Milestone One Report
Machine Learning For Physics - Spring 2024

Hananeh Moballeghtohid

Mohammad Sam Tajik

Dina Mousavi

Introduction

The study of the large-scale structure in non-linear scales is essential as it is tightly connected to the essence of dark matter. The small-scale challenges of cold dark matter could be a hint beyond the standard model of cosmology. In non-linear scales, the matter density distribution is not Gaussian. Consequently, the widely used two-point correlation function is not adequate anymore to capture the matter density field's entire behaviour. Among all statistics beyond correlation functions, nearest-neighbour distribution function seems a promising tool to probe matter distribution in non-linear regime.[1] Moreover, using numerical simulations, we can evaluate and evolve a set of initial conditions of matter and energy that end up merging as the structures observed today, and artificial intelligence and ML methods are becoming more accepted to model relevant features in numerical simulation.[2, 3] Our study employs Machine Learning algorithms to the nearest neighbour (NN) analysis as a new statistical tool for examining the structure and statistics of the cosmic web in an ΛCDM universe.

In this work, we consider the dark matter density as a discrete point process. We performed a simulation with the cosmological code GADGET-2, assuming a ΛCDM Universe. The total number of dark matter particles is 256^3, each with a mass of 1.3 $\times 10^9$ M☉ in a box of comoving length L = 50 h^{-1} Mpc running from z = 100 to z = 0. Our approach involves smoothing the initial dark matter density data (using a window function) to identify peak points in the initial dark matter distribution, which represent the seeds of future halos. By constructing a graph where these peak points are nodes and their nearest neighbours are connected by edges representing distance, we aim to predict the late-time distribution of dark matter halos and their nearest neighbour statistics. This graphical representation not only provides a visual tool for understanding the formation of dark matter halos but also serves as a dataset for training machine learning models.

Drawing inspiration from recent advancements in machine learning applied to cosmology, such as the analysis of dark matter halo structure formation in N-body simulations, our project seeks to predict these graphs using machine learning techniques. By training models on the data derived from our simulations, we aim to predict the structure of dark matter halos with unprecedented accuracy. This endeavour not only contributes to our understanding of the universe's early stages but also paves the way for more sophisticated models that can simulate and predict the evolution of dark matter halos over cosmic time.

Related Articles

1) Mohammad Ansari Fard, Zahra Baghkhani, Laya Ghodsi, Sina Taamoli, Farbod Hassani, Shant Baghram, Structure of cosmic web in non-linear regime: the nearest neighbour and spherical contact distributions, Monthly Notices of the Royal Astronomical Society, Volume 512, Issue 4, June 2022, Pages 5165–5182

2) Analysis of dark matter halo structure formation in N-body simulations with machine learning, Jazhiel Chacón, Isidro Gómez-Vargas, Ricardo Menchaca Méndez, and J. Alberto Vázquez, Phys. Rev. D 107, 123515 – Published 14 June 2023

3) J. Chacón, J. Vázquez, and E. Almaraz. Classification algorithms applied to structure formation simulations. Astronomy and Computing, 38:100527, 2022

Dataset

For this project we had three initial dataset. A dataset for dark matter particles at redshift z = 100, a dataset for dark matter particles at redshift z = 0 and a dataset for halos.
The dataset for halos was a dataset consisting of 1273 samples and 4 (or as someone might say 2) features. Namely position in cartesian coordinates and mass

of halos. This dataset was clean without any duplicates or nans. Also correlation between features was small enough. As such it was almost completely copied for our use.

As for another two datasets, each one had about 16.8 million (or to be exact 256 power 3) samples of dark matter particles and 7(or maybe just 3, as you may prefer) features. Namely velocity in cartesian coordinates, position in cartesian coordinates and id for each particle. These two datasets too were without any nans or duplicates and also correlation between features was small again. Unfortunately these datasets both had a problem, the value ‘0’ was repeated for almost exactly half of the samples. We know as a matter of fact that each sample is different from another in our dataframe and similarity in ids doesn’t represent similarity of samples(particles). So if we only need one of these two datasets we can ignore the id feature and remove it from the dataframe. On the other hand if we need both datasets, we can remove samples with id value of ‘0’ and all ids would be unique.then we can add a feature to each particle that shows its redshift(which is 0 or 100) and add two datasets together. There will be about 8 million samples with common id in the integrated dataset(as mentioned in the notebook).