the training data, outperforming the existing fast analytical approximations from standard techniques.

References

Arjona, R., & Nesseris, S. 2020, Phys.Rev.D, 101, 123525. 1910.01529

Howard, E. M. 2017, in Astronomical Data Analysis Software and Systems XXV, edited by N. P. F. Lorente, K. Shortridge, & R. Wayth, vol. 512 of Astronomical Society of the Pacific Conference Series, 245

Kwon, Y., Hong, S. E., & Park, I. 2020, Journal of Korean Physical Society, 77, 49. 2006.06236

Pan, S., Liu, M., Forero-Romero, J., Sabiu, C. G., Li, Z., Miao, H., & Li, X.-D. 2020, Science China Physics, Mechanics, and Astronomy, 63, 110412. 1908.10590

Vafaei Sadr, A., & Farsian, F. 2020, arXiv e-prints, arXiv:2004.04177. 2004.04177

Zorrilla Matilla, J. M., Sharma, M., Hsu, D., & Haiman, Z. 2020, arXiv e-prints, arXiv:2007.06529. 2007.06529